

Medición y evaluación en educación física

Manuel Rodríguez

PREMIO DEL LIBRO UNIVERSITARIO
“EDILUZ 60 ANIVERSARIO” 2021

EDILUZ
Editorial de la
Universidad del Zulia



UNIVERSIDAD DEL ZULIA

AUTORIDADES

Dra. Judith Aular de Durán
Rectora

Dr. Clotilde Navarro
Vicerrector Administrativo
(encargado del Vicerrectorado Académico)

Dra. Marlene Primera
Secretaria de la Universidad del Zulia
(encargada del Vicerrectorado Administrativo)

Dra. Ixora Gómez
Secretaria (e)

Poeta Dr. Carlos Ildemar Pérez
Director de la Editorial de la Universidad del Zulia



Medición y evaluación en educación física

Una guía para estudiantes y profesionales de la educación física

Manuel Rodríguez

Prohibida su reproducción, adaptación, representación o edición, sin la debida autorización de la Universidad del Zulia o del autor.

Depósito legal ZU2022000112

HECHO EL DEPÓSITO DE LEY

Editorial de la Universidad del Zulia (EDILUZ). Ciudad Universitaria Dr. Antonio Borjas Romero. Facultad de Humanidades y Educación. Sótano del Bloque C. Apartado 526. Teléfonos (424) 5621618

Correo Electrónico: direccionediluz2020@gmail.com

Maracaibo, Venezuela.

Mayo, 2023

**Presentación de la Colección:
Premio del Libro Universitario EDILUZ
60 Aniversario. 2021**

Como director de la Editorial de la Universidad del Zulia (EDILUZ) tomé la decisión de convocar el Concurso del Libro Universitario EDILUZ 60 Aniversario en el año 2021. Este concurso se abrió para la participación de todos los profesores de la Universidad del Zulia, con el fin de promover y difundir el conocimiento humanístico y científico de nuestra Alma Mater, porque existía una deuda de convertir en libros digitales tantos años de investigación a los que sean dedicado la mayoría de nuestros docentes. Alrededor de este premio debemos resaltar dos sucesos históricos importantes. En primer lugar, el reconcomiendo a la trayectoria de una editorial universitaria que alcanza sus 60 años de existencia cumpliendo denodadamente sus compromisos y labores; luego, está la aparición de la pandemia de la COVID-19, que nos llevó a cambiar nuestros hábitos cotidianos, y que a pesar de ello no detuvo el funcionamiento de nuestra editorial en pleno confinamiento y restricciones de todo tipo, creándose propuestas alternativas como fue la convocatoria de este concurso del libro universitario.

En este Concurso de Libros Universitarios EDILUZ 60 Aniversario, se evidenció el compromiso de los profesores y el alto nivel académico de la universidad, resaltando la excelencia investigativa, a pesar de la tremenda crisis a la que ha sido sometida la universidad pública en Venezuela durante más de dos décadas.

De los 57 libros participantes en el Concurso de Libros Universitarios EDILUZ 60 Aniversario, los jurados prestigiosos seleccionaron 16 obras para declararlas ganadoras, y cuyo premio consiste en la publicación digital de las mismas, y que hoy es una realidad cumplida. Con el éxito de esta iniciativa editorial se demuestra que EDILUZ está a tono con los tiempos y su forma de gerenciar responde a las exigencias de innovación y transformación impuestas a nivel mundial, con el fin de derribar barreras y ampliar los horizontes que ofrece la globalización de la digitalidad y las nuevas tecnologías en beneficio del desarrollo universitario.

Reciban todos los participantes y ganadores nuestras felicitaciones.
Sigamos construyendo el futuro de nuestra Universidad del Zulia.

**Poeta Carlos Ildemar Pérez
Director de la Editorial**

Maracaibo, 18 de julio, 2022.

Presentación

En la educación física y, en general, en todos los ámbitos relacionados, una de las cuestiones más interesantes y, al mismo tiempo, más confusa, es el tratamiento que se le deba dar a los resultados de las mediciones que se hacen tanto en el ámbito educativo como en el deportivo.

En el campo de la actividad física, la constante son los datos numéricos. Por ejemplo, en béisbol, entre otros, es el promedio de bateo. En fútbol, la cantidad de goles. En golf, el número de golpes para completar los hoyos. En las carreras atléticas, el tiempo logrado. En lanzamientos, la distancia alcanzada. En gimnasia o en los saltos de trampolín en deportes acuáticos, los puntos otorgados por los jueces.

Esta situación, ejemplificada con el ámbito deportivo también es valedera para el área de la educación física, aunque orientada más hacia el aprendizaje que hacia el rendimiento. Por ejemplo, al realizar prácticas de control del ritmo cardíaco, el resultado son números. Cuando se efectúa un trabajo por estaciones, se registra en número de estaciones, repeticiones y vueltas. Al medir la capacidad aeróbica corriendo una milla se registra el tiempo, o al ejecutar la prueba PACER se registra la cantidad de vueltas. Incluso, el dominio de las destrezas deportivas o la realización de actividades recreativas como construir una cometa o elevarla, y aprender un baile, tienen registros numéricos. En todo caso, es imprescindible que se midan los logros de carácter físico, deportivo y/o recreativo que se alcancen como resultado del proceso didáctico, o de enseñanza – aprendizaje.

El presente material constituye una guía, una orientación, que está enfocado en el desarrollo de mecanismos y herramientas para el procesamiento de datos como proceso y producto de la enseñanza, el aprendizaje, el entrenamiento, y/o la participación. Está conformado por la compilación de diversos materiales y visiones, que se creen de gran utilidad para el desarrollo científico del campo de la actividad física.

Está dirigido, especialmente, a los estudiantes y profesionales de la educación física, tratando de hacer una contribución a su desarrollo y formación continua.

Índice

	Pág.
Presentación	5
Capítulo 1	Bases de la medición y evaluación en educación física 7
Capítulo 2	Fundamentos estadísticos para medir y evaluar la actividad física 26
Capítulo 3	Medidas de tendencia central 49
Capítulo 4	Medidas de variabilidad 60
Capítulo 5	Conversión de puntajes 75
Capítulo 6	Correlación y regresión 98
Capítulo 7	Fiabilidad y objetividad 112
Capítulo 8	Validez 126
Capítulo 9	Nociones estadísticas para la investigación de la actividad física 139
Capítulo 10	Diseño y selección de pruebas de condición física 184
Capítulo 11	Desarrollo de pruebas de destreza deportiva 202
Capítulo 12	La evaluación cualitativa en educación física 212
ANEXOS	237
Referencias	299

Capítulo 1. Bases de la medición y evaluación en educación física

¿Qué es la evaluación?

La evaluación constituye la tercera fase del proceso didáctico y es un tema que siempre despierta interés en el ámbito educativo. A pesar que la evaluación educativa ha sido discutida ampliamente en cada oportunidad que se convoque y que los docentes se muestran ávidos, plantean interrogantes e inquietudes sobre las innovaciones, realmente el centro de la evaluación ha estado en la actuación del estudiante y no en el proceso de enseñanza – aprendizaje en sí.

En el contexto educativo el mayor esfuerzo debe dirigirse a la evaluación de las situaciones de enseñanza – aprendizaje, evitando el sesgo cuando se utiliza la evaluación para otros fines tal como el mantenimiento de la disciplina o la asignación de una calificación por asistencia, por ejemplo. La evaluación tiene importancia en la determinación del proceso didáctico realmente implementado, porque sus resultados deben tener carácter público con la posibilidad de ser supervisados y valorados por los estudiantes y los demás docentes, además de los padres y las autoridades educativas. En todo caso, se debe hacer al estudiante consciente y partícipe de las decisiones de la vida escolar y en ello la evaluación juega un rol protagónico.

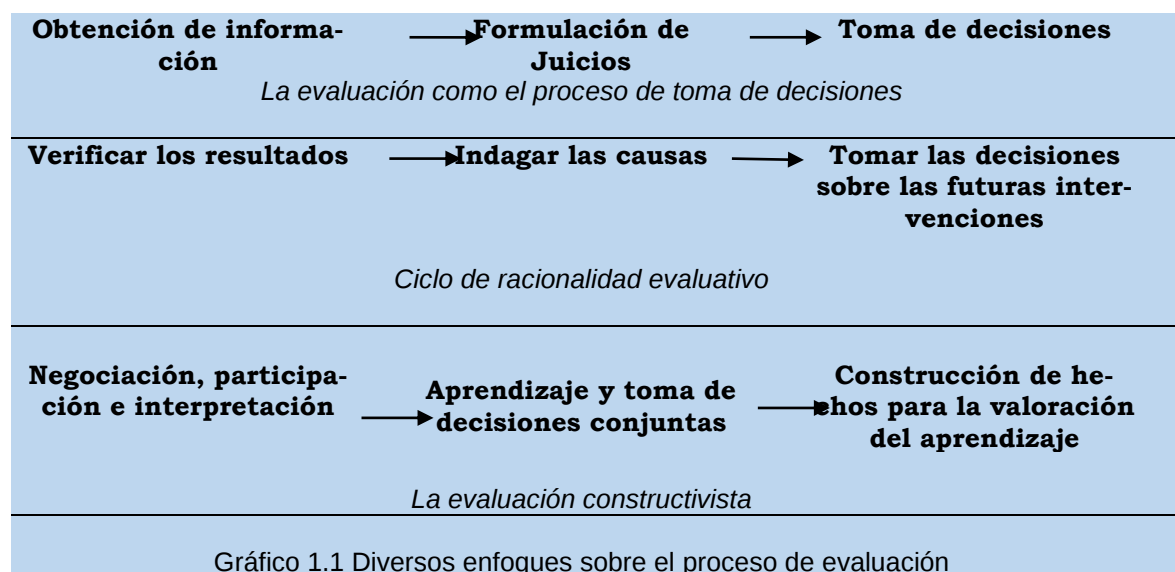
Todo docente debe estar consciente que la evaluación es un momento excepcional, privilegiado, para un aprendizaje de calidad. Después de la planificación y la metodología, no se trata de esperar a ver los resultados de lo ya hecho, porque se puede interpretar que sólo es una obligación administrativa o que se ha dado por concluida la tarea de enseñar.

Corresponde, en primer lugar, la clarificación de términos. La evaluación puede tener diferentes significados si se confunde con los términos test, calificación o medición. Evaluar, simplemente, es asignar un valor a algo, juzgar. Normalmente en la educación se juzga un estudiante, profesor o una asignatura. Algunas formas de juzgar se originan de respuestas a interrogantes como las siguientes: ¿Cómo salta Juan? ¿Se ha despertado en mi grupo nuevos intereses como resultado del trabajo en equipo para jugar? ¿Cuáles técnicas del voleo no domina todavía David? Al responder se están emitiendo juicios, por lo que es difícil imaginar que haya educación sin que se hagan juicios. Ello implica la obtención de información y la emisión de juicios, así como la toma de decisiones: el juicio de que a Juan le cuesta saltar se puede utilizar para decidir en qué grupo de salto colocarlo o qué materiales asignarle. De allí la definición de evaluación: es el proceso de obtención de información y de su uso para formular juicios que a su vez se utilizarán para tomar decisiones.

En otro contexto, evaluar significa verificar los resultados e indagar sus causas. Al verificar los resultados obtenidos tras la intervención docente e indagar las causas de los

éxitos y/o fracasos registrados, es evidente que no se trata de archivar la información obtenida sino la continuación o modificación, parcial o total, del proceso cumplido.

También se presenta la evaluación constructivista, concebida para el mejoramiento del proceso y superación de fallas y no como la etapa final para aprobar o aplazar al estudiante. Es un enfoque alternativo, respondiente, que se da por un proceso interactivo de negociación participativa, utilizándose métodos interpretativos. Se revaloran los papeles del evaluador y del evaluado, al punto que la participación lleva ambos a aprender y tomar decisiones. El evaluador pierde el carácter de juez y se convierte en investigador de procesos para construir hechos en relación con el acto de aprender y su valoración.



Ámbitos de la evaluación

La evaluación presenta diversos ámbitos de aplicación que constituyen un conjunto sistémico de relaciones recíprocas con una característica especial: al tomar una decisión en un ámbito, se afecta el resto de los ámbitos. Los ámbitos responden a preguntas como: ¿Para qué evaluar? ¿Quién debe evaluar? ¿Quién debe ser evaluado? ¿Qué evaluar?

¿Para qué evaluar? Para informar y orientar las decisiones que se toman durante el proceso de evaluación. Fundamentalmente, los fines de la evaluación son mejorar las decisiones con relación al aprendizaje de cada estudiante, informar sobre su progreso a los padres, otorgar los certificados a los estudiantes y a la sociedad, y mejorar la calidad de la enseñanza en general. Indudablemente, se debe informar honestamente a la sociedad acerca del valor, capacidad y conocimiento de los individuos al término del período de enseñanza – aprendizaje, por cuanto alguien tiene que asegurarle a la sociedad una garantía para su desarrollo y bienestar.

¿Quién debe evaluar? Deben participar todos los actores que tengan alguna información útil que aportar, en función de la descripción cualitativa y/o cuantitativa de lo logrado o no en la escuela. Todos los destinatarios teóricos de los resultados de la educación son potenciales evaluadores: docentes, estudiantes, directivos, supervisores, padres y representantes, diseñadores de la política educativa, el gobierno. Con ello se busca lograr el incremento de la objetividad de la evaluación en su conjunto, para completar los aspectos de la evaluación escolar a plenitud. En cuanto a la autoevaluación, no se trata de eliminar un posible poder arbitrario del docente para exaltar el poder del aprendiz. Se trata de no imponer arbitrariedades en el proceso didáctico de la evaluación ya que, si el estudiante propone la nota que cree que merece, debe razonarla ante el responsable de firmar su calificación, para poder darle la validez correspondiente, pública y oficial.

¿Quién debe ser evaluado? La respuesta es obvia: el estudiante, ante la cultura vigente en un sistema educativo que se resiste a estimar los resultados del proceso didáctico. La evaluación del estudiante es clave pues su éxito o fracaso determina la eficacia o ineficacia de todo profesor, sistema didáctico, centro educativo o del financiamiento educativo. Si evaluar significa verificar los resultados y analizar sus causas, es imposible racionalizar el proceso didáctico de la evaluación sin explorar los factores que condicionan el rendimiento observado de cada uno. Además de ello, si se limita la observación a lo que el estudiante hace sin revisar los factores previos como la capacidad de los docentes y si los planes son apropiados o no, la educación puede hacerse irracional. Si el docente tiene una actitud favorable a ser evaluado, se presupone un cambio conceptual profundo acerca de qué es y qué debe ser la enseñanza y su cometido sustancial.

¿Qué evaluar? Una vez que se tienen claros cuáles son los objetivos que los estudiantes deberían lograr, el docente se encuentra ante un océano de posibles comportamientos entre los que debe seleccionar, por razones de espacio y tiempo, unos pocos que se suponen indicadores de la presencia o no de los objetivos, surgiendo dos preguntas: (a) ¿Cuáles conductas o producciones de los estudiantes indican el logro de los objetivos? (b) Al ser teóricamente infinitas las conductas o producciones pertinentes para la indicación de un logro, ¿Cuáles son los criterios para seleccionar unos y excluir otros?

Es imprescindible que el docente tenga algún criterio que no sea la arbitrariedad, el azar o la costumbre, para tomar decisiones que den respuesta a esas interrogantes. Esos criterios necesitan dos componentes: primero, una relación de pertinencia con el objetivo cuyo logro se desea evaluar, y segundo, una clarificación de los comportamientos, operaciones o producciones que se podrán mostrar, para realizar la selección y graduación correspondiente.

Generalidades en medición y evaluación de la educación física

Los términos que se usan en la medición y evaluación de la actividad física tienen significados muy específicos porque se refieren a elementos del proceso de toma de decisiones;

dichos términos están relacionados entre sí y tienen a la vez significados distintivos que deben utilizarse correctamente.

La evaluación y sus principales componentes han sido estudiados profusamente en el ámbito de la educación física; las perspectivas en cuanto a su significado estriban, en primer lugar, en la relación entre evaluación y condición física, y, en segundo lugar, el carácter tanto cuantitativo como cualitativo que se debe dar a los instrumentos utilizados en la evaluación escolar. Se debe comenzar por los términos básicos: test o prueba, medición, evaluación.

Toda ciencia se basa en mediciones precisas y consistentes. Ya sea que usted sea o no un científico o un docente de educación física, hay mucho que usted puede lograr al aprender las técnicas de medición adecuadas. La mayoría de las mediciones son de valores cuantitativos: distancia, fuerza y tiempo. Ahora, contar la frecuencia de ocurrencia de eventos es otra forma de obtener valores cuantitativos.

Test. Se refiere a todos los instrumentos, protocolos o técnicas usadas para medir la cantidad o cualidad de unas propiedades o atributos que interesan al evaluador o investigador. Un test es una herramienta o instrumento utilizado para efectuar una medición en particular, y puede ser escrito, oral, mecánico, fisiológico, o con otra variante. Un test es un tipo específico de medición. Así, la estatura y el peso resultan de una medición, pero no son tests. Los tests se usan para medir la adquisición y retención de conocimientos o habilidades físicas y/o mentales.

Medición. Tiene el propósito fundamental de recolectar información. La medición es un acto de valoración que, por lo general, se expresa mediante la asignación de un número a la cualidad o característica que se valora. La medición determina el grado en el cual un individuo posee una característica definida (velocidad, fuerza, etc.); es decir, es el proceso de recolección de información sobre una propiedad, atributo o cualidad en particular. La medición utiliza el test o prueba siguiendo un protocolo para lograr su propósito: durante la administración de una prueba, la medición surge cuando se obtienen los puntajes. La medición es el proceso de comparación un valor con un estándar. Por ejemplo, podemos comparar nuestro propio peso (la fuerza de gravedad sobre nuestro cuerpo) con el estándar en libras o kilogramos cada vez que subamos a la balanza. Cuando un docente de educación física aplica la prueba de salto largo a los estudiantes, se está cumpliendo el proceso de medición. El estándar con el que el salto se va a comparar es la distancia (metros y centímetros, pies y pulgadas). Cuando el docente usa un instrumento (en este caso, una cinta métrica) para determinar que un estudiante ha saltado 5,20 metros, este bit de información se llama dato. El proceso de medición implica cuatro pasos: a) Se identifica y define el objeto a ser medido; b) Se identifica y define el estándar con el cual el objeto medido será comparado; c) Se hace la comparación del objeto con el estándar; d) Se hace una declaración cuantitativa de la relación del objeto con el estándar.

Evaluación. La evaluación es una afirmación acerca de la calidad, mérito o valor de lo que se ha medido e implica la toma de decisiones. La evaluación es el proceso de determinación del valor de los datos y es un proceso dinámico de toma de decisiones centrado en los cambios que se han realizado. Este proceso involucra tres fases: (a) Recolección de

información o medición de una habilidad, (b) Juicio de valor sobre esa información de acuerdo con algún estándar; (c) la toma de decisiones basándose en los datos anteriores. La evaluación es el proceso de interpretación de la medición efectuada y la determinación de algún mérito o valor en la realización de los sujetos; la evaluación es mucho más amplia que el test y la medición y la evaluación los abarca a ambos. Realmente, la evaluación es un proceso de toma de decisiones subjetivo en el cual las puntuaciones recolectadas son evaluadas para determinar hasta qué punto los objetivos están siendo logrados.

Relación entre medición y evaluación. La medición y la evaluación son conceptos interdependientes. La evaluación es un proceso que utiliza la medición, y el propósito de la medición es recolectar información para la evaluación. En el proceso de evaluación la información se interpreta de acuerdo a unos estándares establecidos para que se tome una decisión; el éxito de la evaluación depende de la calidad de los datos recolectados. Si los resultados no son consistentes (fiables) ni verdaderos (válidos), es imposible una evaluación precisa. El proceso de medición es un paso muy importante en la evaluación ya que una medición de calidad conduce a una evaluación precisa. Las personas son diferentes y varían en tamaño y forma corporal, velocidad, fuerza y muchas otras cualidades: la medición determina el grado en el cual una persona posee una característica definida. La precisión de la medición es esencial. Si la medición no es precisa, los datos no son fiables. Los datos de la medición también deben ser reproducibles, es decir, una segunda medición en las mismas condiciones debe producir el mismo resultado que la primera medición. Los datos que no son precisos ni consistentes no tienen valor. Es posible que no representen las verdaderas condiciones de los sujetos que se miden.

Las pruebas o tests son los instrumentos utilizados para recolectar información. Las pruebas varían en un espectro entre altamente objetivas y altamente subjetivas. Las pruebas que son más objetivas son aquellas que tienen un sistema de registro definido y son aplicadas por evaluadores entrenados. Una prueba escrita de selección múltiple, un cronómetro, un calibrador de pliegues cutáneos o un electrocardiograma, tienen un sistema de registro definido, y se necesita que los evaluadores estén entrenados para asegurar unas mediciones objetivas. Una prueba es altamente subjetiva cuando le falta un sistema de registro estandarizado, ya que introduce una fuente de error en la medición. En la medición de la actividad física se utilizan mediciones objetivas siempre que sea posible ya que son más fiables que las mediciones subjetivas.

Un ejemplo en cuanto a las diferencias entre medición y evaluación es la determinación del consumo de oxígeno. Se puede medir el máximo consumo de oxígeno de una persona (VO_2 máx., una medida de capacidad aeróbica, de varias maneras, por ejemplo, al obtener el rendimiento en una prueba máxima sobre la correa sin fin, recogiendo y analizando los gases espirados. También se pueden recoger los gases espirados en un ciclo máximo sobre el ergómetro. Por otra parte, un sujeto puede efectuar un esfuerzo submáximo en la correa sin fin o en el ergómetro, y luego predecir el VO_2 máx., considerando el ritmo cardíaco y/o el rendimiento alcanzado. Finalmente, se puede medir la distancia que corre una persona en 12 minutos o el tiempo que le toma recorrer 1,5 millas (2414 metros).

Cada una de esas herramientas da resultados en cifras, tal como el % de oxígeno y anhídrido carbónico espirados, el ritmo cardíaco, los minutos empleados, o los metros recorridos. Haber valorado o medido el máximo VO_2 con alguna de esas herramientas no significa que ya se ha evaluado. Obtener y reportar datos tiene poco significado a menos que los resultados se comparen con algún tipo de valor. Aquí entra la evaluación al proceso. Si se le aplica la prueba de consumo de oxígeno a alguien que probablemente sepa poco de ello podría preguntar: “¿Cómo salí?” Si se le señala, “Su máximo consumo de oxígeno es $30 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ ”, simplemente se le dice muy poco. Es necesario suministrar una forma de evaluación ya que una afirmación o conclusión sobre el desempeño de los sujetos introduce el elemento de mérito o calidad de la ejecución, por ejemplo, si se compara el resultado con una tabla de rendimiento.

Por lo general las mismas medidas se aplican en la actividad física, donde comúnmente medimos la distancia (qué tan altas son las personas o qué tan lejos pueden saltar), la fuerza (cuánto pesan o cuánto levantan), el tiempo (qué tan rápido pueden correr o cómo pueden correr a un ritmo determinado en una cinta de correr) y la frecuencia (cuántos pasos se necesitan al caminar durante 10 minutos, cuántas veces late el corazón en un minuto, cuántas repeticiones se hacen de un ejercicio). Se considera que estas mediciones son objetivas por cuanto el resultado se ha registrado con algún tipo de instrumento como un cronómetro o una cinta métrica, por lo que el juicio del evaluador se necesita más en el procedimiento que en el resultado. Otro tipo de mediciones se consideran cualitativas o subjetivas porque requieren fundamentalmente del juicio del evaluador para determinar la calidad del desempeño como en las rutinas de gimnasia o de nado sincronizado.

Evaluación formativa y evaluación sumativa

La evaluación se efectúa bajo dos modalidades, la evaluación formativa y la evaluación sumativa. Las evaluaciones formativas son evaluaciones iniciales e intermedias, tal como la aplicación de un pretest y la subsiguiente evaluación de sus resultados. La evaluación formativa que se puede producir durante la instrucción, el entrenamiento o el proceso de investigación. La continuidad de los procedimientos de evaluación, medición e información sobre el desempeño es esencial para lograr las metas formuladas en un programa vinculado al desempeño humano, ya sea educativo, recreativo o de entrenamiento deportivo. Por ejemplo, luego de una cirugía del hombro, una de las metas propuestas debe ser la recuperación del rango de amplitud del movimiento. Un terapeuta físico puede valorar ese rango y sugerir diferentes actividades para su mejoramiento, observando continuamente el progreso del paciente. Por ello, la evaluación formativa y continua no implica un examen formal; la simple observación y la información del desempeño entre el aprendiz y el instructor, por lo general, son apropiadas.

En la evaluación formativa se hace un juicio de los logros en las diferentes etapas del aprendizaje. La investigación en el aprendizaje motor ha demostrado que el feedback es una de las más poderosas variables en el aprendizaje, y la evaluación formativa lo utiliza en el transcurso del proceso didáctico o instruccional. La evaluación formativa se inicia con el período académico y continúa a lo largo del proceso didáctico. Esto implica dividir

la instrucción en pequeñas unidades de aprendizaje y evaluar el dominio del estudiante de esas unidades durante la instrucción; se busca determinar el grado de dominio de una tarea de aprendizaje dada para entonces puntualizar en la parte de la tarea que no se domina. La fortaleza de la evaluación formativa es su utilización para suministrar el feedback tanto como se pueda en el transcurso de todo el proceso.

Las evaluaciones sumativas son evaluaciones finales que, típicamente, se hacen al término de una unidad instruccional o de entrenamiento deportivo. Por ejemplo, un estudiante debe interesarse por la evaluación sumativa – el nivel alcanzado – que obtiene al finalizar el semestre. La evaluación sumativa de un atleta de clase mundial ocurrirá durante los Juegos Olímpicos, con la obtención o no de una medalla y su ubicación en la clasificación mundial. La evaluación sumativa se utiliza para decidir si se han alcanzado los objetivos generales. La diferencia entre la evaluación formativa y la sumativa parece estar únicamente en el momento en que se recolecta la información; sin embargo, es el uso real de los datos recolectados lo que distingue la evaluación como formativa o sumativa. Por ello, los mismos datos pueden ser usados como evaluación formativa o sumativa.

Tanto la evaluación formativa como la evaluación sumativa son utilizadas en los programas de entrenamiento de la condición física exitosos. El incremento de la intensidad y/o duración de un ejercicio aeróbico puede usarse para suministrar el feedback (es decir, evaluación formativa) al participante que ha mejorado, y puede convertirse en una fuente de motivación para continuar en el programa de ejercicios. Una prueba de condición física posterior a una etapa del entrenamiento puede aplicarse como evaluación sumativa para juzgar si las metas de condición física han sido alcanzadas o si el programa de entrenamiento fue exitoso.

Una manera simple y útil de ejemplificar la forma de aplicar la evaluación formativa y sumativa, es un programa de control del peso. Supongamos que se ha determinado el porcentaje de grasa y el peso corporal de una persona. La evaluación formativa señala que esta persona tiene un 30 % de grasa corporal y requiere perder 5 Kg. para llegar al 25 %. Se le recomienda una dieta y un programa de ejercicios diseñado para producir una pérdida de peso de $\frac{1}{2}$ Kg. por semana durante 10 semanas. Cada semana se le pesa y se le mide el % de grasa, dándole un feedback basado en la evaluación formativa e informándole su progreso semanal. Al final del programa de 10 semanas se mide nuevamente el peso corporal y el % de grasa, evidenciándose la evaluación sumativa: ¿Fueron logradas las metas de reducción de peso y del % de grasa planteadas al inicio? En la tabla 1.1 se resumen las semejanzas y diferencias entre la evaluación formativa y la evaluación sumativa.

Observación y recolección de información

Una de las inquietudes más persistentes en relación con la evaluación la constituye el tipo de información que debe registrarse y la manera de hacerlo. Se ha considerado que los instrumentos pueden ser múltiples y complementarios entre sí, dependiendo de los aprendizajes previstos.

	Formativa	Sumativa
Propósito	Feedback al estudiante y al docente sobre el progreso del estudiante en el transcurso del proceso instruccional	Certificación o calificación al final de la unidad, semestre o curso
Momento	Durante la instrucción	Al final de la unidad, semestre o curso
Énfasis de la evaluación	Conductas definidas explícitamente	Categorías de conductas más amplias o combinación de varios desempeños específicos
Estándar	Referidos a criterios	Referidos a normas o referidos a criterios
Tabla 1.1 Semejanzas y diferencias entre evaluación formativa y sumativa		

La observación es uno de los recursos y estrategias claves para diferenciar la actuación de los estudiantes, tratando de obtener información de la actuación dirigida y/o espontánea de los cursantes por su trabajo individual, en equipo o en pequeños grupos, en la puesta en común o en los debates generales, en las formas de participación, charlas, entrevistas y en la atención individual. También es importante realizar el análisis de las producciones o trabajos de los estudiantes, la aplicación y revisión de pruebas, los trabajos de campo e investigación, la autoevaluación y la coevaluación, el registro de actitudes y valores, que son en realidad formas de clarificar el nivel y consistencia del aprendizaje del estudiante. Esta labor se basa en registros descriptivos y analíticos, listas de cotejo y de verificación, escalas de estimación, entrevistas, pruebas escritas, ensayos, exposiciones, anecdóticos, instrumentos utilizados frecuentemente en la evaluación cualitativa y formativa.

En educación física, más que utilizar los instrumentos de observación en los que se registra la actividad de los estudiantes durante la clase, la observación del docente está dirigida, en primer lugar, a determinar el nivel de intensidad y de dominio que el estudiante está demostrando a objeto de proveerle de la información necesaria y suficiente (feedback) que les lleve a conocer inmediatamente cuál es el resultado de su actuación y qué hacer a continuación. En segundo lugar, a registrar en un formato diseñado para tal fin los elementos de la actuación relacionados con el objetivo de la clase, cuestión que se debe hacer al final de la lección si es posible: mucho tiempo después puede llevar al olvido de lo sucedido. El docente debe ser un observador entrenado, capaz de evaluar con mayor objetividad que los propios participantes, especialmente niños y adolescentes. Dado que la actividad física debe ser examinada y expresada en la medida que ocurre, esto le provee al estudiante de información más precisa que las mediciones posteriores, particularmente porque estas son afectadas por la duración.

El error de la observación. La evaluación se apoya en varias premisas sencillas, directas y lógicas, que dan base a la medición y evaluación. Los individuos tienen capacidades que

se pueden medir, es decir, logran puntuaciones reales. Pero la observación de los resultados de los sujetos contiene errores. Se supone que la puntuación obtenida a partir de la observación de un individuo no es la misma que su puntuación real ya que erramos cuando medimos, de manera tal que la medición comprende tanto la puntuación observada como el error, que puede ser positivo o negativo.

Si una puntuación es mayor que la puntuación real de la persona, el error es positivo. Si la puntuación obtenida es menor, el error es negativo. No se puede saber nunca cuál es la puntuación real de una persona o la cuantía de error que existe en una puntuación dada. Existen procedimientos para estimar la puntuación real de un individuo y su error, pero nunca sabremos exactamente el grado de error cometido en un caso específico: la puntuación real siempre permanecerá desconocida.

Asimismo, hay tres tipos de errores comunes en los observadores, exigiéndoles que estén conscientes de ellos y que aprendan a detectarlos en sus propias observaciones. El primer error es el efecto halo, que se produce cuando un observador está influido por la impresión que le causa la persona que observa: si la impresión es buena lo califica con alta puntuación en todas las características que se juzgan y, viceversa, si es poco favorable la calificación es baja.

El segundo error que cometen los observadores es la tendencia a la respuesta personal: unos observadores tienden a otorgar calificaciones altas a todos – el error de generosidad – y otros tienden a calificar bajo a todos, el error de severidad. También puede existir el error de tendencia central al calificar a todos medianamente. El error por tendencia también se manifiesta en la utilización del anecdotario cuando sólo se anota el lado positivo o negativo de las anécdotas.

El tercer tipo de error es el error lógico. El observador no entiende la relación entre las variables que interactúan en una situación concreta y, en consecuencia, sus observaciones son influidas por lo que él piensa de las relaciones entre las características de las variables. Por ejemplo, valoran a los estudiantes muy estudiosos y minusvaloran los poco estudiosos, porque suponen que existe una fuerte relación entre la inteligencia y las buenas notas. Muchos docentes suponen que los niños callados y dóciles son buenos estudiantes, o que el estudiante que habla mucho en los temas colectivos sabe mucho de la materia.

Evaluación referida a normas y referida a criterios

Para tomar las decisiones en evaluación el docente debe disponer de alguna forma de orientarse. La evaluación es el proceso de darle significado a una medición juzgándola con respecto a algún estándar. Los dos tipos de estándar ampliamente utilizados son los estándares referidos a normas y los estándares referidos a criterio. Los estándares referidos a criterio son útiles para el establecimiento de estándares de desempeño para todos, mientras los estándares referidos a normas son valiosos para las comparaciones entre individuos cuando la situación requiere un grado de selectividad.

Los estándares referidos a normas se desarrollan al evaluar un gran número de personas de un grupo definido, por ejemplo, adolescentes de 14 años de edad. Luego se emplea la estadística descriptiva para desarrollar los estándares. Uno de los métodos para desarrollar normas es el de rangos percentiles; este tipo de normas refleja el porcentaje de integrantes del grupo que se espera que obtenga un puntaje por debajo de un puntaje dado. Por ejemplo, un tiempo de 11:31 minutos en la carrera de una milla para un niño de 11 años de edad está en el percentil 25, es decir, sólo el 25% corre más lento mientras se espera que el 75% de los niños de 11 años superen ese tiempo. Una decisión evaluativa que utilice valoraciones con base en normas significa que se reportará cuán bien se compara el desempeño de un sujeto con respecto al desempeño de otros, quizá con gente del mismo sexo, edad o clase. Así, se podría reportar que un consumo de oxígeno de $30 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ es relativamente pobre con respecto a su grupo de edad y sexo en particular.

Un estándar referido a criterios es un estándar predeterminado que muestra que el individuo ha logrado un nivel deseado de rendimiento. Es diferente a un estándar referido a norma en que el rendimiento de un individuo no se compara con el de otros individuos, sino sólo contra el estándar. Por ejemplo, se pudiese reportar el desempeño de una persona con respecto al criterio que se le pide lograr. Supongamos que se midió el VO_2 máx., de alguien que tuvo un ataque cardíaco. El médico podría estar interesado en determinar si el paciente puede lograr un VO_2 máx de $25 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$. Este es el caso de una valoración con base en criterios; se tiene poco o ningún interés en comparar unos con otros, sino que la comparación es con un estándar o criterio predeterminado.

Fiabilidad, objetividad, validez

Hay ciertas características esenciales para una medición: sin ellas, se puede tener poca fe en las mediciones y en realidad se hace poco uso de ellas o no se hace. La característica más importante de una medición es la validez. La segunda cualidad más importante de una medición es la fiabilidad. Una medida fiable no cambia constantemente en un corto período de tiempo. Es decir, si un individuo cuya habilidad no ha cambiado se mide dos veces, por ejemplo, hoy y en uno o dos días, los dos puntajes serán idénticos. Para que una medición tenga validez, debe ser fiable. El error de medición disminuye la exactitud de la puntuación observada, lo que se puede considerar como inconsistencia: entre una y otra medida no hay acuerdo, no son consistentes. El término estadístico que se usa para describir la inconsistencia es la falta de fiabilidad. Fiabilidad es, pues, un término que describe la consistencia que existe entre las medidas, la ausencia del error de medición. Por otra parte, una puntuación en matemática puede ser fiable pero no serviría si lo que se necesita es información sobre capacidad anaeróbica. La información que se consigue a través de la medición debe ser fiable y apropiada, y también debe ser la información que se necesita, es decir, los puntajes obtenidos.

Otra característica importante de una medición es la objetividad. La objetividad a veces se llama fiabilidad del evaluador, porque se define en términos del acuerdo entre jueces competentes sobre el valor de una medición. Por lo tanto, si dos jueces que evalúan al mismo individuo en la misma prueba no logran ponerse de acuerdo sobre un puntaje,

las mediciones carecen de objetividad y ninguno de los puntajes tiene fiabilidad ni validez. La falta de objetividad, entonces, reduce tanto la fiabilidad como la validez. La objetividad es necesaria para que haya fiabilidad, y la fiabilidad es necesaria para que haya validez. Entonces, al determinar si existen estas tres características esenciales en una medición, se determina primero la objetividad y la confiabilidad y luego se estima la validez.

La definición de validez está en la interpretación de la puntuación de la prueba. Por lo tanto, las interpretaciones basadas en los puntajes de las pruebas son válidas (correctas). Por ejemplo, la interpretación en cuanto a la cantidad de fuerza de la cintura escapular que se basa en los puntajes de la flexión de codos (push-up) es válida. Para que exista validez, los puntajes de las pruebas deben ser fiables, pero la fiabilidad no garantiza interpretaciones válidas. Obviamente, para que las interpretaciones de los puntajes de las pruebas sean válidas, la prueba o el instrumento deben generar mediciones consistentes. Así, si las personas van a tomar la prueba de correr la milla y han tenido experiencia previa en la prueba, la carrera de la milla arrojará puntajes fiables (estables día a día) a partir de los cuales se pueden hacer interpretaciones válidas de la capacidad aeróbica, pero alguna interpretación con respecto a la velocidad de la carrera no sería válida. Pero, ¿qué puntaje de carrera de milla no fue fiable (cambiante día en día)? Un día, una persona corrió la milla en 13 minutos, pero dos días después corrió la milla en 10 minutos. La interpretación de 13 minutos sería notablemente diferente de la interpretación de 10 minutos, y ninguna interpretación puede ser correcta. Las interpretaciones de los puntajes de la carrera de la milla no serán válidas en esta situación. Si las interpretaciones basadas en los datos de un instrumento de prueba o medición no son válidas, los datos son inútiles. Cualquier decisión que se tome sobre los individuos con base en estos datos, será errónea.

La evaluación cualitativa

Actualmente, en muchos países los programas de estudio en la educación primaria y secundaria por lo general reflejan el enfoque constructivista, con una evaluación multidireccional y centrada en procesos que considera todos los elementos, relaciones e influencias que permiten el mejoramiento continuo. La necesidad de redimensionar el proceso de enseñanza – aprendizaje y lograr cambios que ayuden a superar dificultades para la incorporación de las personas al mundo del trabajo y a la sociedad en general, hace imperativo el uso de métodos y actividades que tomen en cuenta al niño como centro del saber, al maestro como mediador y a la escuela en integración con su comunidad.

La enseñanza es concebida como un proceso planificado e intencional cuya finalidad es orientar y dirigir las experiencias de los estudiantes, atendiendo los hechos y valores de sus culturas y vidas. El aprendizaje se define como un proceso continuo, permanente, individual, constructivo, que implica convalidar las aproximaciones cada vez más acabadas a los objetos de conocimiento. Ambos procesos deben darse en forma conjunta e interactiva; el docente debe planificar y provocar las condiciones necesarias para producir los encuentros y así favorecer la construcción del saber por parte del estudiante. El enfoque tradicional de evaluación tiene carácter unidireccional, donde el único evaluador

es el docente, el único medido más no evaluado es el estudiante, y el único fin de la evaluación es administrativo: se evalúa para calificar y la decisión será la aprobación o no del estudiante. La visión unidireccional de la evaluación centra la atención en el producto, no en los procesos, y eso no permite una auténtica evaluación continua ya que se hace énfasis en la memorización con la ausencia de conocimientos verdaderamente consolidados. La instrucción y el aprendizaje están separados, constituyen acciones paralelas que en ningún punto se cruzan.

La evaluación de los aprendizajes fundamentada en el enfoque cualitativo hace del contexto del proceso de enseñanza – aprendizaje un espacio para la reflexión, comprensión y valoración de los avances, intereses, aspiraciones e interpretaciones de quienes participan en la acción educativa. Se utiliza la etnografía como metodología, interesándose en lo que hacen las personas dentro del grupo. La evaluación es un proceso de investigación para recoger y analizar, continuamente, evidencias sobre lo que acontece en el acto de enseñar y aprender. Con la evaluación cualitativa se describen, narran e interpretan los avances, aciertos y errores acontecidos en el proceso de construcción de los conocimientos, y se intenta comprender los fenómenos y situaciones como un todo tomando en cuenta el desarrollo evolutivo de los niños para ajustar la enseñanza y sentar las bases para la construcción de los conocimientos más complejos. Se evalúan tanto los significados personales o experiencias previas, así como el proceso de construcción del conocimiento, relacionados con el contexto y las competencias de grado, evidenciadas en los indicadores respectivos. A este respecto, la evaluación cualitativa se define como multidireccional, cualitativa, naturalista y constructivista.

- Multidireccional: Se evalúan todos los actores y factores que intervienen en los procesos de enseñar y aprender. Su objeto no es clasificar a los estudiantes en excelentes, regulares, etc. El propósito es que todos alcancen las competencias propuestas en cada una de las áreas, según sus posibilidades, unos primeros y otros más tarde, que es el verdadero criterio de efectividad.
- Cualitativa: Es descriptiva y busca explicaciones relacionadas con los resultados, considerando la influencia del contexto social y familiar en los procesos. La finalidad es permitir que los que hayan logrado las competencias o hayan ido más allá mantengan el nivel logrado, y los que no lo hayan logrado deben recibir un tratamiento adecuado y oportuno para alcanzar la nivelación.
- Naturalista: Se entiende la evaluación como una capacidad natural y, por lo tanto, inherente a la condición humana. Es un derecho inalienable de todo aquel que aprende o participa en la enseñanza, de valorar su trabajo.
- Constructivista: Los cambios que se dan como consecuencia del aprendizaje no se producen en forma lineal, sino que la nueva información incrementa los niveles de complejidad. El alumno valora los hechos y entiende sus datos como representaciones de los procesos del aprendizaje en construcción.

En el caso de la medición y evaluación en el área de educación física, es necesario diferenciar las condiciones en que se trabaja en el aula de clase de las condiciones en que se trabaja en la instalación deportiva donde se imparte regularmente una clase de educación física. Mientras en el aula de clase las actividades se desarrollan en grupo trabajando en el tema asignado, o individualmente en algún tipo de consulta bibliográfica u otras

actividades, al mismo tiempo que el docente observa, registra y orienta las actividades cumplidas y el comportamiento de aprendizaje del estudiante, en la clase de educación física los estudiantes están en movimiento continuo en el proceso de adquisición de las destrezas de movimiento ya sea relacionados con la psicomotricidad, los movimientos básicos, las destrezas deportivas e incluso, con la condición física relacionada con la salud. En educación física los procesos didácticos son mucho más complejos en el sentido de que el movimiento continuo es el principal factor a considerar para que el estudiante practique, automatice, demuestre y aplique los diferentes tipos de destrezas. El docente, al contrario del docente de aula, efectúa el registro de sus observaciones después de finalizar el encuentro, ya que durante la clase debe estar atento al movimiento de los estudiantes y de la información o feedback que debe hacerle llegar a cada uno al observar sus movimientos.

¿Cómo obtener la información necesaria para la evaluación cualitativa? Al haber énfasis en la evaluación cualitativa, lo cual dirige la atención hacia la forma y tipo de información que se desea recolectar, por lo que el docente debe preparar conscientemente los instrumentos a utilizar.

En primer lugar, al diagnosticar se debe localizar la información disponible que se encuentra en las notas del propio docente, de otros docentes o de los padres, o en los archivos del plantel. Luego, es necesario verificar la fiabilidad y validez de esa información respondiendo las siguientes cuestiones: ¿Encaja la información disponible con la información requerida? ¿Es exacta la información? ¿Hay fuentes de error en la información? ¿Qué utilidad tiene la información?

Si la información no está disponible, se decide cuándo, dónde y cómo obtenerla, y se seleccionan las técnicas e instrumentos apropiados. Las técnicas pueden ser la observación, el interrogatorio, el análisis y los tests. Entre los instrumentos de observación están los anecdóticos, las listas de control, las escalas de estimación, las listas de cotejo y las hojas de registro. Los instrumentos para interrogar pueden ser cuestionarios, entrevistas, entre otros. Los instrumentos de análisis se refieren a realizaciones como las exposiciones y demostraciones, productos no – verbales como mapas y pinturas, las tareas de adquisición y de repaso. Los tests pueden ser elaborados por el docente o aplicarse tests estandarizados.

En el caso de las técnicas de la evaluación cualitativa, se debe considerar que la observación requiere de concentración y habilidad para obtener información válida y fiable. Es indispensable la utilización de los mejores instrumentos de observación disponibles y estar muy familiarizados con su empleo; observar constantemente, verificando y comparando la información obtenida, es vital. Para observar con propiedad se debe ser objetivo. La mayor fuente de error en la información obtenida por observación es el sesgo personal y los prejuicios del observador. Es importante concentrarse en el comportamiento observado o sus características específicas, en lo que se percibe (se ve, se oye, se toca) en lugar de lo que se siente (actitudes y opiniones).

El docente debe centrarse en los comportamientos significativos. Cuando se observa es muy fácil distraerse, y el observador empieza a ver cosas distintas a las que intenta

observar. El docente debe ser discreto y pasar inadvertido al observar; sin atraer la atención hacia sí con gestos de aprobación o no, o mostrando sus notas. Es importante que los estudiantes no se den cuenta que se les observa, incluso anotando las observaciones después del encuentro. Por lo demás, se debe observar a menudo, sistemáticamente, como parte natural de las actividades diarias. Sólo con la práctica constante se convertirá en un observador justo y exacto, e incrementará su capacidad de observación. Efectuando muchas observaciones se tienen más posibilidades de obtener indicadores precisos de cómo son sus estudiantes y lo que son capaces de hacer.

Propósitos de la evaluación en educación física

Los docentes e investigadores de la actividad física, los promotores de la salud y quienes trabajan en la industria de la condición física, deben comprender el proceso de evaluación, medición y aplicación de pruebas porque tienen el deber de tomar decisiones evaluativas diariamente, además de prepararse para responder acerca de cuáles son las mejores herramientas para interpretar las mediciones realizadas y el desempeño del ejecutante.

Cada docente descubrirá durante sus estudios y en el ejercicio profesional que necesita conocimiento y destreza para aplicar la medición y la evaluación en forma correcta y efectiva. Así como en otras profesiones, es necesario tener una profunda comprensión de los propósitos del proceso de medición y evaluación en educación física, empezando por los propósitos generales de dicho proceso: ubicación, diagnóstico, predicción, motivación, evaluación del logro, mejoramiento, evaluación del programa.

Ubicación. Tanto un test como otro tipo de evaluación inicial le permiten al profesional agrupar a los estudiantes de acuerdo con sus habilidades, en grupos instruccionales en clase o para el entrenamiento deportivo.

Diagnóstico. Los resultados de los tests por lo general se utilizan para determinar las habilidades y deficiencias de estudiantes, atletas y otros participantes en programas de condición física., cuando se les aplica una o varias pruebas. Con la ubicación se determina el lugar del individuo con respecto a los demás participantes, mientras el diagnóstico se utiliza para especificar las deficiencias individuales y la subsiguiente elaboración de un programa de baja intensidad inicial.

Predicción. Una de las metas de la investigación científica es predecir los futuros resultados partiendo de datos actuales o anteriores. Los puntajes de un test pueden predecir el futuro del estudiante en una universidad. Por ejemplo, un especialista del ejercicio físico puede utilizar las cualidades físicas, las mediciones de la resistencia cardiovascular, la presión sanguínea, el porcentaje de grasa u otros factores vinculados a la condición física, para predecir el riesgo de desarrollo de enfermedades cardiovasculares. En otro caso, el desempeño en el último partido de fútbol puede utilizarse para predecir el éxito o no en los próximos eventos. Así se puede ayudar a los interesados en seleccionar las actividades que sea más probable que lleguen a dominar.

Motivación. Se necesita del proceso de medición y evaluación para motivar a los estudiantes y participantes en cualquier programa de actividades físicas, ya que la gente necesita del reto y estímulo que obtienen al evaluar sus logros. No habría atletas si sólo se practicara, sin oportunidades de competencia, y es difícil estar motivado si se desconocen los resultados de las pruebas y su evaluación.

Evaluación del logro. En un programa educativo, se debe formular un conjunto de objetivos que oriente la evaluación del nivel de logro de los participantes. El mejoramiento del desempeño humano es un objetivo importante en los programas de entrenamiento e instrucción, pero ese desempeño es muy difícil evaluarlo con claridad y precisión. El desempeño exitoso en un conjunto de pruebas iniciales puede estimular la búsqueda de niveles más altos de desempeño. La valoración del logro es una tarea de la evaluación sumativa que requiere del proceso de medición y evaluación.

Mejoramiento. Es la diferencia en el desempeño durante un período, entre el nivel inicial explorado y el nivel final alcanzado. El desempeño puede ser medido a lo largo de la unidad, semestre o año escolar, y ser valorado tanto por el docente como por el estudiante. Para un estudiante con un nivel inicial bajo, ser capaz de verificar su propio progreso es particularmente importante.

Evaluación del programa. El docente está obligado a evaluar el programa que dicta, de manera tal que justifique los procesos instruccionales cumplidos. La meta de la evaluación del programa es demostrar, en forma transparente, el logro exitoso de los objetivos del programa ante las autoridades. Al docente de educación física se le puede exigir que demuestre que su grupo está practicando la condición física en forma apropiada. Para hacerlo, el docente puede comparar los resultados del test de condición física con los resultados de otros estudiantes del distrito o municipio escolar o con las tablas de normas nacionales, si

trabajo del docente profesional podrían de capacidad efectiva y evaluación del

Propósito: Desarrollo de destrezas neuromusculares
Objetivo general: Desarrollo del equilibrio
Objetivo terminal: Demostración de equilibrio estático en posición estacionaria vertical
Objetivo específico: Equilibrarse estáticamente al menos durante 20 segundos, en uno de los tres intentos del test Bass Stick

existieran. El y su futuro como depender de su aplicar una apropiada programa.

Tabla 1.2 Vinculación entre los propósitos y los objetivos específicos

Principios de la evaluación en educación física

Existen un conjunto de principios que sustentan y orientan la acción docente con relación a la determinación del progreso del estudiante. Algunos de estos principios ya han sido reseñados, tal como la evaluación en base a normas o criterios y la evaluación formativa y sumativa. A continuación, se señalan otros importantes principios de la evaluación en educación física.

La medición y la evaluación deben utilizarse como medios y no como fines. La medición y la evaluación no son fines en sí mismas y aplicar pruebas por aplicarlas no sólo constituye una pérdida de tiempo sino una verdadera obstrucción al proceso educativo en general. A los resultados de las mediciones efectuadas se les debe dar un uso claro y definido, ya que la medición y la evaluación son los medios por los cuales el docente demuestra el cumplimiento de su tarea de educar a los estudiantes.

La medición y la evaluación deben relacionarse con metas y objetivos. La medición debe conducir hacia el propósito de evaluar los resultados, a la luz de las metas educativas. De esta manera, los objetivos generales orientan el proceso educativo hacia los objetivos específicos. Por ejemplo, un propósito y un objetivo general relacionados con el desarrollo neuromuscular, llevan hacia objetivos específicos como el dominio de los movimientos básicos (Tabla 1.2).

La medición y la evaluación deben ser utilizadas para determinar el valor de equipos, materiales y métodos. La medición permite revisar el mejoramiento de los estudiantes con respecto a los métodos de enseñanza. Con la determinación de la eficiencia de los equipos y materiales mediante mediciones repetidas y su posterior análisis, es posible identificar y eliminar equipos deteriorados, así como abandonar métodos contraproducentes para el desarrollo del programa.

La medición y la evaluación deben utilizar tanto técnicas objetivas como subjetivas. Al efectuar mediciones objetivas para evaluar el dominio de un programa se tiene una gran oportunidad para determinar la efectividad de su aplicación, lo cual es importante para los estudiantes, representantes y autoridades educativas. Sin embargo, no se debe olvidar que en educación física algunas tareas requieren de la evaluación cualitativa, y el docente debe esforzarse por realizar este tipo de evaluación tan objetiva como sea posible, utilizando los instrumentos apropiados.

La medición y la evaluación son precedidas por la instrucción y la práctica. A menos que el test o prueba esté diseñada para medir el nivel inicial del estudiante, la medición y la evaluación deben efectuarse después del proceso de instrucción, asegurándose que

sienta confianza y familiaridad con las condiciones de la prueba. Más aun, la confiabilidad y la validez de un test son mayores cuando los estudiantes están familiarizados con los ítems de las pruebas y han superado las variaciones iniciales del desempeño cuando se aprende un tema. En muchos casos, la observación y un sano juicio son suficientes al comienzo de la instrucción, lo que permite al docente evaluar subjetivamente el nivel inicial del grupo sin tener que emplear un tiempo valioso instruyendo el test y sin desconcertarlos con una prueba formal.

La medición y la evaluación deben ser conducidas de manera profesional. En el proceso de medición y evaluación, todas las condiciones deben ser iguales para todos los estudiantes dentro de un grupo asegurando así la confiabilidad y validez de las pruebas, ya que un cambio en las condiciones puede producir resultados diferentes. Si un docente intenta motivar a un estudiante durante una prueba con palabras de estímulo o información especial, se debe tratar a todos los estudiantes de la misma manera. Esto también significa que las instrucciones de la prueba deben ser claras y explícitas, para que los resultados sean precisos.

La medición y la evaluación deben ser conducidas por personal entrenado. En tanto sea posible, sólo el personal adiestrado y experimentado debe conducir los procesos de medición y evaluación. Los docentes deben procurar constantemente el mejoramiento en la aplicación de las pruebas, evaluando su utilidad real para el proceso educacional. Como la medición no es infalible, se debe reconocer que la formación de esas personas es la clave de las evaluaciones exitosas, y no el dispositivo de evaluación por sí mismo. Un test que reporta alta confiabilidad y validez puede dejar de tener valor si no es administrado correctamente.

La medición y la evaluación consideran la totalidad individual y del ambiente. La evaluación debe ser interpretada en términos del individuo total en relación con su ambiente. Un estudiante que ha actuado excepcionalmente bien por varios meses, pero repentinamente comienza a hacerlo mal debido a circunstancias más allá de su control, tal como lo problemas emocionales derivados de la muerte de un familiar, debe ser evaluado sobre una base diferente.

La medición y la evaluación de la educación física deben funcionar armónicamente con la filosofía escolar. Para la implementación armónica del programa de educación física y los consiguientes procesos de medición y evaluación, debe aprovecharse cada oportunidad para adecuarse a la filosofía educativa, así como coordinar debidamente con las demás asignaturas. No es raro encontrar una filosofía educativa diferente a la filosofía personal del docente. Cuando esto sucede, lo mejor es que el docente busque la manera de resolver el conflicto, ya sea trabajando para que haya un cambio en la filosofía escolar o con la modificación de su propia filosofía.

En resumen, la relación entre test, medición y evaluación indica que el docente de educación física tendrá que tomar decisiones con respecto a los métodos de recolección e información de los datos arrojados por la medición, como lo representa el siguiente gráfico.

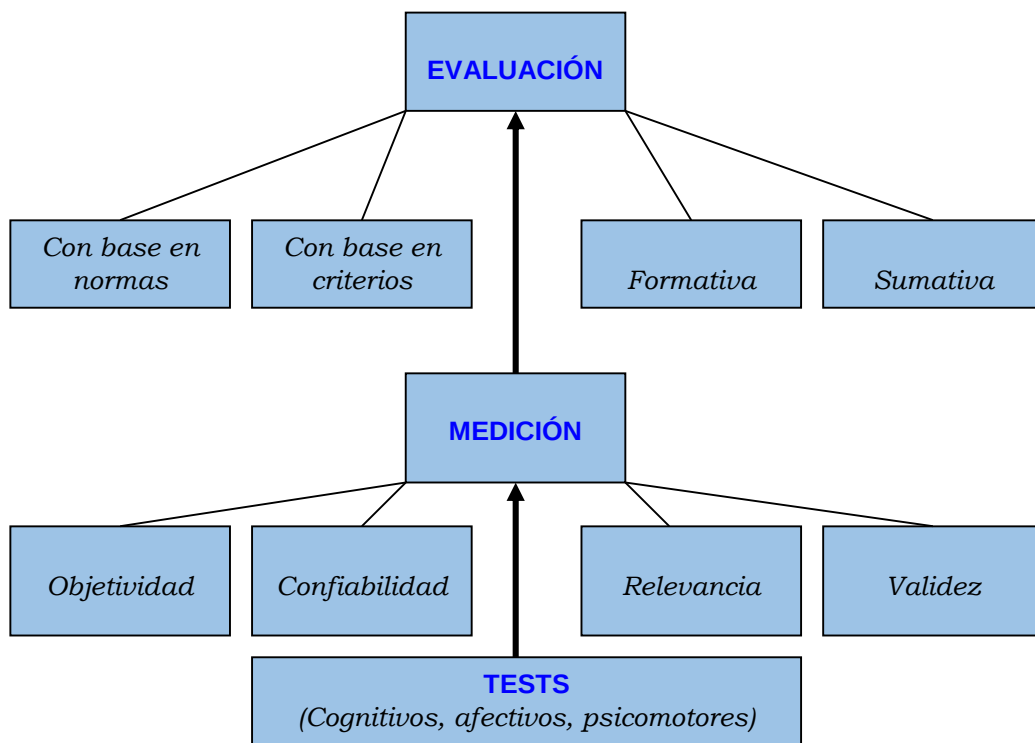


Gráfico 1.2 Relación entre test, medición y evaluación

EJERCICIOS

1. Muchos docentes emplean los términos medición y evaluación indistintamente. Defina cada término.
2. ¿Cuáles son las diferencias claves entre medición y evaluación?
3. Explique cómo se relacionan la medición y la evaluación.
4. ¿Cuáles son las diferencias entre evaluación referida a criterios y evaluación referida a normas?
5. Defina y diferencie entre evaluación formativa y evaluación sumativa.
6. Muchos docentes creen que los alumnos producirán un mejor desempeño si se utilizan los dos tipos de evaluación, la formativa y la sumativa. ¿Qué cree usted?
7. Señale 2 propósitos de la evaluación en educación física y explíquelos.
8. De 2 ejemplos de situaciones reales relacionados con los principios de la evaluación en educación física.

Capítulo 2. Fundamentos estadísticos para medir y evaluar la actividad física

Cuando se quiere determinar el nivel de condición física de una persona o de un grupo de personas, se recurre a un conjunto de procedimientos que permitan establecer en qué lugar se encuentra de acuerdo con su demostración o desempeño. Si se aplica una prueba de velocidad, por ejemplo, correr 30 metros planos, el resultado de un sujeto debe ser comparado con el resultado de otros sujetos de manera tal que permita la identificación de su ubicación con respecto a la velocidad exhibida por los demás en esa prueba.

Esto implica, en primer lugar, el desarrollo o selección de una prueba apropiada para medir el logro con relación a unos objetivos previamente establecidos. Luego de la aplicación de prueba, los datos obtenidos se procesan para proceder a su comparación. Todo ello se basa en una medición precisa y consistente que requiere el dominio o aprendizaje de una serie de técnicas. Los científicos del movimiento y del ejercicio miden las cualidades del desempeño humano en un laboratorio o en el campo, mientras los docentes y los entrenadores realizan las mediciones solo en el campo, y ambos resultados requieren de la aplicación de técnicas estadísticas para su procesamiento, análisis y comprensión.

Una vez que el resultado del desempeño (los puntajes o datos de las pruebas) ha sido recolectado, debe ser analizado. Este análisis se refiere a la comparación de los valores obtenidos con un estándar, ya sea un estándar referido a normas o un estándar referido a criterios, y esto es básicamente el proceso de medición y evaluación. Al medir el peso de una persona, se compara con una escala en kilogramos o en libras. La distancia obtenida en un salto se compara con una escala en metros o en pies y pulgadas. Si un docente utiliza una cinta métrica para señalar que un estudiante alcanzó una distancia de 5,20 metros, este resultado se llama dato.

Hay muchas razones por las cuales analizamos conjuntos de puntajes de exámenes. Para un grupo grande, una simple lista de puntajes no tiene sentido. Solo al condensar la información y aplicarle términos descriptivos podemos interpretar el desempeño general de un grupo, su mejora de año en año o desde el comienzo de una unidad de enseñanza o entrenamiento, o su desempeño en comparación con otros grupos de antecedentes similares. El análisis de puntaje también se utiliza para evaluar el logro individual. Una vez que se ha obtenido información sobre el desempeño general de un grupo, se puede evaluar el logro de un individuo con relación a él, o con respecto a un criterio. El análisis también ayuda al instructor o al terapeuta a desarrollar estándares de rendimiento, ya sea con fines de evaluación o simplemente para que las personas sepan cómo les está yendo.

Datos, variables, constantes

Los datos o puntajes constituyen el resultado de la medición. Existen varias razones que justifican el análisis de los datos como un conjunto. Para un grupo de personas muy grande, una simple lista de los datos no tiene significado. Sólo condensando la información y aplicando elementos descriptivos se puede interpretar el desempeño general del grupo, su mejoramiento de año a año, o comparar el desempeño desde el comienzo de un período de enseñanza o de entrenamiento al final, o con otros grupos de similar experiencia.

El análisis de los datos también se utiliza para evaluar los logros individuales. Una vez que la información sobre el desempeño general del grupo se obtiene, cualquier logro individual puede evaluarse con relación al comportamiento del grupo. Asimismo, el análisis de los datos ayuda al evaluador a desarrollar estándares de desempeño, ya sea con propósitos de evaluación o simplemente para que los individuos conozcan cómo lo están haciendo. Por otro lado, es esencial que haya precisión al realizar la medición. Si la medición no es precisa, los resultados no son verdaderos. Los datos de la medición también deben ser reproducibles; esto es, al realizar una segunda medición bajo las mismas condiciones deben producir los mismos resultados de la primera medición. Los datos que no son precisos ni son consistentes, no tienen validez: esos datos no representan las condiciones reales de los sujetos que están siendo medidos.

Variable. Cuando medimos el desempeño humano, medimos variables. Una variable es una característica de una persona, lugar o cosa que puede asumir más de un valor. Por ejemplo, una característica de una persona que varía es el tiempo que tarda en correr una milla. Otros ejemplos de variables son el peso, la altura, el número de tiros libres de baloncesto realizados en 10 intentos, el consumo de oxígeno y la frecuencia cardíaca durante una prueba de cinta, o los ángulos en la articulación de la rodilla durante varias fases de una zancada. Por ejemplo, si la estatura en centímetros de tres personas es de 170, 178 y 171, estos tres datos están dentro de la variable estatura que se está midiendo. Tenga en cuenta que diferentes personas obtienen puntajes diferentes en la misma variable, y una sola persona puede desempeñarse de manera diferente cuando se le mide la misma variable más de una vez. Los datos varían entre las personas y también dentro de las personas. Debido a que las variables cambian, debemos monitorearlas o medirlas con frecuencia si necesitamos conocer el estado actual de la variable.

Constante. Una característica que puede asumir sólo un valor se llama constante. Como una constante nunca cambia, una vez que la medimos con precisión no la tenemos que volver a medir. El número de jugadores en un equipo oficial de béisbol es una constante. Deben ser nueve, ni más ni menos. Las distancias en los eventos de pista son constantes. En una carrera de velocidad de 100 metros, la distancia es siempre de 100 metros. Las distancias de base a base en un campo de béisbol y desde el montículo del lanzador hasta el plato son constantes. Muchas características anatómicas son constantes. En general, una persona tiene solo una cabeza, un corazón, dos pulmones y dos riñones. Por supuesto, las enfermedades o los accidentes pueden desfigurar el cuerpo y, por lo tanto, cambiar algunas de estas características, pero típicamente la anatomía del cuerpo humano es constante. Probablemente la constante

más conocida en matemáticas es π . π es el número de veces que el diámetro de un círculo se puede colocar alrededor de su circunferencia, y siempre es 3,14159.

Tipos de datos

Es importante determinar cuándo recolectar y registrar los puntajes de una prueba, así como seleccionar la técnica de análisis de los datos. Para ello se debe conocer el tipo de datos que arrojará la prueba. Los datos pueden ser clasificados de varias maneras. En primer lugar, tenemos los datos continuos y discretos.

Datos continuos. Los datos continuos, que son la mayor parte en el área de la actividad física, tienen un número potencialmente infinito porque pueden ser medidos con variados grados de precisión. Entre dos valores de unos datos continuos existen innumerables valores que se pueden expresar como fracciones. En el caso de una carrera de 100 metros planos los datos se podrían registrar en segundos y décimas de segundos, pero también podrían registrarse en segundos y centésimas de segundo o en segundos y milésimas de segundo. La cantidad de peso que una persona puede levantar podría anotarse en kilogramos, o en gramos, o ambos, y así sucesivamente. La distancia, la fuerza y el tiempo son variables continuas. Dependiendo de la precisión del instrumento de medición, la distancia puede ser medida tan corta como una millonésima de un centímetro, o tan amplia como un año luz. El tiempo puede ser medido en milisegundos, en días o en siglos.

La mayoría de los puntajes continuos se redondean a la unidad de medida más cercana al anotarlo. Por ejemplo, el puntaje de un estudiante que corre 100 metros en 10,57 segundos se registra como 10,6 porque 10,57 está más cerca de 10,6 que de 10,5. Por lo general, cuando un número se redondea a la unidad de medida más cercana, se incrementa solo cuando el número que se elimina es 5 o más. Por lo tanto, 11,45 se redondea a 11,5, mientras que 11,44 se registra como 11,4. Un método menos común es redondear a la última unidad de medida, otorgando el siguiente puntaje más alto solo cuando ese puntaje se haya logrado. Por ejemplo, una persona que levanta un peso 8 veces, pero no puede completar el noveno levantamiento recibe una puntuación de 8. Asimismo, si un estudiante hace la prueba PACER y lleva 49 vueltas y se detiene a mitad del recorrido sin alcanzar la vuelta 50, se le anotan 49 vueltas.

Datos discretos. Los datos discretos están limitados a un número específico de valores y regularmente no se pueden expresar en fracciones. Los datos discretos se limitan a ciertos números, por lo general enteros. Acertar una serie de lanzamientos o de tiro al blanco con una escala de 5-4-3-2-1-0 arroja datos discretos, porque un intento sólo puede recibir un puntaje de 5, 4, 3, 2, 1 o 0; un puntaje de 4.7 o de 1.67 es imposible. Cuando se cuenta gente, siempre se utilizan datos enteros porque es imposible fraccionar una persona. El número de pulsaciones en un minuto o el número de cestas logradas en baloncesto también son ejemplos de datos discretos. Uno no puede contar la mitad de una pulsación o dividir una canasta convertida.

Los datos discretos y continuos se expresan en forma cuantitativa. Los datos cuantitativos consisten en números que representan conteos o mediciones, como las tallas de 120 o 125 centímetros; se debe tener cuidado y señalar explícitamente la unidad de medida utilizada refiriendo, por ejemplo, “todos los tiempos están en centésimas de segundo”, “el peso se expresa en kilogramos”. Los datos cualitativos se enuncian por medio de categorías o atributos distinguiéndose por características no numéricas, como los colores de los ojos negros y azules, o los cabellos lisos y rizados.

Escalas. La clasificación de los datos se puede hacer de otra forma, a través de escalas o niveles de medición. Estas escalas determinan el procedimiento estadístico a utilizar, ya que es inútil efectuar cálculos inapropiados si el tipo de datos recolectado no se ajusta a los requerimientos exigidos. Por ejemplo, ¿qué utilidad tendría promediar los números de los jugadores de un equipo de béisbol? Al efectuar una medición se asigna un número para representar esa medición, se trate del peso, la altura, la distancia o el tiempo. Los datos se clasifican en cuatro escalas o niveles de medición: nominal, ordinal, intervalo y razón.

Escala nominal. En la escala nominal se clasifican los datos por su nombre, como los jugadores de fútbol de campo por su posición: porteros, defensas, o los atletas por el género masculino y femenino. En la escala nominal se utilizan las categorías para identificar y diferenciar algún elemento o característica de los datos, por lo que no hay noción de orden, magnitud o tamaño, y los datos no se pueden utilizar para efectuar cálculos.

Escala ordinal. En la escala ordinal los datos se ubican por un orden en concordancia con las condiciones previamente establecidas, tal como el orden de llegada en una carrera atlética, el orden de la clasificación en una competencia ciclística después de varias etapas, el orden jerárquico establecido en algunos deportes como el tenis de campo, o la ubicación de los equipos por juegos ganados y perdidos. En este tipo de escala no es posible determinar las diferencias entre los valores de los datos o tales diferencias carecen de significado ya que, por ejemplo, la diferencia entre el primero y el segundo puede ser muy pequeña, y entre el tercero y el cuarto muy grande.

Escala de intervalos. La escala de intervalos tiene unidades de medición comunes o iguales, es decir, intervalos que representan la misma distancia entre ellos en una escala, sólo que no tienen un puntaje cero (0) real. Hay la misma distancia entre los grados de calor medidos en centígrados, pero un grado 0 de calor no indica la falta de calor sino el punto en el que el agua se congela. Al considerar un salto largo medido en centímetros un puntaje de 0 indica que no hubo una distancia cualquiera alcanzada, por lo que en esta prueba el 0 es real. Al contrario, en una prueba de conocimiento un puntaje de 0 puntos no indica la ausencia total de conocimientos, sino que no hubo respuestas correctas. En el deporte, la escala de intervalos se utiliza para juzgar el desempeño en gimnasia, nado sincronizado o salto del trampolín; si un gimnasta obtiene 9 puntos no significa que es dos veces mejor que aquel que obtuvo 4.5 puntos. Un puntaje de 0 no significa la falta de destreza, sino que la demostración se realizó sin la suficiente destreza como para esta sea premiada con un alto puntaje. En algunos casos, pueden presentarse datos negativos como en la medición de temperatura al señalarse -18 grados, o al expresar negativamente los años antes de Cristo.

Escala de razón. La escala de razón es la más completa escala de medición y está en el más alto nivel de la clasificación de los datos por escala. La escala de razón posee un orden, igual distancia entre los posibles puntajes, y utiliza el 0 para representar la ausencia de valor. El 0 está presente, por ejemplo, en la medición del peso de las personas, en la distancia de un lanzamiento, o en el tiempo de una carrera. Las unidades son equidistantes unas de otras por lo que pueden ser comparadas entre sí ya que el punto de partida 0 hace que las razones o cocientes tengan significado; todas las mediciones de distancia, fuerza y tiempo se basan en escalas de razón. Si una persona mide 1.80 metros y otra mide 0.90, se puede decir que la primera persona es dos veces más alta que la segunda. Los números negativos no son posibles en la escala de razón. Nadie puede correr en segundos negativos, pesar menos de 0 kilogramos, anotar puntos negativos en baloncesto deporte, ni batear un promedio menor de 0.

ESCALA	RESUMEN
Nominal	Emplea categorías. Los datos no pueden arreglarse de acuerdo con un orden.
Ordinal	Las categorías están ordenadas pero no hay diferencias entre ellas o carecen de significado.
De intervalo	Las diferencias tienen un significado pero no hay un puntaje inicial cero natural y los cocientes no tienen significado.
De razón	Hay un punto de partida 0 natural y los cocientes tienen significado.
Tabla 2.1 Comparación entre las escalas de medición	

Un concepto importante de tener presente es que algunas características de los datos deben verificarse antes de proceder a efectuar las respectivas operaciones matemáticas. Los datos deben ser observados dentro de un continuo que va de la escala nominal a la ordinal a la de intervalo a la de razón. Las escalas de medición son jerárquicas y cada una tiene en sí el nivel o los niveles previos. Así, un número que sea ordinal también es nominal; si un dato está en la escala de intervalos también contiene información de las escalas ordinal y nominal, y si es un dato en la escala de razón contiene información de las demás escalas. Asimismo, sólo los datos de las escalas de intervalo y de razón pueden ser procesados matemáticamente, siendo inapropiado, por ejemplo, hallar el promedio de un conjunto de datos ubicados en una escala ordinal.

Notación

Para representar lo que se quiere lograr con los procedimientos matemáticos, se ha desarrollado un sistema que resume la expresión y procesamiento de los datos, sistema denominado notación. Aunque la notación puede ser muy compleja, es necesario conocer algunos con-

ceptos básicos. En primer lugar, un dato individual se representa con la letra x , que corresponde al registro de la variable observada. Si un sujeto obtiene una marca de 7,78 metros en salto largo, entonces $x = 7,78$, y ello es valedero para cualquier medición en el área de la actividad física y en variables como altura, peso, distancia, tiempo, frecuencia. Σ es la letra griega sigma que indica *la suma de*, y Σx quiere decir $x_1 + x_2 + x_3 + \dots + x_n$, y se lee como *la suma de todos los datos*. N representa el número de datos o sujetos en una población, mientras n es el número de datos o sujetos en una muestra. Σx^2 se lee como la suma de los cuadrados de los datos. Por ejemplo, si tenemos los datos de una prueba de velocidad en 50 metros de cinco sujetos, cada dato se multiplica por sí mismo y luego se suman todos los puntajes elevados al cuadrado. En este caso Σx es 39,66 y Σx^2 es 314,6928.

n	x	x^2
1	7,71	59,4441
2	8,15	66,4225
3	7,85	61,6225
4	7,94	63,0436
5	8,01	64,1601
Σ	39,66	314,6928

Tabla 2.2 Suma de los puntajes y suma de los cuadrados de los valores

Ahora bien, se debe distinguir entre la suma de los cuadrados de los datos y el cuadrado de la suma de todos los datos o puntajes. Este se denota como $(\Sigma x)^2$ lo que indica que la suma de los puntajes se eleva al cuadrado; en el ejemplo la suma de los puntajes es 39,66 por lo que $(\Sigma x)^2$ es $(39,66)^2 = 1572,9156$, es decir, el cuadrado de la suma de todos los datos. La media se representa mediante la letra M para los valores muestrales, así como con la letra x con una barra arriba de ella, mientras que la media de los valores de una población se representa con el símbolo μ (la letra griega mu). En algunos casos, los estadísticos productos de muestras se representan con letras latinas, mientras que los parámetros poblacionales con letras griegas. Recuérdese que, si en las operaciones existen paréntesis, se procede a resolver primero los exponentes, luego las multiplicaciones y divisiones y, por último, la suma y las restas. Otros símbolos de notación se verán adelante. En resumen, algunos de los símbolos empleados en la notación estadística son:

Símbolo	Significado
Σ	Suma
x	Representa cada dato individual
n	Representa la cantidad total de personas o datos incluidos en la prueba, cuando se trata de muestras
N	Representa la cantidad total de personas o datos incluidos en una prueba, cuando se trata de poblaciones
Σx	Representa la suma de los datos
Σx^2	Suma de los cuadrados de los datos
$(\Sigma x)^2$	Cuadrado de la suma de todos los datos

$\mathcal{M} = \frac{\sum x}{n}$	Representa la media de los datos muestrales
$\mu = \frac{\sum x}{N}$	Representa la media de los datos de una población
Tabla 2.3 Algunos de los símbolos de la notación estadística	

Organización de los datos

Una vez que se ha aplicado una prueba, los datos recolectados poseen una gran cantidad de información de un valor potencial. Antes de que se pueda generar una información valiosa, se deben organizar los datos en forma lógica, útil y comprensible, a través de un formato denominado tabla de distribución de frecuencias, que se puede presentar de tres maneras: la distribución de orden jerárquico, la distribución simple de frecuencias y la distribución agrupada de frecuencias.

Una distribución de frecuencias se construye por varias razones. En primer lugar, ofrecen la posibilidad de resumir grandes conjuntos de datos; en segundo lugar, se logra la comprensión de la naturaleza de los datos al observarlos en forma global, y, en tercer lugar, se sientan las bases para construir gráficas acerca de los mismos datos. Una distribución de frecuencias o tabla de frecuencias organiza los datos en forma de lista, ya sea de manera individual o por grupos de intervalos, junto con las frecuencias correspondientes a los datos.

Distribución de orden jerárquico. Una distribución de orden jerárquico es una lista ordenada de los datos en una columna. Ofrece una rápida visión de la distancia o variabilidad de los datos y se identifican fácilmente los puntajes extremos, de los más altos a los más bajos. Una distribución de orden jerárquico es útil cuando la cantidad de datos es poca, no mayor de 20. Al ordenárseles en una lista los datos deben mostrarse en una página, lo cual hace apropiada esta distribución.

Al tomar una prueba como ejemplo, tenemos la ejecución de la prueba de flexiones de codo en suspensión por parte de 15 estudiantes, con los siguientes datos originales: 12, 10, 9, 8, 2, 5, 18, 15, 14, 17, 13, 12, 8, 9, 16 (Tabla 2.4)

n	x	x
1	12	2
2	10	5
3	9	8
4	8	8
5	2	9
6	5	9
7	18	10
8	15	12
9	14	12
10	17	13
11	13	14
12	12	15

13	8	16
14	9	17
15	16	18
Tabla 2.4 Datos originales		Tabla 2.5 Distribución de orden jerárquico

La observación de estos datos puede resultar confusa y difícil de interpretar por su falta de organización. Al ser colocados de menor a mayor es más fácil su interpretación por parte del docente o entrenador (Tabla 2.5). Recuérdese que el símbolo x representa la variable de flexiones, lo que logró cada sujeto, y n el número de participantes. Veamos la tabla 2.4 de datos originales y la tabla 2.5 de distribución de orden jerárquico.

Cuando los datos originales se organizan en una tabla de distribución de orden jerárquico, es más fácil percibir la diferencia entre el mejor y el inferior que alcanza un rango de 16 flexiones, bastante amplio. También se observa una gran diferencia en la capacidad para realizar las flexiones y que hay varios pares de sujetos con la misma capacidad, los que realizaron 8, 9 y 12 repeticiones.

Distribución simple de frecuencias. Si el número de datos que se procesa es demasiado amplio, es inconveniente elaborar una lista continua porque sería demasiado larga y sobrepasaría el tamaño de la hoja, con todas las dificultades que ello acarrea. Esa gran cantidad de datos se organiza en una distribución simple de frecuencias, ordenando la variable que se estudia (x) en una columna y en otra columna la frecuencia (f) que indica en número de datos que corresponde al valor x . Supongamos que tenemos los resultados de una prueba de salto largo sin impulso con los siguientes 50 datos (Tabla 2.6):

n	x	n	x	n	x	n	x	n	x
1	168	11	170	21	137	31	160	41	229
2	142	12	142	22	165	32	180	42	208
3	173	13	173	23	193	33	138	43	198
4	119	14	147	24	173	34	198	44	193
5	117	15	173	25	173	35	229	45	158
6	147	16	125	26	158	36	213	46	191
7	191	17	165	27	168	37	183	47	185
8	180	18	191	28	211	38	211	48	163
9	152	19	193	29	165	39	201	49	142
10	173	20	178	30	122	40	196	50	150
Tabla 2.6 Datos originales de la prueba de salto largo sin impulso, medida en centímetros									

Para construir la tabla de distribución simple de frecuencias, se trazan las columnas que contienen los valores ordenados de x con el valor más bajo en la celda superior, así como las frecuencias (f) que corresponden al número de sujetos que lograron ese valor. En las tablas 2.7 y 2.9 se incluyen dos columnas para la interpretación de los resultados: la columna del porcentaje (%) correspondiente a cada valor, y la columna del % acumulado que se inicia desde el primer puntaje (que es el inferior o menor) y se va acumulando hasta el mayor o mejor.

<i>x</i>	<i>f</i>	%	% acumulado
117	1	2	2
119	1	2	4
122	1	2	6
125	1	2	8
137	1	2	10
138	1	2	12
142	3	6	18
147	2	4	22
150	1	2	24
152	1	2	26
158	2	4	30
160	1	2	32
163	1	2	34
165	3	6	40
168	2	4	44
170	1	2	46
173	6	12	58
178	1	2	60
180	2	4	64
183	1	2	66
185	1	2	68
191	3	6	74
193	3	6	80
196	1	2	82
198	2	4	86
201	1	2	88
208	1	2	90
211	2	4	94
213	1	2	96
229	2	4	100

Tabla 2.7 Distribución simple de frecuencias de la prueba de salto largo sin impulso

<i>n</i>	<i>x</i>	<i>n</i>	<i>x</i>	<i>n</i>	<i>x</i>	<i>n</i>	<i>x</i>	<i>n</i>	<i>x</i>
1	7,75	11	7,90	21	7,95	31	8,57	41	8,00
2	7,80	12	7,27	22	7,40	32	8,25	42	7,40
3	7,78	13	8,36	23	8,60	33	7,41	43	7,51
4	7,14	14	7,31	24	7,27	34	7,21	44	7,16
5	7,97	15	7,20	25	7,75	35	8,44	45	7,83
6	7,09	16	7,85	26	8,18	36	8,00	46	7,94
7	7,94	17	8,02	27	7,34	37	8,19	47	7,75

8	8,12	18	7,34	28	7,75	38	7,99	48	7,84
9	8,91	19	7,48	29	8,96	39	7,81	49	7,14
10	8,00	20	7,84	30	7,94	40	7,61	50	8,36

Tabla 2.8 Datos originales de la prueba de 50 metros en velocidad, en segundos y centésimas

Así, en la tabla 2.7 el puntaje de 173 centímetros fue alcanzado por 6 sujetos que representan el 12% de todos los puntajes mientras que el 58% de los sujetos alcanzaron 173 centímetros o menos.

En el caso de una prueba de velocidad cuyos datos originales o crudos están en la tabla 2.8, en la tabla 2.9 de distribución de frecuencias van de mejor (menor) a inferior (mayor) y se observa que la separación entre los dos puntajes extremos, el menor y el mayor, es amplia por lo que es preferible construir una distribución agrupada de frecuencias.

<i>x</i>	<i>f</i>	%	% acumulado
7,09	1	2	2
7,14	2	4	6
7,16	1	2	8
7,20	1	2	10
7,21	1	2	12
7,27	2	4	16
7,31	1	2	18
7,34	2	4	22
7,40	2	4	26
7,41	1	2	28
7,48	1	2	30
7,51	1	2	32
7,61	1	2	34
7,75	4	8	42
7,78	1	2	44
7,80	1	2	46
7,81	1	2	48
7,83	1	2	50
7,84	2	4	54
7,85	1	2	56
7,90	1	2	58
7,94	3	6	64
7,95	1	2	66
7,97	1	2	68
7,99	1	2	70
8,00	3	6	76
8,02	1	2	78
8,12	1	2	80
8,18	1	2	82
8,19	1	2	84
8,25	1	2	86
8,36	2	4	90
8,44	1	2	92
8,57	1	2	94
8,60	1	2	96
8,91	1	2	98

8,96	1	2	100
------	---	---	-----

Tabla 2.9 Distribución simple de frecuencias de la prueba de velocidad

Distribución agrupada de frecuencias. Cuando se trabaja con una gran cantidad de datos, y esto es lo que sucede con los docentes y entrenadores, no es viable elaborar una lista continua con todos los valores, por lo cual es necesario agruparlos para reducir la longitud de esa lista.

En una distribución agrupada de frecuencias los puntajes originales x se organizan en una columna en forma agrupada, con una columna contigua donde se anota la cantidad de datos incluidos en cada grupo. Para conformar los grupos primero se debe decidir cuántos valores se incluirán en cada uno. Por lo general, se ha establecido que 15 es la cantidad ideal de grupos para cualquier tipo de prueba, aunque realmente puede variar entre 10 y 20 grupos, pero ello es una decisión arbitraria. Si la cantidad de grupos es menor de 10 los datos en cada grupo son demasiados, y si la cantidad de grupos es mayor de 20 los datos en cada grupo son pocos y la tabla tiende a alargarse demasiado.

Para establecer el tamaño del intervalo o número de valores para cada grupo, se determina en primer lugar el rango de los datos precisando el mejor puntaje y el puntaje inferior de la totalidad de datos, sustrayéndolos y dividiendo el resultado entre 15 (cantidad de valores o posibilidades que tienen los sujetos de obtener un puntaje en cada grupo).

$$i = \frac{\text{Rango}}{15}$$

En el caso de la prueba de salto largo sin impulso en la tabla 2.7, el mejor puntaje fue de 229 centímetros y el puntaje más bajo o inferior fue de 117, por lo que se calcula el intervalo de la siguiente manera:

$$i = \frac{229 - 117}{15} = \frac{112}{15} = 7,5 = 8$$

La agrupación de los datos es más fácil si el intervalo es un número entero por lo que se procede a redondear el resultado, en este caso, a un intervalo de 8 que será el número de valores que se incluyan en cada grupo. Por otro lado, se incluye una columna para registrar la cantidad de datos y luego expresarlos en números en la columna de frecuencias.

Una vez obtenido el intervalo de 8, se organizan los grupos empezando por el puntaje menor, 117, para conformar el primer grupo con los puntajes entre 117 y 124, dentro del cual hay 3 sujetos ya que ellos saltaron de manera tal que no superaron los 124 centímetros. Seguidamente se conforman los demás grupos de la misma manera, hasta que el puntaje más alto, 229 centímetros, se incluya en un grupo. Así, tenemos que en último grupo se encuentran dos sujetos y ambos saltaron 229 centímetros (Tabla 2.10).

Grupo	<i>r</i>	<i>f</i>
117-124	///	3
125-132	/	1
133-140	//	2
141-148	+++	5
149-156	//	2
157-164	////	4
165-172	+++ /	6
173-180	+++ ////	9
181-188	//	2
189-196	+++ //	7
197-204	///	3
205-212	///	3
213-220	/	1
221-228		0
229-236	//	2
	Σ	50
<i>i</i> = 8		
Tabla 2.10 Distribución agrupada de frecuencias de la prueba de salto largo sin impulso		

Se debe tomar en cuenta que, una vez que se haya determinado el intervalo o anchura y se vayan a crear los grupos, si el intervalo es un múltiplo de 5 el grupo menor debe incluir el dato más bajo y el límite inferior debe ser un múltiplo de 5. Por ejemplo, si en una prueba el puntaje más bajo es 13 y el intervalo es 15, el grupo comienza con 10 y estaría conformado de 10 a 24, el segundo grupo sería 25 – 39, y así sucesivamente. Si el intervalo no es múltiplo de 5 el grupo se comienza a conformar con el puntaje más bajo. En ambos casos, los grupos se conforman hasta que alcance al puntaje del otro extremo.

En el caso de las pruebas como velocidad y agilidad el intervalo se establece en decimales, por lo general en centésimas de segundo. En la prueba de capacidad aeróbica de correr una milla, el tiempo se registra en minutos y segundos, y luego se llevan a segundos. En estos casos la tabla se organiza igualmente de menor a mayor puntaje y el menor puntaje es el mejor por tratarse de una prueba de tiempo. Considerando los datos de la tabla 2.9 se calcula el intervalo de la siguiente manera:

$$i = \frac{8,96 - 7,09}{15} = \frac{1,87}{15} = 0,1246 = 0,13$$

Con el intervalo de 0,13 se construye la tabla de distribución de frecuencias como se observa a continuación en la tabla 2.11. Al empezar la tabla con el menor puntaje que en esta prueba es el mejor (7,09) y utilizar el intervalo de 0,13, el primer grupo estaría entre 7,09 y 7,21 segundos con la cantidad de 6 sujetos que lograron un tiempo entre ambos valores, y así sucesivamente hasta el grupo donde se encuentre el puntaje más alto (8,96).

Grupo	<i>r</i>	<i>f</i>
7,09 – 7,21	//// /	6
7,22 – 7,34	////	5
7,35 – 7,47	///	3
7,48 – 7,60	//	2
7,61 – 7,73	/	1
7,74 – 7,86	//// //// /	11
7,87 – 7,99	//// //	7
8,00 – 8,12	////	5
8,13 – 8,25	///	3
8,26 – 8,38	//	2
8,39 – 8,51	/	1
8,52 – 8,64	//	2
8,65 – 8,77		0
8,78 – 8,90		0
8,91 – 9,03	//	2
Σ		50
<i>i</i> = 0.13		
Tabla 2.11 Distribución agrupada de frecuencias de la prueba de velocidad		

Límites y punto medio de los grupos. Cada grupo que se haya conformado presenta unos límites aparentes, unos límites reales y un punto medio. Si consideramos la tabla 2.10 y tomando el primer grupo, 117 – 124, tenemos que estos dos puntajes extremos son el límite aparente inferior y el límite aparente superior, respectivamente. El límite real inferior de este grupo es 116,5 y el límite real superior es 124,49 ya que 124,5 es el límite real inferior de la siguiente clase. Recuérdese que estos datos han sido registrados en centímetros; si la distancia se hubiese registrado en metros y centímetros indudablemente que los límites reales se tendrían que establecer en milímetros. La primera clase en metros sería 1,17 – 1,24 y su límite real inferior sería 1,165 y el límite real superior 1,2449 por lo cual es preferible registrar las marcas en números enteros como se presenta en la prueba de salto sin impulso ya que, en todo caso, resulta perfectamente claro para los participantes que se les exprese que saltaron un metro con 17 centímetros o 117 centímetros.

Este concepto es importante cuando los datos originales contienen decimales, especialmente si la prueba es medida en segundos y milésimas de segundo. En este caso la prueba se registra en segundos y centésimas de segundos, mientras los límites reales se establecen en milésimas de segundo. Así sucede en la prueba de velocidad de la tabla 2.11. En el primer

grupo los límites aparentes son sus puntajes extremos, 7,09 y 7,21; el límite real inferior es 7,085 y 7,2149 es el límite superior. Esto se ha establecido de esta manera para ubicar con claridad a un sujeto dentro de un grupo cuando ha realizado una marca, por ejemplo, de 7,213, lo que permite discernir que está dentro de este grupo y no del siguiente.

El punto medio del grupo se determina al promediar los dos puntajes extremos. Siguiendo con los ejemplos anteriores, tenemos el punto medio del grupo 117 – 124 de la prueba de salto sin impulso y del grupo 7,09 – 7,21 de la prueba de velocidad:

$$PM = \frac{117 + 124}{2} = 120,5$$

$$PM = \frac{7,09 + 7,21}{2} = 7,15$$

La determinación del punto medio orienta la elaboración de los gráficos, como se verá en el próximo punto. La tabla 2.12 refleja el límite real inferior, el límite aparente inferior, el punto medio, el límite aparente superior y el límite real superior de la prueba de salto largo sin impulso.

Límite real inferior	x Límite aparente inferior	Punto medio	x Límite aparente superior	Límite real superior
116,5	117	120,5	124	124,49
124,5	125	128,5	132	132,49
132,5	133	135,5	140	140,49
140,5	141	144,5	148	148,49
148,5	149	152,5	156	156,49
156,5	157	160,5	164	164,49
164,5	165	168,5	172	172,49
172,5	173	176,5	180	180,49
180,5	181	184,5	188	188,49
188,5	189	192,5	196	196,49
196,5	197	200,5	204	204,49
204,5	205	208,5	212	212,49
212,5	213	216,5	220	220,49
220,5	221	224,5	228	228,49
228,5	229	232,5	236	236,49

Tabla 2.12 Límites y punto medio de las pruebas de salto largo sin impulso

La tabla 2.13 refleja el límite real inferior, el límite aparente inferior, el punto medio, el límite aparente superior y el límite real superior de la prueba de velocidad.

Límite real inferior	X Límite aparente inferior	Punto medio	X Límite aparente superior	Límite real superior
7,085	7,09	7,15	7,21	7,2149
7,215	7,22	7,28	7,34	7,3449
7,345	7,35	7,41	7,47	7,4749
7,475	7,48	7,54	7,60	7,6049
7,605	7,61	7,67	7,73	7,7349
7,735	7,74	7,80	7,86	7,8649
7,865	7,87	7,93	7,99	7,9949
7,995	8,00	8,06	8,12	8,1249
8,125	8,13	8,19	8,25	8,2549
8,255	8,26	8,32	8,38	8,3849
8,385	8,39	8,45	8,51	8,5149
8,515	8,52	8,58	8,64	8,6449
8,645	8,65	8,71	8,77	8,7749
8,775	8,78	8,84	8,90	8,9049
8,905	8,91	8,97	9,03	9,0349

Tabla 2.13 Límites y punto medio de las pruebas de velocidad

Distribución de los datos en gráficos

Los gráficos constituyen una manera útil de presentar la información que generan los datos. Son representaciones visuales de los datos que ofrecen una visión que probablemente no se pueda observar fácilmente en una tabla de distribución de frecuencias o en una hoja de cálculo de algún programa informático, por lo que constituyen una versión gráfica de la tabla. Ya que hay diferentes tipos de puntajes, los gráficos facilitan la visión de los datos al agruparlos ya que adquieren una forma general que denota su distribución de acuerdo con los resultados, permitiendo conocer la naturaleza de la distribución de un conjunto de datos grande.

Por lo general, todo gráfico se organiza en dos ejes, el eje horizontal X o eje de las abscisas, y el eje vertical Y o eje de las ordenadas. En el eje de las ordenadas se colocan las frecuencias mientras que en el eje de las abscisas se colocan los puntajes, de manera tal que se cruzan en el gráfico el puntaje con el número de sujetos (frecuencia) que alcancen ese puntaje, señalándose con un punto el lugar donde se produce ese cruce. En el eje de las abscisas sólo se señala el punto medio de cada grupo; en este eje los puntajes van de izquierda a derecha, así como los valores van del puntaje inferior al mejor. Debe recordarse, sin embargo, que, en algunas pruebas como las pruebas de velocidad y destreza, el mejor puntaje no es el más alto sino el más bajo. A continuación, se presentan dos tipos de gráficos, el histograma y el polígono de frecuencias; aunque hay mayor variedad de gráficos, en ellos se visualiza la presentación de los datos y cómo se distribuyen de acuerdo con la tabla de distribución de frecuencias. En los programas informáticos como Excel y SPSS hay una gran variedad de

gráficos que se pueden seleccionar de acuerdo con el objetivo planteado y el estilo que se le desee dar.

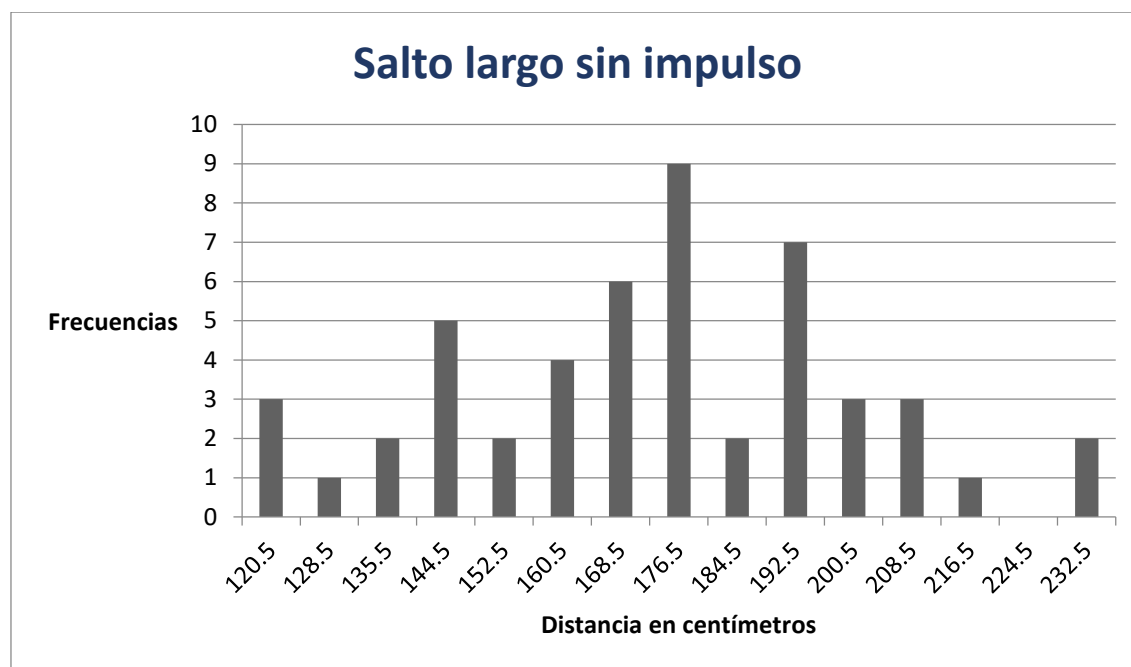


Figura 2.1 Histograma del salto largo sin impulso

Histograma. El histograma es un tipo de gráfico muy utilizado, también denominado gráfico de barras. Las barras sobre el eje de las abscisas representan los grupos, y la altura de cada barra la determina la frecuencia o número de sujetos en cada grupo, ubicados en el eje de las ordenadas.

Veamos los datos de las pruebas ya presentadas con anterioridad, que se distribuyen en la figura 2.1. Se observa que los puntajes pueden variar entre 120 y 232 centímetros aproximadamente, el centro está en 176 centímetros, y que hacia el extremo izquierdo los primeros cuatro grupos reúnen 11 sujetos, hacia el extremo derecho los últimos cuatro grupos contienen 6 sujetos, mientras que en los siete grupos centrales hay 33 sujetos, lo cual significa que la mayoría se halla alrededor del grupo central.

En la figura 2.2 los datos se han invertido para observar hacia la derecha las mejores marcas; los puntajes se distribuyen entre 8,97 y 7,15 aproximadamente, de acuerdo con el punto medio del grupo que origina la barra correspondiente. Los primeros cuatro grupos de la izquierda tienen apenas 4 sujetos, y los últimos cuatro grupos de la derecha registran 16 sujetos. Los siete grupos centrales presentan 30 sujetos, observándose que los datos se encuentran ligeramente cargados hacia la derecha.

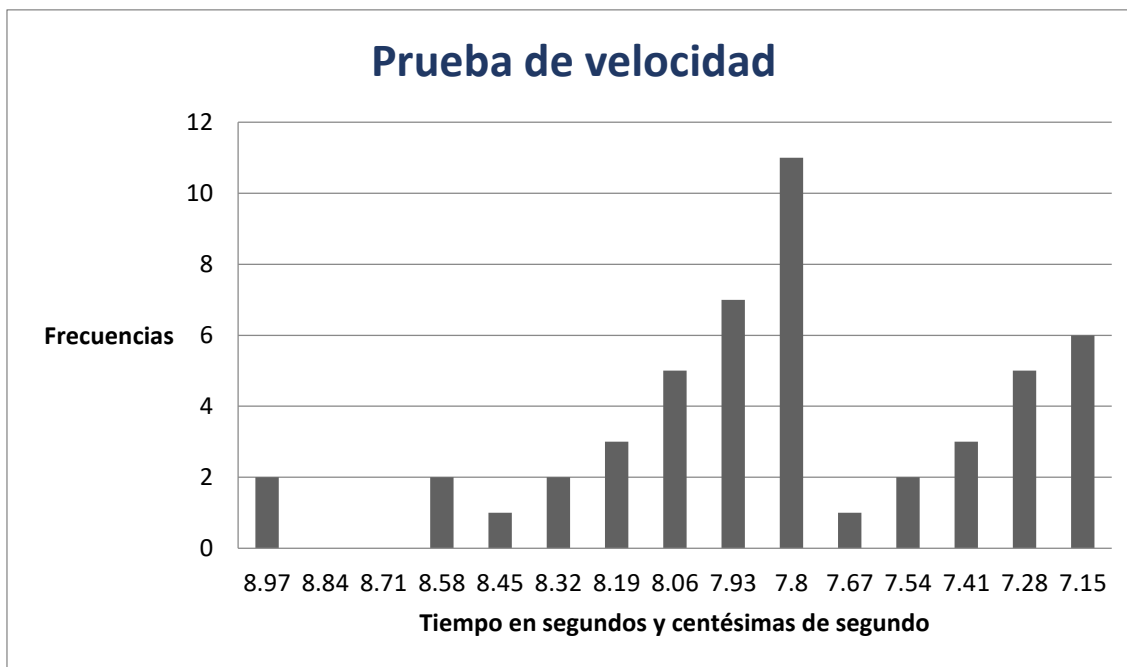


Figura 2.2 Histograma de la prueba de velocidad

Polígono de frecuencias. Otro tipo de gráfico que se puede construir a partir de la tabla de distribución de frecuencias es el polígono de frecuencias, que es una línea gráfica de los puntajes señalados en el eje X en coincidencia con las frecuencias del eje Y. Se conectan los puntos localizados por arriba de los puntos medios en coincidencia con la frecuencia correspondiente a cada grupo, conexión que se hace a través de segmentos lineales. Los datos y las frecuencias se ubican de la manera ya descrita anteriormente. Obsérvense los polígonos de frecuencia de las pruebas señaladas en las figuras 2.3 y 2.4.



Figura 2.3 Polígono de frecuencias de la prueba de salto sin impulso

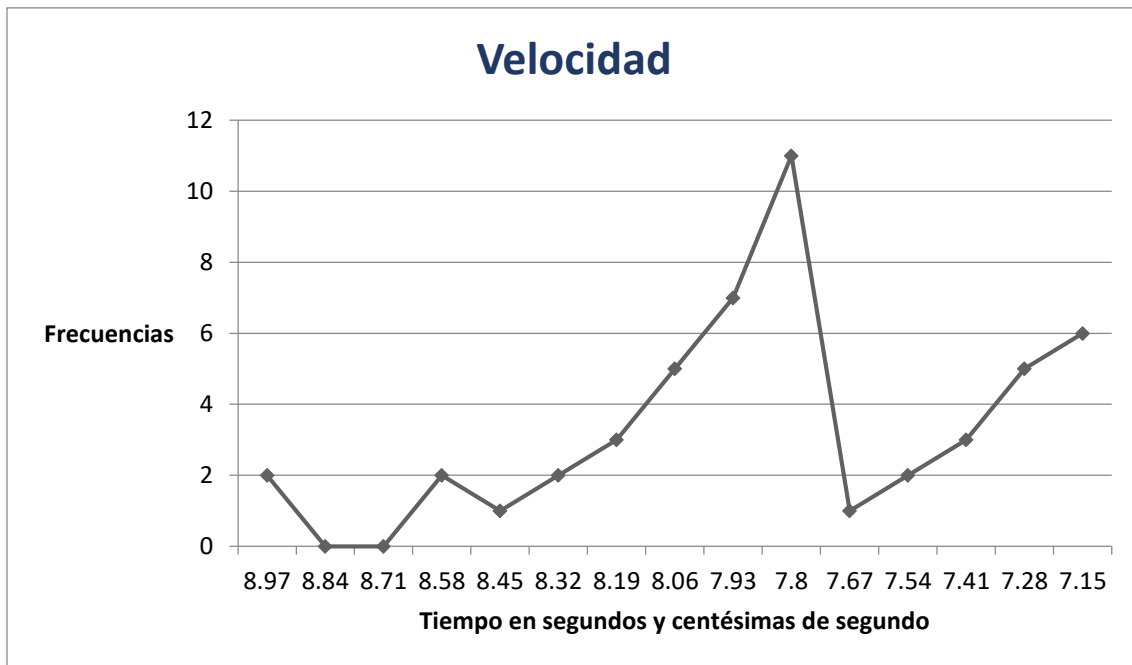


Figura 2.4 Polígono de frecuencias de la prueba de velocidad

Gráfica de puntos. Muchos conjuntos de datos cuantitativos están formados de números que no se pueden separar fácilmente en categorías o intervalos. Entonces se hace necesaria una forma diferente de graficar este tipo de datos. La gráfica más sencilla para datos cuantitativos es la gráfica de puntos. Para un conjunto pequeño de mediciones, por ejemplo, el conjunto 2, 6, 9, 3, 7, 6, se puede simplemente graficar las mediciones como puntos en un eje horizontal. Esta gráfica de puntos, generada por el programa Minitab, se muestra en la figura 2.5. Para un conjunto grande de datos, como el de la figura 2.6, la gráfica de puntos puede ser poco informativa y tediosa de interpretar.

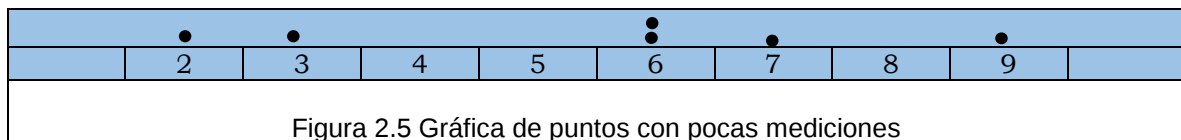


Figura 2.5 Gráfica de puntos con pocas mediciones

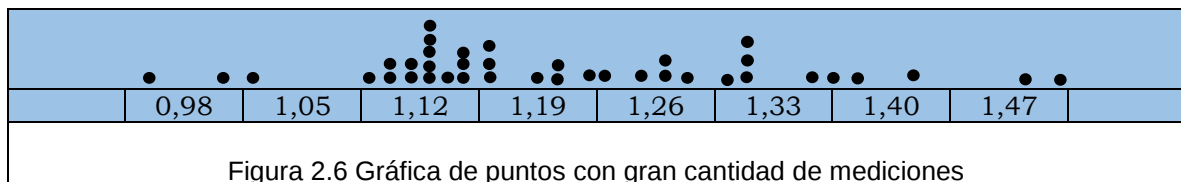


Figura 2.6 Gráfica de puntos con gran cantidad de mediciones

Distribución normal

Cuando se grafica una distribución de frecuencias se forma una línea curva que ofrece una perspectiva visual de la tendencia de los datos y que no son fácilmente observables en la tabla de distribución de frecuencias. En estadística, la curva más conocida es la curva normal o curva en forma de campana, que se caracteriza por una distribución simétrica de los datos a

partir del centro hacia los extremos, por lo que la frecuencia de los datos es predecible por cuanto declinan cada vez más en la medida en que se alejan del centro. En la tabla 2.14 tenemos los datos de la estatura de 1000 mujeres y en la figura 2.7 el histograma correspondiente.

Distribución normal

Al inicio las frecuencias son bajas, luego alcanzan un nivel máximo y disminuyen nuevamente

Estatura (en pulgadas)	Frecuencia	Distribución normal
56,0 – 57,9	10	← Al inicio las frecuencias son bajas
58,0 – 59,9	64	
60,0 – 61,9	178	
62,0 – 63,9	324	← Aumentan hasta un máximo
64,0 – 65,9	251	
66,0 – 67,9	135	
68,0 – 69,9	32	← Disminuyen nuevamente
70,0 – 71,9	6	

Tabla 2.14 Distribución de frecuencias de la estatura de 1000 mujeres

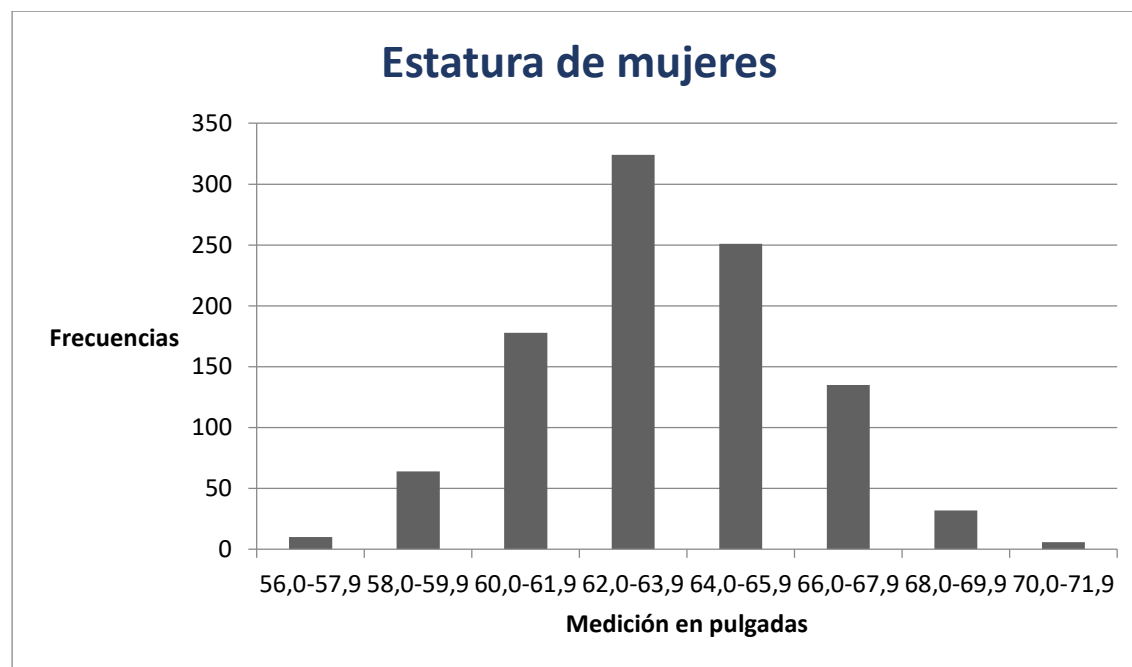
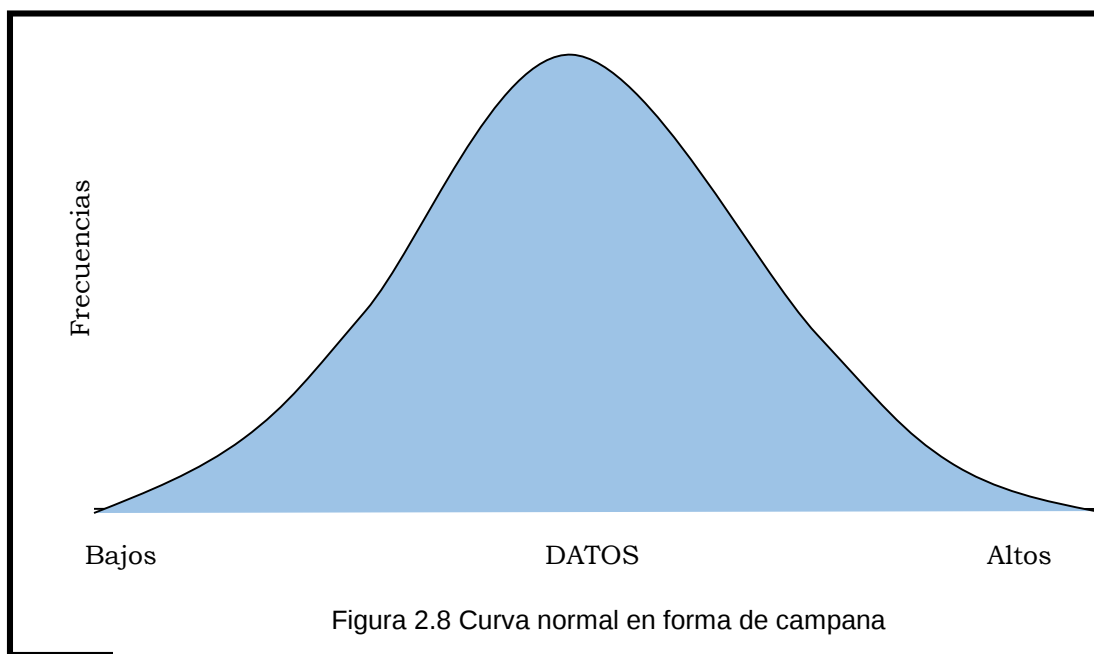
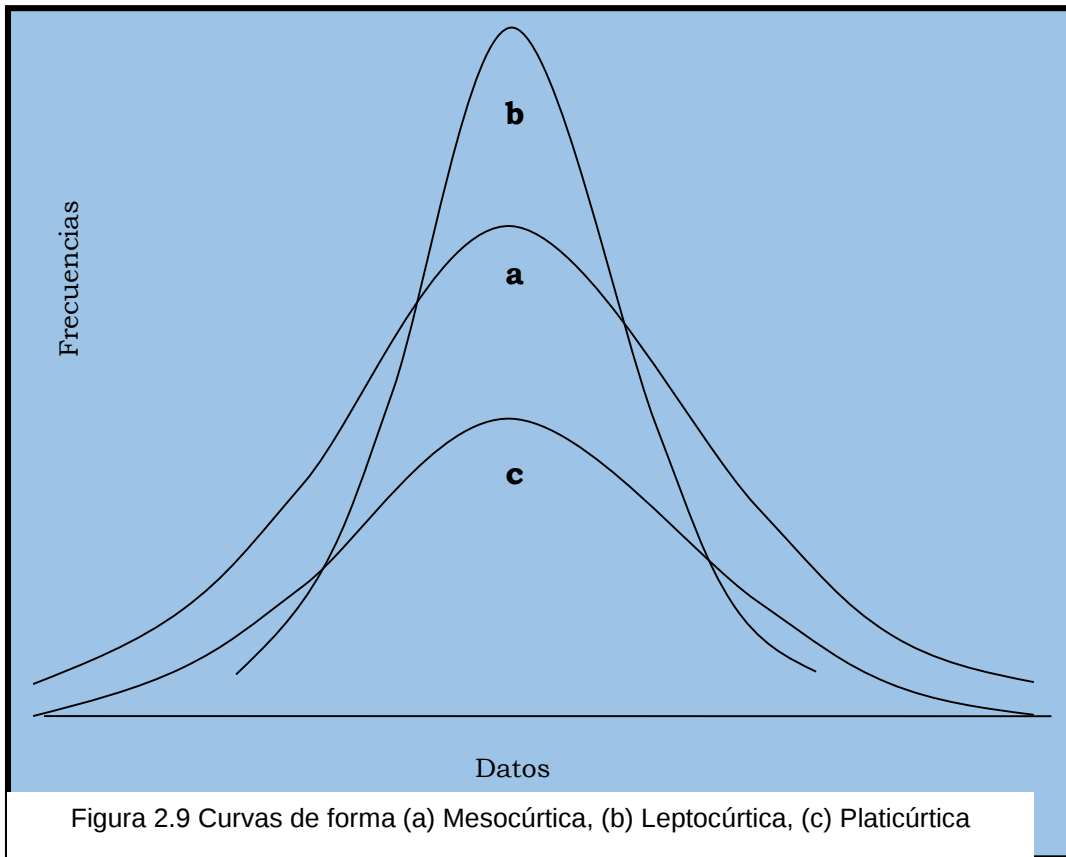


Figura 2.7 Histograma de la estatura de 1000 mujeres

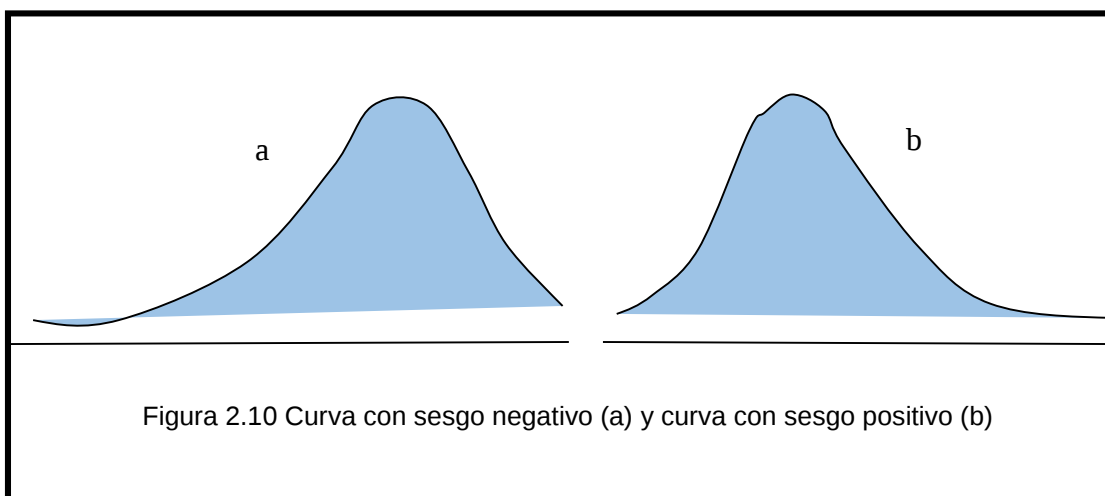
Se deduce del histograma con los datos de la estatura de las mujeres que se trata de una distribución normal con forma de campana como se observa en la figura 2.8. Nótese que los lados finales o colas de la curva son simétricos, representando los extremos bajos y altos de la escala, donde la cantidad de datos declina ostensiblemente.



La distribución normal puede adquirir diversas formas. Si la mayoría de los datos se encuentran el centro y se alejan simétricamente hacia las colas, la curva se denomina mesocúrtica. Si el rango de los datos es limitado y los puntajes extremos se hallan hacia el centro, la curva se denomina leptocúrtica o curva en forma de pico. Si, al contrario, el rango de los datos es muy amplio con poca cantidad en el centro la curva se denomina platicúrtica, de forma llana. Estas curvas la vemos en la figura 2.9.



Por otra parte, los datos se podrían distribuir de una manera que no es normal, es decir, las colas de la curva no son simétricas. Cuando una desproporcionada cantidad de datos se carga hacia uno de los lados de la curva, hacia el inicio o final de la escala, se dice que es una curva sesgada. La distribución de datos está sesgada si no es simétrica y se alarga más hacia un lado que hacia el otro. Este sesgo puede tener dos formas. El sesgo negativo, con los datos sesgados hacia la izquierda, posee una cola izquierda más larga más prolongada. El sesgo positivo, con los datos sesgados a la derecha, posee una cola derecha más larga. Este tipo de curva la vemos en la figura 2.10.



Otras formas de curva que no son normales pueden originarse de acuerdo con la distribución de los datos. Las curvas bimodales o multimodales poseen dos o más modas, creando dos o más ápices. La curva rectangular se origina cuando hay frecuencias iguales en cada grupo. La curva en J se produce si hay una gran cantidad de datos hacia algún extremo y pocos en el centro, mientras que la curva en U implica una gran cantidad de datos hacia ambos extremos y pocos en el centro. Estas distribuciones no son comunes y deben ser observadas cuidadosamente al producirse.

EJERCICIOS

Los siguientes puntajes corresponden a una prueba de agilidad realizada por un grupo de estudiantes cuyos tiempos se tomaron en segundos y décimas de segundo.

X	X	X	X	X	X	X
12,1	13,0	15,9	11,8	11,4	10,4	12,8
12,9	9,3	11,0	10,5	10,1	11,6	11,6
13,1	11,4	11,7	11,7	13,4	12,5	12,8
15,4	13,8	14,2	9,6	12,6	13,2	12,5
12,2	11,8	12,9	12,1	11,1	13,3	12,0
12,3	10,9	10,3	14,4	9,1	11,5	12,4
11,9	13,4	9,8	15,3	13,0	10,6	14,6
11,8	12,7	12,0	12,1	12,6	12,3	12,8
10,4	10,8	13,1	10,3	12,7	12,9	12,4
10,8	14,8	14,7	11,2	11,9	13,5	13,6
13,6	11,7					

1. Determine el puntaje más alto y el más bajo, y calcule el rango.
2. Determine el tamaño del intervalo más apropiado para agrupar los datos.
3. Forme los grupos y registre la cantidad de puntajes por grupo. Desarrolle una columna de frecuencias y una columna de frecuencia acumulada. Establezca el límite aparente inferior de del grupo más bajo. ¿Cuál es N?
4. ¿Cuáles son los límites reales del grupo 12,0 – 12,4?
5. Dibuje un histograma de frecuencias.
6. Dibuje un polígono de frecuencias. ¿Son datos normales?
7. ¿Qué le puede suceder al polígono de frecuencias si se agregan más datos?
8. Utilizando un programa de estadística computarizado, introduzca los datos en una columna y produzca un histograma y un polígono de frecuencias.
9. Investigue, por lo menos, otros tres tipos de gráficos y dibújelos con base en los datos de la prueba de agilidad.

Capítulo 3. Medidas de Tendencia Central

Si se desea conocer la fuerza de un determinado número de personas, se puede aplicar una prueba de campo tal como la elevación de pesos, en dos oportunidades al menos. Si no hay diferencias en los resultados entre la primera y segunda prueba se puede decir que son similares, pero si son diferentes es difícil describir el grado de diferencia entre ambas si solo se tiene una imagen o una idea de esas diferencias. Una forma de superar estos problemas es usar medidas numéricas, que se pueden calcular a partir de una muestra o de una población de mediciones. Se pueden usar los datos para calcular un conjunto de números que los transformarán, empezando por la distribución de frecuencias. Estas mediciones se llaman parámetros cuando se asocian con la población y se denominan estadísticas cuando se calculan a partir de mediciones muestrales.

Las medidas de tendencia central son valores que describen el punto medio o central de un conjunto de datos. Se deben conocer las características del centro por lo que se busca obtener un número que represente ese valor central ya que una medida de tendencia central es un valor que se encuentra en el centro o la mitad de un conjunto de datos.

Estas medidas suministran una importante información ya que indican los puntos hacia los cuales los datos tienden a estar concentrados, permitiendo calcular la relación entre un puntaje dado y esos puntos medios de un conjunto de datos. Una de las primeras mediciones numéricas importantes es una medida de centro o medida de tendencia central, es decir, una medida a lo largo del eje horizontal que localiza el centro de la distribución la cual indica los puntos hacia los cuales los datos o puntajes tienden a estar concentrados. Si los datos se reflejan en un eje horizontal (abscisas) presentando los valores de x , los datos son distribuidos a lo largo de esta recta horizontal. Supongamos que 21 sujetos hicieron una prueba de levantamiento de pesos en kilogramos, alcanzando ese valor como un promedio de diez intentos y presentando los siguientes resultados: 55, 30, 45, 15, 35, 30, 35, 35, 60, 55, 45, 20, 45, 20, 40, 40, 45, 35, 30, 40, 50. Una vez organizados los datos, la distribución de la cantidad de sujetos con su respectivo puntaje se presenta en la figura 3.1.

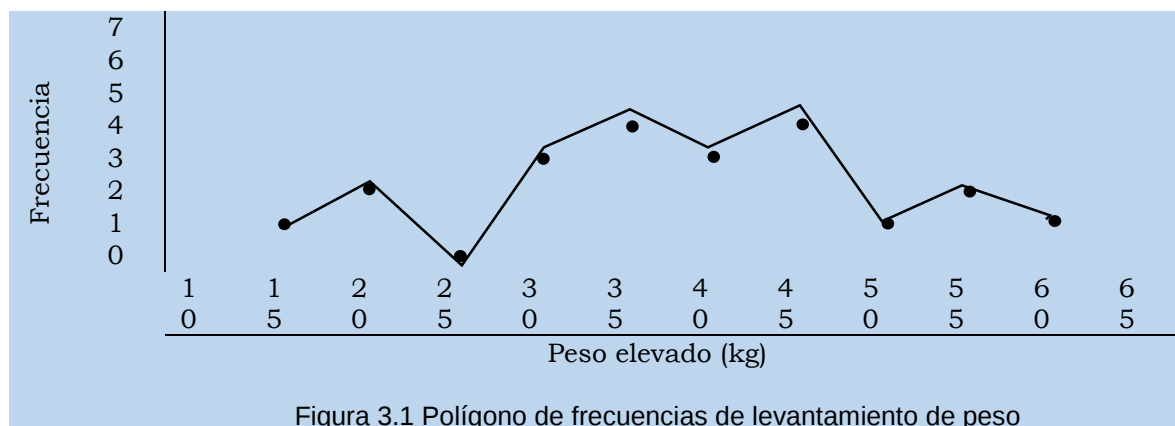


Figura 3.1 Polígono de frecuencias de levantamiento de peso

Por lo general se consideran tres medidas de tendencia central: la media, la mediana y la moda.

Media

Para determinar el centro de las mediciones se utiliza el promedio aritmético del conjunto de mediciones; es frecuente que esta medida se conozca como media aritmética o simplemente media. Para distinguir entre la media para la muestra y la media para la población, usamos la letra M para una media muestral y el símbolo μ para la media de una población. En este texto se utilizará fundamentalmente la media muestral.

La media es la medida de tendencia central más apropiada para el procesamiento de los datos de intervalo y de razón. No es viable calcular la media de un grupo de datos ordinales porque en ellos se desconoce la distancia relativa entre los puntajes.

La media aritmética o promedio de un conjunto de n mediciones es igual a la suma de las mediciones dividida entre n . La media (M) se calcula al sumar los puntajes ($\sum x$) y dividirlos entre el número de sujetos (n); Σ es la letra griega sigma mayúscula y pide sumar los valores que se encuentran a su derecha, por lo que Σx significa la suma de todas las mediciones de x .

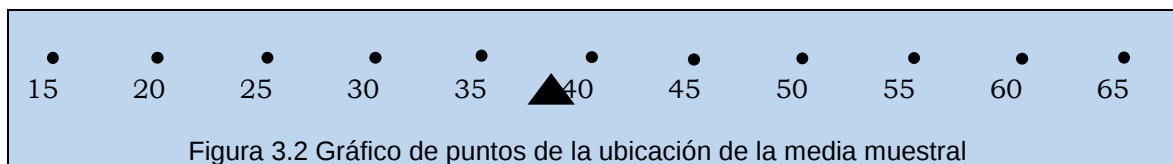
$$M = \frac{\Sigma x}{n}$$

Para ejemplificar tomemos los puntajes del levantamiento de peso. De acuerdo con la imagen el promedio pudiese hallarse alrededor de 40 repeticiones en el levantamiento de peso. Para hallar la media muestral se calcula:

$$M = \frac{55+30+45+15+35+30+35+35+60+55+45+20+45+20+40+40+45+35+30+40+50}{21} =$$

$$M = \frac{803}{21} = 38,2$$

Al trasladar los resultados a un gráfico de puntos tendríamos lo siguiente:



La media muestral de 38,2 sería un punto de equilibrio de todos los puntajes y corrobora la imagen de la figura 3.1 del polígono de frecuencias, donde hacia el centro del gráfico hay una elevación de la curva por haber mayor cantidad de sujetos con capacidad para levantar ese volumen de kilogramos, así como la menor cantidad de sujetos hacia los extremos.

Así, tenemos que la media es la medida de tendencia central de mayor uso en diferentes tipos de estudio; la media aritmética o promedio es la medida numérica que más se utiliza para describir datos ya que le da peso a cada puntaje de acuerdo a su distancia con respecto a otros puntajes en un conjunto de datos.

Debido a esta característica la media es la más sensible de todas las medidas de tendencia central por cuanto un ligero cambio en los datos no afecta a la moda ni la mediana, pero siempre cambia la media. La gran ventaja de la media es que considera toda la información en el procesamiento de los datos generando bases adicionales para diferentes cálculos con la consiguiente producción de información relevante.

Una forma de establecer la media una vez realizado su cálculo es redondear el resultado; un ejemplo lo tenemos con los siguientes datos:

12, 15, 18, 13, 10, 14, 14, 12, 16, 21, 15

x
12
15
18
13
10
14
14
12
16
21
15
$\Sigma x = 160$
Tabla 3.1 Cálculo de la media

$$M = \frac{\Sigma x}{n} = \frac{160}{11} = 14,545454545454 = 14,6$$

Como se observa en la tabla 3.1, no es necesario ordenar los datos de menor a mayor o viceversa, para calcular la media. Los datos son significativos solo al decimal más cercano a los resultados, por lo que el valor medio se calcula en este caso a la décima más cercana. Redondear la media a una cifra significativa más allá de los datos proporciona a los usuarios de los datos la información de que la media está más cerca de 14,6 que de 14,5. Los datos son significativos si la media se redondea a décimas al tratarse de

pruebas con puntajes en enteros (como en este ejemplo) aunque el resultado esté en centésimas, o se redondea a centésimas si los resultados están en milésimas, por cuanto esa cercanía es más adecuada a ese tipo de dato. A esto debemos agregar que en las pruebas de tiempo (velocidad, agilidad, etc.) la media se redondea en segundos y centésimas que están más cercanas al tiempo real tomado en segundos y milésimas.

Los mismos principios del cálculo de la media señalados anteriormente se aplican ante una distribución de frecuencias, modificando la fórmula para multiplicar cada valor por su frecuencia. La fórmula es la siguiente:

$$M = \frac{\Sigma(fx)}{n}$$

Esta ecuación requiere que cada puntaje original (x) se multiplique por su frecuencia, y los valores de fx se suman para ser divididos entre n , el número de sujetos en la prueba. Veamos el siguiente ejemplo.

x	f	fx
1	3	3
2	3	6
3	4	12
5	7	35
7	6	42
9	4	36
10	1	10
$n = 28$		$\Sigma = 144$
$M = 144 / 28 = 5,12857$ $M = 5,1$		
Tabla 3.2 Cálculo de la media de una distribución simple de frecuencias		

Para una distribución agrupada de frecuencias la media se calcula de manera similar, sólo que el puntaje x es reemplazado por el punto medio del grupo:

$$M = \frac{\Sigma(fx)_{pm}}{n}$$

donde x_{pm} es el punto medio de cada grupo. Tomando la distribución agrupada de frecuencias de la prueba de salto largo sin impulso y su punto medio, con $n = 50$, calculemos la media para datos agrupados (Tabla 3.3).

Grupo	Punto medio	f	$(fx)_{pm}$
117-124	120,5	3	361,5
125-132	128,5	1	128,5
133-140	135,5	2	271,0
141-148	144,5	5	722,5
149-156	152,5	2	305,0
157-164	160,5	4	642,0
165-172	168,5	6	1011,0
173-180	176,5	9	1588,5
181-188	184,5	2	369,0
189-196	192,5	7	1347,5
197-204	200,5	3	601,5
205-212	208,5	3	625,5
213-220	216,5	1	216,5
221-228	224,5	0	0,0
229-236	232,5	2	465,0
$\Sigma(fx)_{pm}$			8655,0

Tabla 3.3 Cálculo de la media en una distribución agrupada de frecuencias

$$M = \frac{\Sigma(fx)_{pm}}{n} = \frac{8655}{50} = 173,1$$

El procesamiento a mano, por medio de una calculadora, puede ser largo y tedioso. Las computadoras tienen programas que permiten realizar cualquier cálculo con una gran cantidad de datos en poco tiempo, tan solo introduciendo los datos y realizando las operaciones correspondientes. Si no se tiene una computadora el cálculo de la media varía ligeramente, especialmente si se tiene una distribución simple o agrupada de frecuencias.

Mediana

Una segunda medida de tendencia central es la mediana (m), que es el valor de la posición media en el conjunto de mediciones ordenada de menor a mayor. La mediana m de un conjunto de n mediciones es el valor de x que cae en la posición media cuando las mediciones son ordenadas de menor a mayor; es el valor intermedio en un conjunto de datos una vez que los datos hayan sido ordenados: la mitad de los puntajes estarán por debajo de la mediana y la otra mitad por arriba. La mediana también es un puntaje típico porque es el puntaje que mejor representa la mayoría de los demás valores.

En una distribución ordenada donde el número de sujetos es impar la mediana es el puntaje intermedio en el rango; por ejemplo, si tenemos los resultados en la tabla 3.4 de la prueba de lanzamiento medida en metros en la columna A (18, 14, 21, 12 y 19), los puntajes se ordenan y se determina la mediana. En la columna B se agrega el puntaje 24 a los datos anteriores y se muestra la cantidad de seis valores por lo que la mediana se encuentra entre dos puntajes.

A	B
12	12
14	14
Mediana → 18	18 ← Me-
19	19
21	diana
	21
	24

Tabla 3.4 La mediana

La observación central en la columna A, marcada con una flecha, es el centro del conjunto o sea $m = 18$. El valor 18 se encuentra en la tercera posición partiendo desde ambos extremos por los que ocupa el lugar central del conjunto de cinco datos.

La determinación de la mediana no toma en cuenta el valor de los puntajes sino el lugar que ocupa en la distribución ordenada. Esta posición se puede calcular con una sencilla fórmula:

$$m = \frac{n + 1}{2}$$

En el caso de la columna A de la tabla, el número de sujetos es 5 por lo que tenemos lo siguiente:

$$m = \frac{5 + 1}{2} = \frac{6}{2} = 3$$

La mediana se encuentra en la tercera posición de los datos, que corresponde al valor 18. Con respecto a la columna B, donde $n = 6$, al realizar la operación $(6 + 1) / 2$ el resultado es la posición 3,5 de la distribución, que se corresponde con el puntaje 18,5 como ya se mencionó. En la columna B tenemos un conjunto de seis valores y ahora hay dos observaciones centrales. Para hallar la mediana, escoja un valor a la mitad entre esas dos observaciones centrales, 18 y 19, por cuanto hay tres valores hacia cada extremo:

$$m = \frac{18 + 19}{2} = 18,5$$

Como el resultado es 18,5 y este puntaje no existe, se asume como mediana el puntaje superior que en este caso es 19. Otra fórmula para calcular la mediana es la siguiente:

$$m = lri + \left(\frac{50 - \% \text{ acumulado bajo } X}{\text{Porcentaje de } X} \right) UM$$

donde X es el puntaje donde está acumulado al menos el 50%, lri es el límite real inferior, % acumulado por debajo de X , porcentaje de X es el % donde se encuentre X , y UM es la unidad de medida expresada por los puntajes. Por ejemplo, en una carrera de velocidad en 100 metros tenemos los siguientes resultados:

Puntaje X	Frecuencia	% de X	% acumulado
10,0	2	5,00	5,00
10,1	3	7,50	12,50
10,2	4	10,00	22,50
10,3	7	17,50	40,00
10,4	8	20,00	60,00
10,5	9	22,50	82,50
10,7	5	12,50	95,00
11,0	2	5,00	100,00

Tabla 3.5 Datos para cálculo de la mediana

En $X = 10,4$ está acumulado al menos el 50%, el límite real inferior (lri) es 10,35 (10,4 – 0,05), el % acumulado por debajo de X es 40, y la UM es 0,1 ya que los puntajes están expresados en décimas de segundo. Al sustituir y efectuar el procedimiento tenemos:

$$\begin{aligned}
 m &= 10,35 + \left(\frac{50 - 40}{20} \right) (0,1) = 10,35 + \left(\frac{10}{20} \right) (0,1) \\
 &= 10,35 + (0,5)(0,1) = 10,4
 \end{aligned}$$

Por las características anteriormente señaladas, la mediana es más apropiada para estudiar datos ordinales y con datos sesgados, por ejemplo, cuando hay un puntaje extremo que es radicalmente diferente a los demás puntajes. Veamos un ejemplo: Si nueve personas realizan cada uno 40 abdominales y uno hace 200, la mediana del grupo es 40 pero el promedio es 56; aun si el puntaje más alto fuese 1000, la mediana continuaría siendo 40 pero el promedio variaría sustancialmente. Por ello se dice que la mediana es más representativa de la mayoría de los puntajes que el promedio cuando se presentan puntajes extremos en un conjunto de datos.

Moda

Otra forma de localizar el centro de una distribución es buscar el valor de x que presenta la mayor cantidad de datos, esta medida del centro se denomina moda: la moda es el puntaje que se presenta con mayor frecuencia en los datos que se estudian. La moda no se

calcula mediante alguna fórmula, sino que se determina al organizar e inspeccionar los datos revisando una lista ordenada de puntajes para visualizar cuál de los puntajes es el que más se repite, ya sea que estén organizados en una distribución simple de frecuencias o en una distribución agrupada de frecuencias.

En una distribución simple de frecuencias el puntaje con mayor número de frecuencias es la moda, pero si los datos se organizan en una distribución agrupada de frecuencias la moda es el punto medio del grupo donde hay más frecuencias. Por otra parte, cualquier distribución puede tener dos o más modas, si dos o más puntajes o grupos tienen la misma frecuencia, denominándose bimodales o multimodales.

Veamos algunos resultados y la moda en cada caso:

a) Velocidad (1/100)	b) Salto (cm)	c) Lugar en competencia
9,35	85	1
6,11	85	3
6,05	85	4
7,27	91	5
7,48	91	7
6,11	91	8
8,03	98	9
9,00	97	10
Tabla 3.6 Determinación de la moda		

En la primera prueba, velocidad, la moda es 6,11; en el salto hay dos modas, 85 y 91, mientras que en la tercera no hay moda. La moda no es una medida estable ya que si se agrega o elimina un puntaje puede cambiar considerablemente; por ejemplo, en la tabla 25 si en vez del puntaje 98 fuese otro puntaje 91 en el salto, la moda ya no serían dos sino una, precisamente 91 con cuatro frecuencias. Si se tienen solo datos nominales la moda es la única medida de tendencia central aplicable; para datos numéricos las otras medidas de tendencia central son más apropiadas.

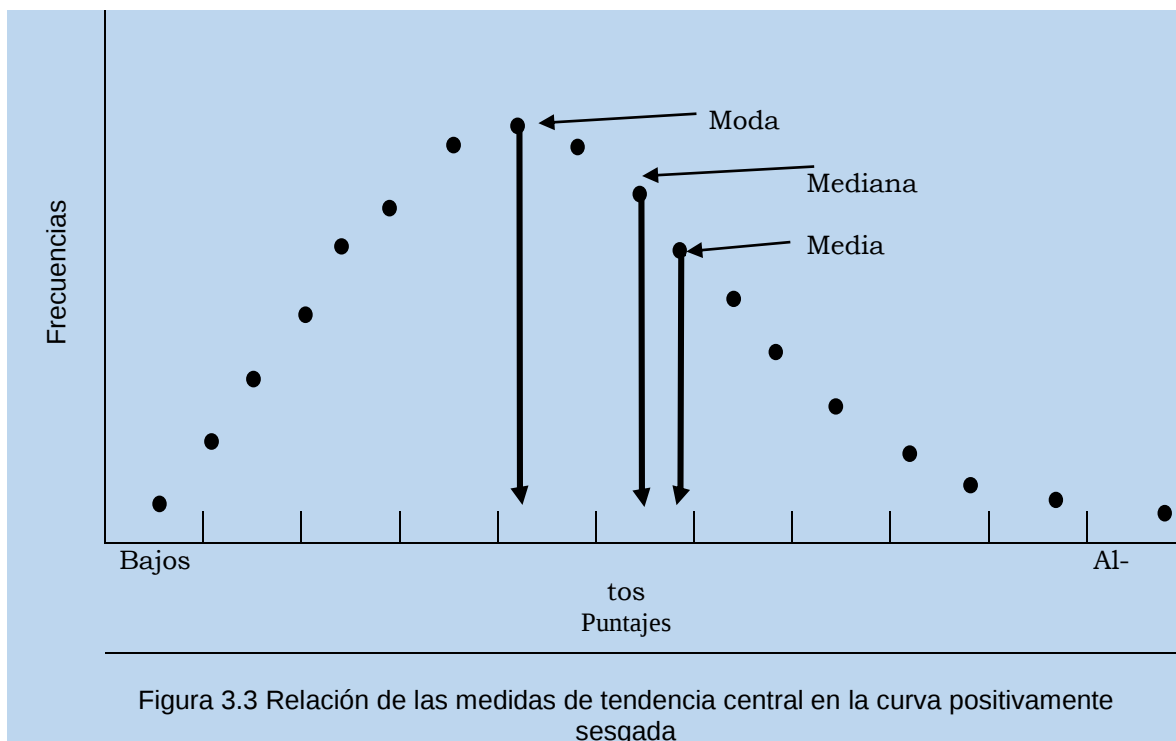
Cuando las mediciones en una variable continua se han agrupado como histograma de frecuencia o de frecuencia relativa, la clase con el pico más alto o frecuencia se llama clase modal, y el punto medio de esa clase se toma como la moda. La moda por lo general se usa para describir conjuntos grandes de datos, mientras que la media y la mediana se usan para conjuntos de datos grandes y pequeños.

Es posible que una distribución de mediciones tenga más de una moda, como ya se mencionó. Por ejemplo, si fuéramos a procesar la cantidad de gente practicando actividades físicas los días domingos, podríamos obtener una distribución bimodal, posiblemente reflejando una mezcla de jóvenes adultos y personas de tercera edad en la pobla-

ción, pudiendo presentarse distribuciones bimodales de tamaños o pesos reflejan una mezcla de mediciones tomadas, por ejemplo, de mujeres y hombres. En cualquier caso, un conjunto o distribución de mediciones puede tener más de una moda.

Relación entre Media, Mediana y Moda

Cuando los datos están distribuidos normalmente, las tres medidas de tendencia central caen el mismo valor o cerca de él. Pero cuando los datos están sesgados, estas tres medidas ya no son tan idénticas. La figura 3.3 demuestra la relación entre las tres medidas en una distribución sesgada positivamente.



Tenga en cuenta que las puntuaciones más altas están más alejadas de la moda que las puntuaciones más bajas. Estas características de las curvas sesgadas positivamente desplazan la mediana y la media a la derecha de la moda. En una curva sesgada positivamente, las tres medidas de tendencia central se leen de izquierda a derecha en el siguiente orden: moda, mediana, media (Figura 3.3). En una curva sesgada negativamente, el orden se invierte: media, mediana, moda.

Al decidir cuál de las tres medidas usar, considere lo siguiente:

- Utilice la moda solo si se necesita una estimación aproximada de la central y los datos son normales o casi.
- Use la mediana si (a) los datos están en una escala ordinal, (b) se necesita el puntaje medio del grupo, (c) se necesita el puntaje más típico o (d) la curva está muy sesgada por puntajes extremos.

- Utilice la media si (a) la curva es casi normal y los datos son del tipo intervalo o razón, (b) se debe considerar toda la información disponible de los datos (es decir, el orden de las puntuaciones y sus valores relativos), o (c) se realizarán cálculos adicionales, tales como desviaciones estándar o puntajes estándar.

EJERCICIOS

1. Calcule la moda, la mediana y la media para los dos conjuntos de datos:

A)

X
2
2
4
5
8
8
8
9
11
11
12
12
13
15
17

B)

X	f
8	3
9	5
10	9
11	10
12	8
13	7
14	7
15	5
16	4
17	2

2. Introduzca los datos de la pregunta 1 en un programa de computación y compare los resultados con los cálculos hechos con calculadora. Los datos de la tabla B deben incluirse tantas veces como lo señale la columna de frecuencias.
3. Calcule la mediana para cada conjunto de datos:
 - a. 1, 13, 12, 1, 8, 4, 5, 10, 2, 5, 6, 8, 9
 - b. 7, 13, 2, 1, 1, 9, 12, 5, 6, 11, 4, 10
 - c. 5, 1, 4, 8, 14, 7, 1, 2, 5
4. En una clase de educación física de la Universidad del Zulia hay 24 estudiantes los cuales trabajan a destajo o tiempo convencional con un salario muy bajo. Si se incluye al profesor con un salario superior, ¿cuál medida es mejor para describir el ingreso del grupo de estas 25 personas, la media o la mediana? ¿Por qué?
5. Catorce estudiantes de Fisiología del Ejercicio midieron la presión sanguínea alta de la misma persona cuyas lecturas se presentan a continuación (en mmHg). ¿Qué se puede destacar de este conjunto de datos de acuerdo con la media y la mediana?
138, 130, 135, 140, 120, 125, 120, 130, 130, 144, 143, 140, 130, 150
6. Es conocido que los hombres tienden a pesar más y a ser más altos que las mujeres. El índice de masa corporal (IMC) se basa en la estatura y el peso. Si tomamos en cuenta la media y la mediana de los valores del IMC que se presentan, ¿Hay diferencias notables?
Hombres: 23.8, 23.2, 24.6, 26.2, 23.5, 24.5, 21.5, 31.4, 26.4, 22.7, 27.8, 28.1
Mujeres: 19.6, 23.8, 19.6, 29.1, 25.2, 21.4, 22.0, 27.5, 33.5, 20.6, 29.9, 17.7

Capítulo 4. Medidas de Variabilidad

Los conjuntos de datos pueden tener el mismo centro, es decir, la media, pero con aspecto diferente por la forma en que los números se dispersan desde el centro hacia los extremos. Considere las dos distribuciones que se muestran en la figura 4.1. Ambas distribuciones están centradas, pero hay una gran diferencia en la forma en que las mediciones se dispersan o varían. Las mediciones de la figura 4.1, a) varían de 2 a 5; b) varían de 0 a 8.

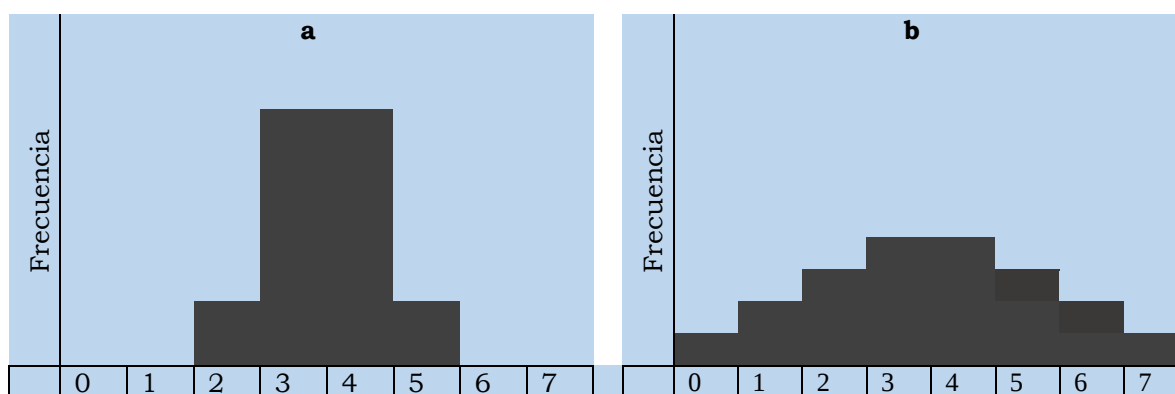


Figura 4.1 Distribución centrada de los puntajes, a) entre los puntajes 2 y 5; b) entre 0 y 7

Las medidas de variabilidad o variación describen un conjunto de datos en términos de su separación o heterogeneidad. Dado que el sólo uso de la media no da detalles sobre la diferencia entre los puntajes, debe conocerse cómo difieren los datos entre sí en un conjunto de datos o con respecto a otro conjunto de datos. La variabilidad o dispersión es una muy importante característica de datos. Por ejemplo, si los jugadores de baloncesto lanzan tiros libres en las prácticas, la variación extrema en la precisión de los lanzamientos en situaciones de juego real causaría un alto porcentaje de lanzamientos no anotados. Por el contrario, si se estuviera tratando de discriminar entre lanzadores de tiro libre buenos y malos, tendría problemas si en el juego se produjeran anotaciones de tiro libre con poca variación, lo cual hace muy difícil la discriminación.

En otra situación deportiva de carácter hipotético, tenemos dos equipos de baloncesto (A y B) que están empatados a 60 puntos al finalizar la primera parte del juego. Se presentan los cinco jugadores de ambos equipos que han anotado (Tabla 4.1)

Equipo	Anotación de los jugadores					Total de puntos
A	6	24	13	5	12	60
B	12	12	12	12	12	60

Tabla 4.1 Variabilidad de la anotación en baloncesto

En esta situación tanto la media como la mediana de los dos equipos es el mismo puntaje, es decir, 12. Si el respectivo informe señala que los dos equipos demostraron igual nivel de destreza porque la media y la mediana son idénticas en ambos casos se estaría incurriendo en un error ya que el equipo B es mucho más consistente. A nivel táctico la pérdida del jugador con 24 puntos afectaría con mayor incidencia al equipo A si el equipo B pierde cualquier jugador.

Cuando se ha calculado la media, la mediana y la moda, se determina fácilmente la tendencia central de los datos. Así como los datos se concentran de acuerdo a su tendencia central, igualmente se van a separar o dispersar alejándose del centro de acuerdo con su variabilidad. Si se conocen la tendencia central y la variabilidad, dos o más conjuntos de datos pueden compararse integralmente. Es interesante comparar la media de dos o más conjuntos de datos para determinar cuál tiene un promedio más alto, pero puede ser más importante comparar la variabilidad de los datos ya que la dispersión diferirá mientras la media puede ser similar.

A continuación, se revisan tres medidas de variación: el rango, la desviación estándar y la varianza. Los dos últimos tienen importancia fundamental porque conducen a cálculos adicionales que responden mejor las dudas sobre los datos. Las medidas de variabilidad pueden ayudarle a crear una imagen mental de la dispersión de los datos. La medida más sencilla de variación es el rango.

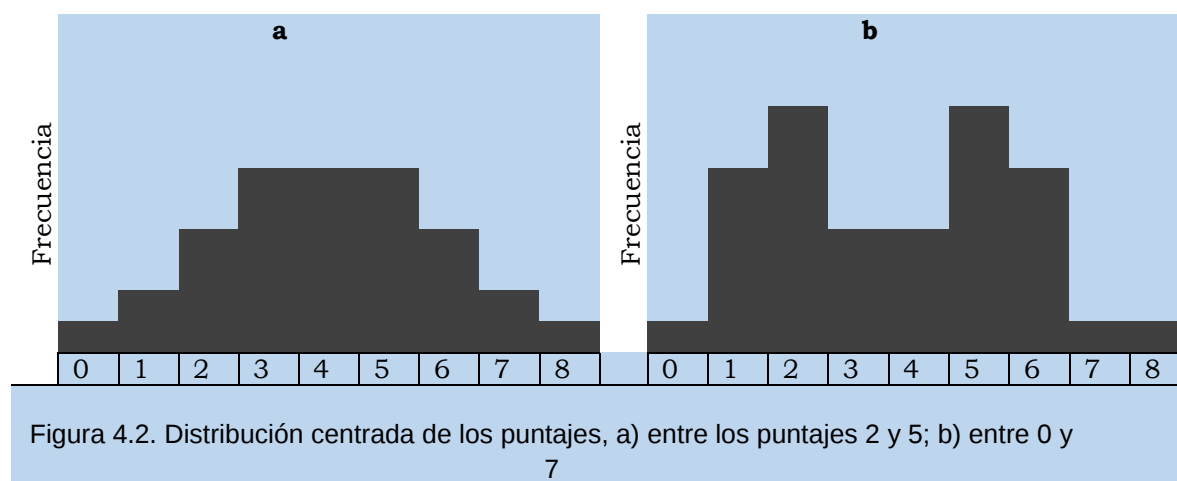
Rango

El rango, R , de un conjunto de n mediciones se define como la diferencia entre la medición más grande y la más pequeña, en un conjunto de datos.

$$\text{rango} = \text{puntaje máximo} - \text{puntaje mínimo}$$

Es bastante simple calcular el rango y como depende de los puntajes extremos de un conjunto de datos, su utilidad es muy poca si se compara con otras medidas de variación porque estas consideran todos los puntajes. Además, el rango es muy inestable como medida de variabilidad ya que cualquier cambio en los puntajes extremos hace que cambie sustancialmente.

Por ejemplo, para los datos de una carrera de velocidad en 30 metros medida en segundos y décimas de segundo, las mediciones varían de 5.6 a 9.4. Por tanto, el rango es $9.4 - 5.6 = 3.8$. El rango es fácil de calcular, fácil de interpretar y es una medida adecuada de variación para conjuntos pequeños de datos. Para conjuntos grandes, el rango no es una medida adecuada de variabilidad. Por ejemplo, las dos distribuciones de frecuencia relativa de la figura 4.2 tienen el mismo rango, pero muy diferentes formas y variabilidad.



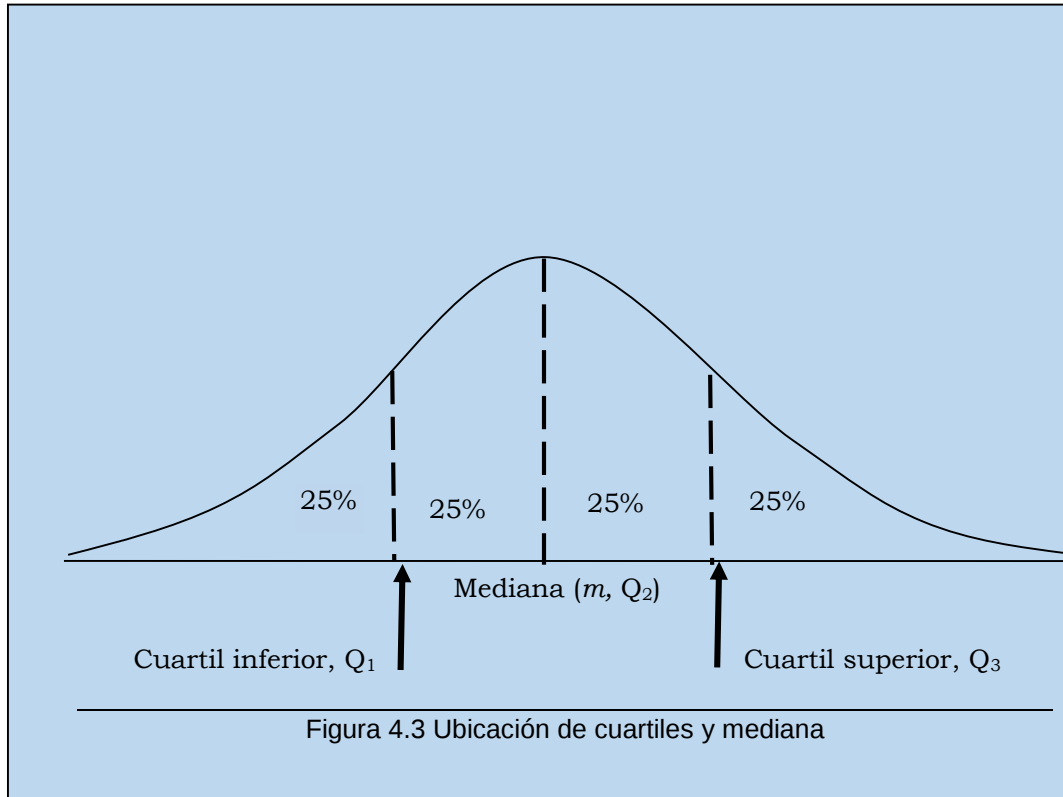
Rango intercuartil

Otra medida de variabilidad que se usa a menudo cuando los datos son ordinales, o cuando los datos de relación o intervalo están muy sesgados, se llama rango intercuartil o IQR. Este rango es la diferencia entre los puntajes brutos en los puntos de percentil 75 (Q_3) y 25 (Q_1). Se calcula de la siguiente manera:

$$IQR = Q_3 - Q_1$$

El cuartil inferior (primer cuartil), Q_1 , es el valor de x que es mayor a un cuarto de las mediciones y es menor que los restantes tres cuartos. El segundo cuartil es la mediana, Q_2 . El cuartil superior (tercer cuartil), Q_3 , es el valor de x que es mayor a tres cuartos de las mediciones y es menor que el restante un cuarto.

El IQR es una medida útil de la dispersión de los datos si el investigador está más interesado en los puntajes medios que en los extremos. Debido a que el rango intercuartil considera solo el 50% de los datos (los puntajes que caen en la mitad del conjunto de datos), no se ve afectado por puntajes altamente divergentes en los extremos, es por eso que es útil para datos asimétricos. Al igual que la mediana, presenta una imagen típica, pero no considera toda la información disponible sobre los datos. Recuerde que la mediana, m , de un conjunto de datos es la medición central y es Q_2 ; esto es, 50% de las mediciones son más pequeñas y 50% son más grandes que la mediana.



Para conjuntos de datos pequeños, con frecuencia es imposible dividir el conjunto en cuatro grupos, cada uno de los cuales contiene exactamente 25% de las mediciones. Por ejemplo, cuando $n = 10$, se necesita tener 2,5 mediciones en cada grupo. Aun cuando usted efectúe esta tarea (por ejemplo, si $n = 12$), hay muchos números que satisfarían la definición precedente y, por lo tanto, podrían ser considerados “cuartiles”. Para evitar esta ambigüedad, usamos la siguiente regla para localizar cuartiles muestrales.

- Cuando las mediciones están dispuestas en orden de magnitud, el cuartil inferior, Q_1 , es el valor de x en la posición $0,25(n+1)$, y el cuartil superior, Q_3 , es el valor de x en la posición $0,75(n+1)$.
- Cuando $0,25(n+1)$ y $0,75(n+1)$ no son enteros, los cuartiles se encuentran por interpolación, usando los valores de las dos posiciones adyacentes.

Por ejemplo, si tenemos un equipo de voleibol y participan en un juego donde los jugadores efectúan un total de remates efectivos, encontrar los cuartiles inferior y superior para este conjunto de mediciones:

9, 1, 1, 9, 2, 7, 6, 6, 5

En primer lugar, ordene las mediciones ($n = 9$) de menor a mayor:

1, 1, 2, 5, 6, 6, 7, 9, 9

En segundo lugar, calcule la posición de Q_1 y Q_3 .

$$\text{Posición de } Q_1 = 0,25 (n+1) = 0,25 (9+1) = 2,5$$

$$\text{Posición de } Q_3 = 0,75 (n+1) = 0,75 (9+1) = 7,5$$

Como estas posiciones no son enteros, el cuartil inferior se toma como el valor $3/4$ de la distancia entre la segunda y tercera mediciones ordenadas, y el cuartil superior se toma como el valor $1/4$ de la distancia entre la octava y novena mediciones ordenadas. Es decir, la posición de Q_1 está entre las posiciones segunda y tercera de las mediciones (entre los puntajes 1 y 2), y la posición de Q_3 está entre las posiciones séptima y octava de las mediciones (entre los puntajes 7 y 9). Por tanto,

$$Q_1 = 1 + 0,75 (2 - 1) = 1 + 0,75 = 1,75$$

$$Q_3 = 7 + 0,25 (9 - 7) = 7 + 0,5 = 7,5$$

El rango intercuartil de este grupo indicaría que el 50% de los puntajes entre Q_3 y Q_1 estaría entre 2 y 6 ($IQR = 7,5 - 1,75 = 5,75$)

Ante este resultado, la dirección técnica del equipo de voleibol podría tomar decisiones ya sea de entrenamiento o de contratación, con respecto a los jugadores de menor rendimiento y que se encuentran por debajo de Q_1 , dentro del cuartil inferior.

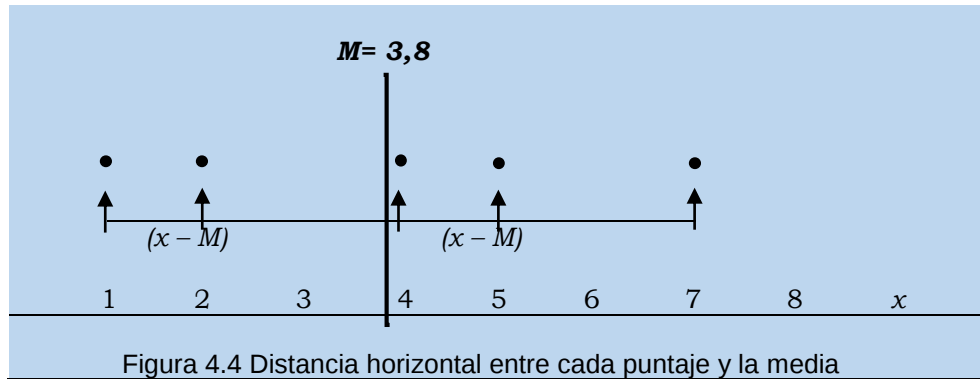
Varianza

Dado que el rango no incluye los demás datos se utilizan otros procedimientos para valorar la variabilidad y que, si consideran cada puntaje, determinando la distancia de cada uno con respecto a la media. Esta distancia se denomina desviación y su denotación es d .

Para ejemplificar veamos las mediciones muestrales 5, 7, 1, 2, 4, mostradas como una gráfica de puntos en la figura 4.4. La media de estas cinco mediciones es:

$$M = \frac{\sum x}{n} = \frac{19}{5} = 3,8$$

La expresión de las mediciones y la media en un gráfico de puntos sería el siguiente:



Las distancias horizontales entre cada medición o puntaje y la media M ayudarán a medir la variabilidad. Si las distancias son grandes, los datos son más dispersos o variables que si las distancias son pequeñas. Si x es un punto particular o medición, entonces la desviación de esa medición desde la media es $(x - M)$. Las mediciones a la derecha de la media producen desviaciones positivas y las de la izquierda, negativas. Los valores de x y las desviaciones para nuestro ejemplo se detallan en las columnas primera y segunda de la tabla 4.1 ($M = 3,8$).

x	$(x - M)$	$(x - M)^2$
1	-2,8	7,84
2	-1,8	3,24
4	0,2	0,04
5	1,2	1,44
7	3,2	10,24
$\Sigma x = 19$	$\Sigma(x - M) = 0,0$	$\Sigma(x - M)^2 = 22,80$

Tabla 4.1 Cálculo de $\Sigma(x - M)^2$

Como las desviaciones en la segunda columna de la tabla contienen información sobre variabilidad, una forma para combinar las cinco desviaciones en una medida numérica es promediarlas. Desafortunadamente, el promedio no funcionará porque algunas de las desviaciones son positivas, otras son negativas, por lo que las desviaciones en la segunda columna de la tabla 4.1 suman cero: la suma de todas las desviaciones debe ser cero, independientemente del tamaño de los puntajes. Es una de las maneras de verificar la precisión de la media. Si la suma de las desviaciones no es cero, la media es incorrecta. Si los puntajes se encuentran por arriba o por debajo de la media entonces tienen un signo que puede ser positivo o negativo. Para eliminar los signos negativos se procede a elevar al cuadrado las desviaciones por lo que todos los valores al cuadrado quedan positivos.

El promedio de las desviaciones al cuadrado provee un método para la determinación de la variabilidad de los datos al tomar en cuenta cada puntaje y su distancia a la media; de la suma de desviaciones cuadradas se calcula una sola medida llamada varianza: es el promedio de las desviaciones al cuadrado desde la media.

Para distinguir entre la varianza de una muestra y la varianza de una población, usamos el símbolo s^2 para una varianza muestral y σ^2 para una varianza de población. La varianza será relativamente grande para datos muy variables y relativamente pequeña para datos menos variables. La varianza de una población de N mediciones es el promedio de los cuadrados de las desviaciones de las mediciones alrededor de su media M . La varianza poblacional se denota con σ^2 y está dada por la fórmula:

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

La mayor parte de las veces, no tendremos todas las mediciones de población disponibles, pero necesitaremos calcular la varianza de una muestra de n mediciones para determinar la distancia entre los puntajes y la media M . La varianza de una muestra de n mediciones es la suma de las desviaciones cuadradas de las mediciones alrededor la media M dividida entre $(n - 1)$. La varianza muestral se denota con s^2 y está dada por la siguiente fórmula:

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1} = \frac{\Sigma d^2}{n - 1}$$

Tomando el resultado de la tabla 4.1 donde la suma de las desviaciones al cuadrado es 22,80 tenemos:

$$s^2 = \frac{22,80}{5 - 1} = \frac{22,80}{4} = 5,7$$

En la siguiente tabla sobre una prueba de flexibilidad en centímetros tenemos dos conjuntos de datos, X y Y , mediante los cuales se ejemplifica el cálculo de la varianza observándose que la media, la suma de los puntajes originales y de la desviación son idénticas, mientras que la suma de las desviaciones al cuadrado y la varianza demuestran que la variabilidad en Y es mucho mayor. En otras palabras, los puntajes en x están más cercanos entre sí y a la media M , mientras los puntajes en y están más separados entre sí y alejados de la media M , más dispersos.

x	d	d ²	y	d	d ²
27	+2	4	35	+10	100
26	+1	1	30	+5	25
25	0	0	25	0	0
24	-1	1	20	-5	25
23	-2	4	15	-10	100
$\Sigma X = 125$	$\Sigma d = 0$	$\Sigma d^2 = 10$	$\Sigma Y = 125$	$\Sigma d = 0$	$\Sigma d^2 = 250$
$M = \frac{125}{5} = 25$			$M = \frac{125}{5} = 25$		
$S^2 = \frac{10}{5} = 2$			$S^2 = \frac{250}{5} = 50$		

Tabla 4.2. Cálculo de la varianza en la prueba de flexibilidad

La varianza es útil en algunos métodos estadísticos relevantes como el análisis de regresión y el análisis de varianza. Por lo general, la varianza tiene una gran desventaja por cuanto las unidades de la varianza son diferentes a las unidades de los datos originales. En el caso de la tabla 30, nos podemos preguntar qué son 50 cm² o 2 cm² si los datos originales están en centímetros. Dada esta característica es más utilizada la desviación estándar para determinar la variabilidad de los datos.

Desviación estándar

Para llevar el valor de la varianza a valores similares de otras medidas de variabilidad, así como a la unidad de valor de los datos originales, se determina la raíz cuadrada de la varianza. El valor resultante se denomina desviación estándar porque se estandariza con respecto a la unidad de valor de los datos originales.

La desviación estándar, simbolizada con la letra *s*, es la medida de variabilidad utilizada conjuntamente con la media que indica la medida en que todos los puntajes difieren o se desvían desde la media; la mayoría de los puntajes difieren de la media y la mayor magnitud de esa diferencia la representa la desviación estándar. El valor de la desviación estándar es positivo y sólo es igual a 0 si todos los valores de los datos son idénticos por lo que nunca es negativa. Si el valor de *s* es alto habrá una mayor cantidad de variación; asimismo, el valor de *s* puede aumentar demasiado si se incluyen valores extremos, muy alejados de los demás.

La desviación estándar se calcula según la siguiente fórmula:

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}$$

Los símbolos $(\sum x)^2$ y $\sum x^2$ en la fórmula son métodos breves para indicar la operación aritmética que es necesario efectuar. En la fórmula para la media muestral $\sum x$ es la suma de todas las mediciones. Para hallar $(\sum x)^2$, eleve al cuadrado cada medición individual y luego súmelas. Deben leerse así:

$\sum x^2$ = Suma de cuadrados de las mediciones individuales

$(\sum x)^2$ = Cuadrado de la suma de las mediciones individuales

Si se toman los puntajes x de la tabla 4.1 para calcular la varianza y la desviación estándar, se procede de la siguiente manera:

x	x^2
1	1
2	4
4	16
5	25
7	49
$\Sigma = 19$	$\Sigma = 95$
Tabla 4.3 Base para el cálculo de s^2 y s	

Con las entradas en la tabla 4.3 que son las mediciones individuales, x , y sus cuadrados, x^2 , junto con sus sumas, se utiliza la fórmula computacional para s^2 , continuando con el procedimiento para calcular primero la varianza:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1} = \frac{95 - \frac{19^2}{5}}{5 - 1} = \frac{95 - \frac{361}{5}}{4} = \frac{95 - 72,2}{4} = \frac{22,8}{4} = 5,7$$

Ahora se calcula la desviación estándar, considerando que la desviación estándar de la muestra, s , es la raíz cuadrada de s^2 :

$$s = \sqrt{s^2} = \sqrt{5,7} = 2,39$$

Otro ejemplo del cálculo de la desviación estándar siguiendo los mismos pasos (calcular $\sum x$, $\sum x^2$, y s), teniendo los siguientes puntajes de los goles marcados por siete sujetos en 10 intentos cada uno: 7, 2, 7, 6, 5, 6, 2. Primero se crea una tabla para introducir los datos en las filas y luego se suman las columnas para determinar Σ .

<i>n</i>	<i>x</i>	<i>x</i> ²
1	7	49
2	2	4
3	7	49
4	6	36
5	5	25
6	6	36
7	2	4
Σ	35	203

Tabla 4.4 Cálculo de la desviación estándar

donde $\Sigma x = 35$, $\Sigma x^2 = 203$ y $n = 7$ según los valores hallados, procediéndose a aplicar la fórmula de *s*:

$$s = \sqrt{\frac{203 - \frac{35^2}{7}}{7 - 1}} = \sqrt{\frac{203 - \frac{1225}{7}}{6}} = \sqrt{\frac{203 - 175}{6}} = \sqrt{\frac{28}{6}} = \sqrt{4.67} = 2,2$$

Si los mismos sujetos hicieran una prueba de abdominales tendríamos los siguientes resultados:

<i>n</i>	<i>x</i>	<i>x</i> ²
1	65	4225
2	60	3600
3	58	3364
4	42	1764
5	51	2601
6	61	3721
7	45	2025
Σ	382	21300

Tabla 4.5 Cálculo de la desviación estándar

$$s = \sqrt{\frac{21300 - \frac{382^2}{7}}{7 - 1}} = \sqrt{\frac{21300 - \frac{145924}{7}}{6}} = \sqrt{\frac{21300 - 20846,286}{6}}$$

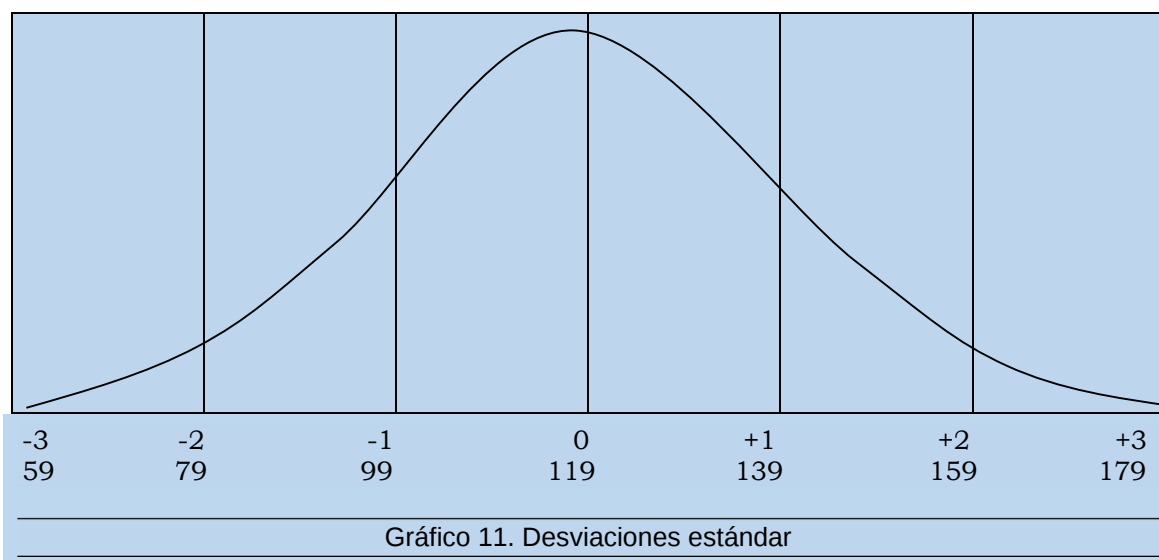
$$s = \sqrt{\frac{453,714}{6}} = \sqrt{75,619} = 8,7$$

Al comparar una prueba con otra se nota que la desviación estándar en la prueba de abdominales es más alta que en la prueba de marcación de goles, indicando que en la

primera prueba hay mayor variabilidad o dispersión. Esta ecuación se aplica a muestras de una población y se corrige con base al concepto de grados de libertad.

La desviación estándar es el método de medición de la variabilidad o dispersión más común y determina la distancia de cada puntaje en particular con respecto a la media del grupo. Si se tienen varias pruebas, a cada una de ellas se le calcula la media y la desviación estándar; luego se pueden hacer estudios para determinar, por ejemplo, la separación relativa de las pruebas o los puntajes estándar, entre otros

• Cuando n es grande y la distribución de los datos es normal, los datos caen dentro de un rango de 5 o 6 desviaciones estándar que se distribuyen a partir de la media de los datos, 3 desviaciones por arriba de la media y 3 desviaciones por debajo de la media, como se muestra en la figura 4.5. Si en una prueba de potencia la media es 119 y la desviación estándar 20, se suma una desviación estándar cada vez para hallar la primera, segunda y tercera desviación por arriba de la media (hacia la derecha en la curva), y se resta una desviación estándar cada vez para hallar la primera, segunda y tercera desviación por debajo de la media (hacia la izquierda en la curva). En las pruebas donde el menor puntaje es el mejor como las pruebas de velocidad se procede, al contrario, se resta hacia la derecha y se suma hacia la izquierda.



Grados de libertad. El grado de libertad es la cantidad de valores de un conjunto de datos que tienen libertad para variar. Si no se colocan restricciones a los datos, entonces todos los valores tienen libertad para variar; esto es, ellos pueden tomar cualquier valor limitados solo por la precisión del instrumento de medición y los valores reales de la población. Pero si se toman los datos a una muestra y se asume que esa muestra representa a la población, se debe colocar una restricción a los datos de la muestra.

Si la media de cuatro valores es 5, la suma de los cuatro puntajes debe ser 20. Siendo los puntajes 2, 3 y 7 los primeros tres puntajes, ¿cuál debe ser el cuarto puntaje para que la suma sea 20? Debe ser 8. El último puntaje no tiene libertad para variar, se limita a un único puntaje para que la suma sea 20. En este ejemplo hay tres grados de libertad; tres de los cuatro puntajes pueden asumir cualquier valor, pero al menos uno debe ser aquel cuyo valor sea la cantidad necesaria para sumar 20. Los grados de libertad para una muestra que representa una población, siempre es $n-1$. En este ejemplo $gl = n - 1 = 4 - 1 = 3$.

EJERCICIOS

1. Los siguientes datos corresponden a una prueba de lanzamiento en baloncesto realizada en 1 minuto. Calcule la media y la desviación estándar.

x	f
20	1
21	2
22	7
23	7
24	13
25	10
26	4
27	2
29	3

2. Introduzca los datos anteriores en un programa de computación. Compare los resultados con los obtenidos por medio de la calculadora.
3. ¿Qué significa para usted una media de 11,67, y una desviación estándar de 2,39 en una prueba de flexiones de codo en apoyo facial?
4. Catorce estudiantes de Fisiología del Ejercicio midieron la presión sanguínea de la misma persona cuyas lecturas se presentan a continuación (en mmHg). Si la presión sanguínea del sujeto permanece constante y los estudiantes aplicaron en forma apropiada la técnica de medición, ¿Cuál debería ser el valor de la desviación estándar?
138, 130, 135, 140, 120, 125, 120, 130, 130, 144, 143, 140, 130, 150
5. A continuación, se presentan los tiempos en una prueba de velocidad de dos secciones de cuarto grado. Compare la variación de los dos conjuntos de datos:
Sección A: 6.5, 6.6, 6.7, 6.8, 7.1, 7.3, 7.4, 7.7, 7.7, 7.7
Sección B: 4.2, 5.4, 5.8, 6.2, 6.7, 7.7, 7.7, 8.5, 9.3, 10.0

Capítulo 5. Conversión de puntajes

Los datos que surgen de la aplicación de las pruebas de condición física y/o de destrezas deportivas requieren de una interpretación objetiva, acorde con los procedimientos científicos y con las expectativas que tienen los evaluados de conocer cuál es su nivel de desempeño actual con respecto a sus marcas anteriores, los resultados de los demás, o de la calificación que obtendrá. Los puntajes que se obtienen al realizar una prueba (puntajes sin procesar, puntajes crudos, puntajes reales, o puntajes originales) por sí solos no tienen significado a menos que se les pueda comparar con algún tipo de valor, y esto sucede en cualquier área del conocimiento y con mediciones o registros tanto cuantitativos como cualitativos.

Por ejemplo, cuando los medios de comunicación en juegos de baloncesto comparan el rendimiento de varios jugadores o equipos, generalmente usan porcentajes en lugar de los datos originales. Se pueden ver los porcentajes de tiro lanzados de algunos jugadores para compararlos con los tiros libres, tiros de dos puntos y tiros de tres puntos convertidos. Los porcentajes se usan porque todos los entienden y permiten comparaciones inmediatas de dos o más jugadores.

Los datos sin procesar dificultarían saber quién es un mejor tirador de tiros libres por los tiros lanzados y convertidos: el jugador A, que hizo 46 de 62 intentos, o el jugador B, que hizo 56 de 78 (datos originales). Pero cuando los datos se convierten en porcentajes (74% para el jugador A, 72% para el jugador B), se pueden hacer fácilmente comparaciones entre jugadores y/o equipos.

Los puntajes estándar se utilizan para evaluar los individuos que han participado en varias pruebas en un determinado período de tiempo y las unidades de medición de cada prueba son diferentes. Si cada sujeto realiza las pruebas de abdominales en un minuto, flexibilidad en el cajón flexómetro y de capacidad aeróbica al correr una milla, ¿Cómo hace el docente para determinar cuál de los sujetos presentó el mejor desempeño? Un entrenador o un especialista en condición física para la salud que trabajen en clubes o empresas, se enfrentarán al mismo problema. Los tres puntajes obtenidos por cada sujeto no pueden procesarse (sumar, promediar) entre ellos porque la unidad de medición difiere de prueba a prueba: los abdominales se miden en repeticiones, la flexibilidad en centímetros, la carrera en segundos.

Este problema se elimina expresando todos los puntajes de las pruebas en una unidad de medición común o puntaje estándar. Ello implica convertir los puntajes de la prueba o puntajes originales en puntajes estándar y luego procesar los puntajes estándar que obtuvo cada sujeto. Por ejemplo, puede decirse que el individuo con la suma más alta de puntajes estándar o su promedio es el mejor de todos.

Si tenemos un estudiante que haga un salto largo de 4 metros y el promedio del grupo es de 5, se sabe que ese puntaje está por debajo de la media, pero ¿cuán lejos? Para evaluar un puntaje original se le compara con una escala donde se conozca las medidas de

tendencia central (media) y la variabilidad (desviación estándar) de la prueba. Después de aprender a calcular la desviación estándar de un conjunto de datos puede surgir una pregunta sobre su significado. La respuesta está en la comprensión de las características universales de una curva normal.

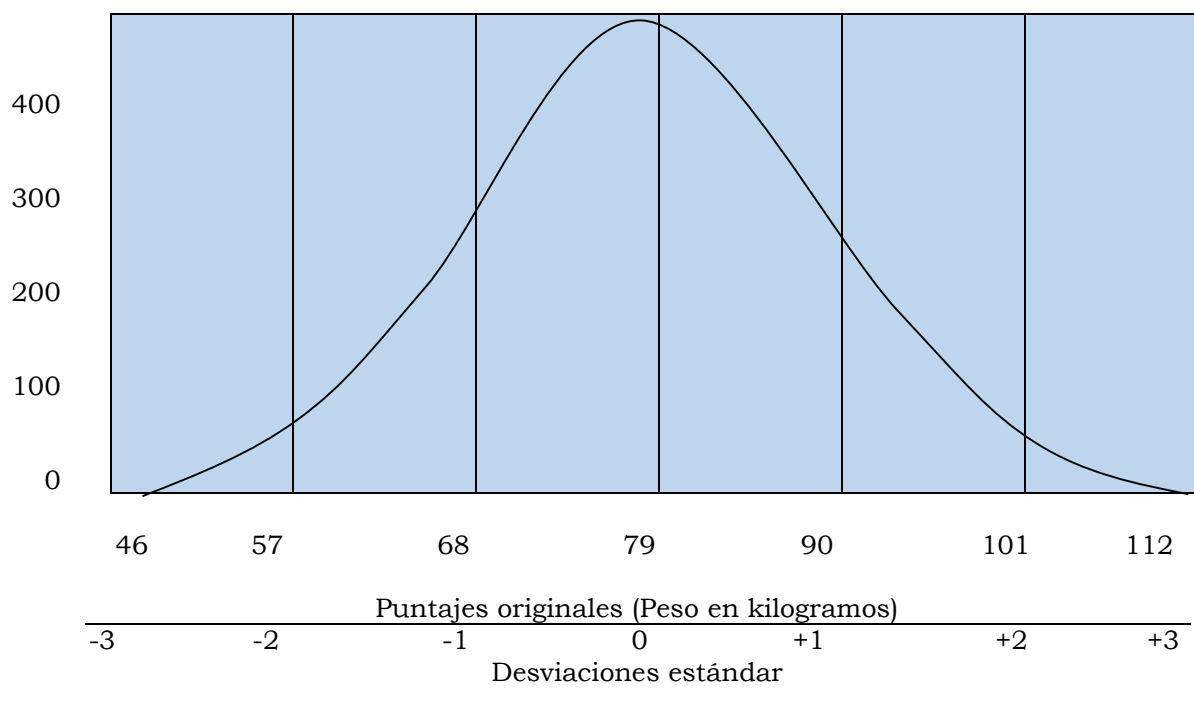


Figura 5.1 Relación entre los puntajes originales, la media y la desviación estándar

Al tener una suficiente cantidad de resultados, la mayor parte de las cualidades y habilidades que se midan asumirán una distribución normal. Si los datos están en una escala de intervalo o de razón, si se tienen suficientes sujetos medidos, y si la curva es normal o muy cercana a la normal, las características de normalidad son aplicables a los datos.

Para ejemplificar esto, supongamos que tenemos la medición del peso de más de 1000 personas del sexo masculino en edad escolar. La media del peso fue de 79 kilogramos y la desviación estándar de 11 kilogramos. Si a la media (79) le sumamos una desviación estándar (11), tenemos 90 y así con las demás desviaciones, es decir, $M \pm 1s$; $M \pm 2s$ y $M \pm 3s$. La figura 5.1 muestra que la mayoría de los sujetos pesan entre 68 y 90 kilogramos (entre -1 y +1) y que la frecuencia más alta está en la media. Unos pocos sujetos pesan menos de 46 kilogramos y otros pocos pesan más de 112. El peso de la mayoría de los sujetos está alrededor de la media. Cuando los valores se alejan de la media en cualquier dirección, hacia los más pesados o los menos pesados, disminuye progresivamente la cantidad de sujetos en cada desviación.

La curva normal es una curva simétricamente bilateral en forma de campana, centrada en un punto que es simultáneamente moda, mediana y media; ese punto es tanto el

puntaje más frecuente como el puntaje por debajo del cual cae la mitad de los puntajes. La curva normal tiene una media de 0 y una desviación estándar de 1.

Acá se presenta la siguiente regla empírica. En una distribución normal al sumar y restar de la media el valor de 1 desviación estándar, se obtiene un rango que abarca aproximadamente el 68% de los puntajes. Si a partir de la media se le suma y resta el valor de la desviación estándar en dos ocasiones se obtiene un rango que abarca aproximadamente el 95% de los puntajes, y al hacerlo sumando y restando el valor de la desviación estándar en tres ocasiones el rango abarca aproximadamente el 99,7% de los puntajes. Esto es cierto sin tener en cuenta la magnitud de la media y la desviación estándar, siempre que la distribución sea normal.

En resumen, la regla empírica se expresa de esta forma:

$M \pm 1s = 68,26\%$ de los puntajes

$M \pm 2s = 95,44\%$ de los puntajes

$M \pm 3s = 99,74\%$ de los puntajes

En la figura 5.2 se observa la distribución normal de los puntajes y la relación existente entre la media, la desviación estándar, el puntaje z, el puntaje T y los percentiles.

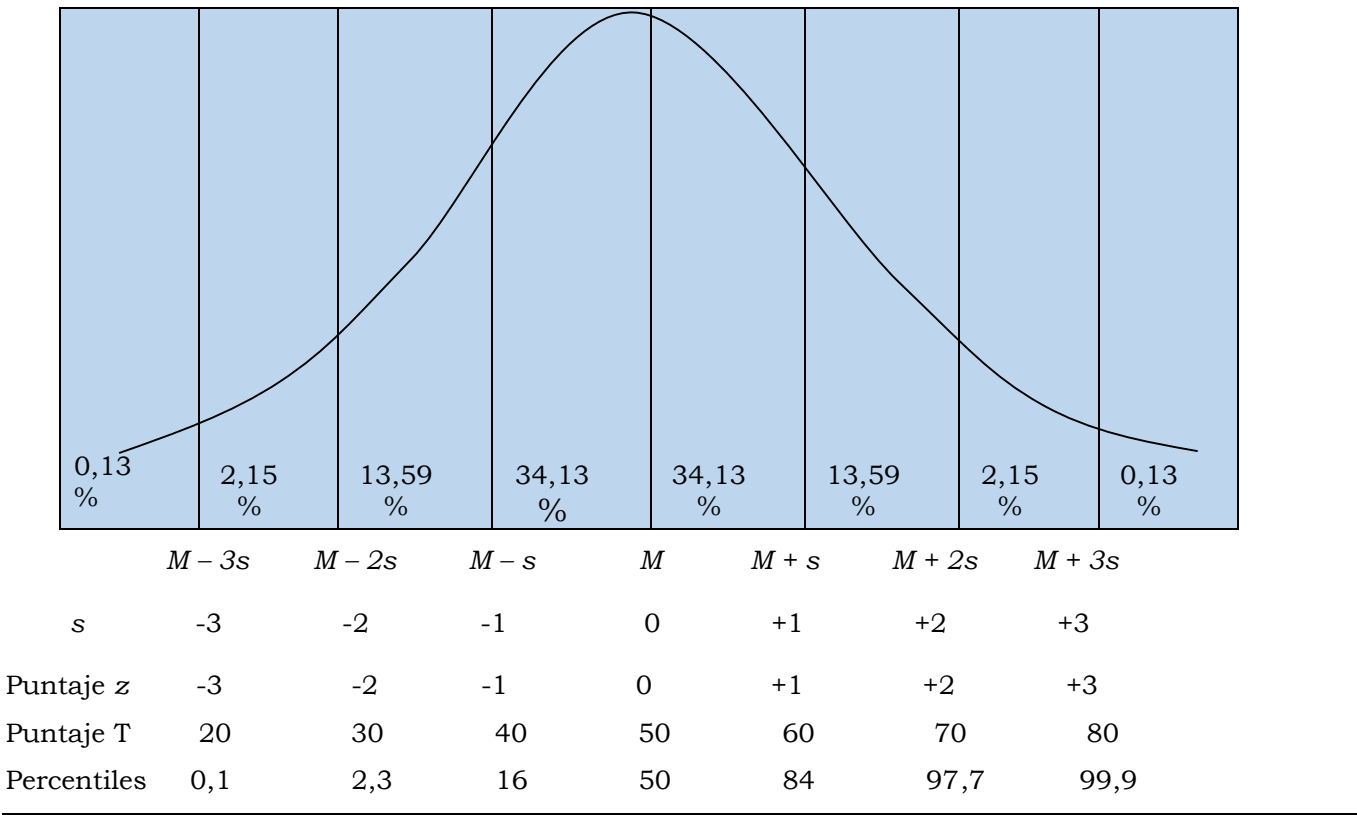


Figura 5.2. Porcentajes de distribución normal de los puntajes y puntajes estándar

Cuando se tienen varias pruebas (abdominales, flexibilidad, carreras de distancia, carreras de velocidad, etc.), los puntajes originales de estas pruebas no se pueden comparar directamente a menos que sean convertidos al mismo tipo de puntaje estándar. No es lógico comparar 50 abdominales con el tiempo de 4:36.3 en la carrera de una milla. ¿Cuál es mejor, un salto largo de 6,10 metros o un tiempo de 10,5 segundos en una carrera de 100 metros? No se puede responder a estas preguntas porque los valores referidos se hallan en diferentes unidades de medición. Para efectuar una comparación apropiada, se debe convertir los puntajes originales en alguno de los siguientes puntajes estándar: percentiles, puntajes z y puntajes T.

Revisemos, en primer lugar, la medición de la posición en el grupo. Hay dos técnicas básicas para procesar e interpretar la medición en términos de porcentajes: el rango percentil y los percentiles. En un principio, es mejor aprender los métodos utilizando pocos datos, papel, lápiz y calculadora, para luego utilizar los paquetes estadísticos en la computadora. Luego se examinará lo referente a puntajes z y puntajes T.

Rangos percentiles

Después de la aplicación de una prueba, todos los sujetos querrán saber su desempeño en comparación con los demás, es decir, el rango alcanzado. Ese rango puede calcularse asignándole al individuo que logró el mejor puntaje el rango 1, al individuo que logró la siguiente mejor marca el rango 2, y así sucesivamente. Pero este tipo de rango tiene poco significado a menos que se conozca el número de individuos en el grupo. Por ejemplo, un rango de 35 es muy bueno si hay 250 sujetos en el grupo, pero es muy malo si la cantidad de sujetos es 37. Los rangos percentiles que indican la posición relativa de un individuo en un grupo, también indica el porcentaje del grupo cuyo puntaje es más bajo que el de ese individuo.

El cálculo de un rango percentil (RP) tiene tres pasos. Primero, se elabora una distribución simple de frecuencias. Segundo, se calcula el rango percentil utilizando la siguiente fórmula:

$$RPx = \% \text{ acum. debajo de } x + \frac{\% \text{ de } x}{2}$$

donde RPx es el rango percentil del puntaje x , $\% \text{ acum. debajo de } x$ es el porcentaje acumulado por debajo del puntaje x , y $\% \text{ de } x$ es el porcentaje propio de x .

En el siguiente ejemplo de la tabla 5.1, determinemos el rango percentil del puntaje 6 de flexiones en apoyo facial realizadas por estudiantes del último año de bachillerato.

x	Frecuencia	Porcentaje	Porcentaje acumulado
0	1	4	4
3	3	12	16
4	6	24	40
5	10	40	80
6	4	16	96
10	1	4	100

Tabla 5.1 Distribución simple de frecuencias en la prueba de flexión y extensión de codos

Para calcular el rango percentil del puntaje 6 en esta prueba, el porcentaje acumulado por debajo de ese puntaje es 80 y el porcentaje del puntaje es 16, por lo que se sustituye la fórmula:

$$RP_6 = 80 + \frac{16}{2} = 80 + 8 = 88$$

Este resultado indica que el rango percentil del puntaje 6 en la prueba de flexión y extensión de codos en apoyo facial es 88, lo cual significa que el 88% de los sujetos hicieron una marca inferior a 6. El porcentaje del puntaje se divide entre dos por cuanto se asume el supuesto de que la mitad de los puntajes realmente alcanzaron un puntaje entre el límite real inferior y el valor correspondiente, por ejemplo, entre 5,5 y 6.

El rango percentil es un tipo de puntaje ordinal que solo indica una posición dentro del grupo; la diferencia en el rango percentil entre un valor y otro puede ser muy amplia por lo cual resulta inapropiado sumar, sustraer, multiplicar y dividir rangos percentiles. Por ejemplo, en la tabla 5.1 el rango percentil de 4 es 28 mientras el rango percentil de 5 es 60, lo que evidencia un cambio desproporcionado de uno a otro, con una diferencia de un 32% entre 4 y 5.

Si a los datos de la tabla 5.1 se le agrega una frecuencia resultando que dos sujetos lograron un puntaje de 10, ¿Cuál es el rango percentil del puntaje 3? Para establecer el porcentaje para cada valor, se procede a dividir cada frecuencia entre el número de sujetos y se divide entre 100; por ejemplo, para la frecuencia 1 la operación es la siguiente: $1/26 = 0.03846 * 100 = 3,846\% = 3,85\%$ (Tabla 5.2)

x	Frecuencia	Porcentaje	Porcentaje acumulado
0	1	3,85	3,85
3	3	11,54	15,39
4	6	23,08	38,47
5	10	38,46	76,93
6	4	15,39	92,32
10	2	7,69	100

Tabla 5.2 Distribución simple de frecuencias en la prueba de flexión y extensión de codos

De acuerdo con la fórmula el rango percentil del puntaje 3 es el siguiente, redondeando a número entero el resultado:

$$RP_3 = 3,85 + \frac{11,54}{2} = 3,85 + 5,77 = 9,62 = 10$$

Ahora bien, supongamos que en un equipo de baloncesto se hace una prueba de tiro libre con los resultados expuestos en la tabla 5.3, y el entrenador desea determinar la forma en que podría definir el equipo titular y los primeros cambios, de acuerdo con los resultados de esta prueba en la que participaron 15 atletas. En primer lugar, ¿Cuál es el rango percentil de 6 canastas? Al ordenar los puntajes de menor a mayor se observa que hay nueve lugares (sujetos) que van de los puntajes 1 a 6 canastas.

x
1
3
3
4
4
5
5
5
6
7
7
7
9
9
10

Tabla 5.3 Lanzamientos de tiro libre en baloncesto, convertidos en 10 intentos.

A continuación, se divide 9 entre 15 y el resultado se multiplica por 100, obteniéndose el percentil correspondiente.

$$\frac{9}{15} = 0,6 * 100 = 60$$

En esta prueba, quien haya logrado 6 canastas se encuentra en el rango percentil 60. En el caso de que con los mismos datos se desee conocer el puntaje correspondiente al rango percentil 60, se divide 60 entre 100 y se multiplica por 15; el resultado nos dice que es el participante que se halle en el noveno lugar de menor a mayor y, por tanto, quien haya convertido 6 lanzamientos.

$$\frac{60}{100} = 0,60 * 15 = \text{Lugar 9} = 6 \text{ lanzamientos}$$

Para determinar el rango percentil de los puntajes repetidos como los cinco lanzamientos convertidos de la tabla 5.3, se cuentan todos los puntajes que en este caso son 8, se divide entre 15 y se multiplica por 100, resultando que el atleta que convierta 5 lanzamientos se halla en el rango percentil 53:

$$\frac{8}{15} = 0,533 * 100 = 53,3 = 53$$

Asimismo, tomando en cuenta estos resultados el entrenador desea seleccionar los 2/3 superiores y eliminar el tercio inferior, para lo cual divide 1 entre 3 y lo multiplica por 15, para hallar la cantidad de sujetos con menor número de lanzamientos convertidos que serán eliminados.

$$\frac{1}{3} = 0,333 * 15 = 5$$

De acuerdo con este resultado los cinco últimos sujetos lograron entre 1 y 4 cestas por lo que la selección la conformarán quienes hayan logrados 5 o más cestas.

Percentiles

Cuando se utiliza el porcentaje, se compara un puntaje original con una escala conceptual de 100 puntos. Un percentil se define como un punto o posición en una escala continua de 100 divisiones teóricas de forma tal que una determinada fracción de los puntajes de la población cae o está por debajo de esa población.

Un puntaje en el percentil 75 es igual o superior a tres cuartos de los puntajes de un conjunto de datos. El percentil 33 es igual o superior que un tercio de los puntajes, pero es inferior a dos tercios de los puntajes. Igualmente, si se logra un percentil 90 significa simplemente que es mejor que el 90% de la gente que participó en la prueba e inversamente, si se logra llegar al percentil 10, el 90% de los participantes son mejores. El puntaje percentil de un estudiante en una prueba indica cómo el puntaje del estudiante se compara con los demás puntajes.

Los percentiles son puntajes estándar. Un puntaje estándar es un puntaje que se ha derivado de unos puntajes originales, reales, y que tiene una base conocida de comparación. En porcentajes, el centro es el 50% y el rango va del 0% al 100%. Si un hombre de edad mediana es capaz de consumir 40 mililitros de oxígeno por kilogramo de peso corporal por minuto, y este es su puntaje real que es difícil de evaluar por sí solo. Pero si se compara ese valor con otros valores de la población de hombres en edad similar y se establece que tiene un puntaje percentil de 65, entonces se concluye que el consumo de oxígeno de esa persona, 40 en este caso, es igual o mejor que el 65% de las personas de esa población. Los puntajes originales o reales se obtienen a través de la medición. Los puntajes estándar derivados desde esos puntajes originales, suministran más información para su evaluación que los puntajes originales.

Los puntajes estándar permiten: (a) evaluar los puntajes originales y (b) comparar dos conjuntos de datos que se basan en unidades diferentes de medición. En este caso, cuál es mejor puntaje, ¿lanzar una pelota de softbol a 45 metros o realizar 30 abdominales en un minuto? Con esta única información, es imposible decirlo. Debe preguntarse, ¿Cuál es el rango? ¿Cuál es el promedio?

Pero cuando se compara el percentil 57 en el lanzamiento de la pelota con el percentil 35 en abdominales, la respuesta es clara. No sólo se determinan los valores relativos de los puntajes, sino que también se conoce el puntaje medio (50) y el rango (0-100). El estudiante que obtuvo esos puntajes es mejor en el lanzamiento que en los abdominales cuando se le compara con los demás estudiantes.

A continuación, se presentan dos formas de calcular los percentiles a partir de la distribución de frecuencias, considerando en primer lugar los porcentajes en cada frecuencia y el porcentaje acumulado. El segundo procedimiento se realiza tomando en cuenta tanto la frecuencia como la frecuencia acumulada.

Primer procedimiento. Las pruebas estándar o normas por lo general se expresan en rangos percentiles que son múltiplos de 5 (5, 10, 15...). Para desarrollar una norma primero se debe determinar el puntaje de la prueba, o percentil, que corresponde a cada RP. Con la siguiente fórmula se calcula el puntaje de la prueba que se corresponde con cada rango percentil:

$$P = lri + \left(\frac{RP - \% \text{ acum. debajo de } x}{\% \text{ de } x} \right) (UM)$$

donde P es el puntaje de la prueba (percentil) que corresponde a RP , x es el puntaje que tiene un porcentaje acumulado menor al valor de RP , lri es el límite real inferior, $\% \text{ acum. debajo de } x$ hasta el puntaje inferior a x , $\% \text{ de } x$ es el porcentaje que corresponde al puntaje, y UM es la unidad de la medición de los datos.

Para la aplicación de la fórmula, primero se construye una tabla de distribución simple de frecuencias, de menor a mayor. Luego se localiza el $\%$ acumulado inmediatamente menor al RP que se busca aplicando la fórmula señalada. En la tabla 5.4 se presenta un ejemplo con la prueba de abdominales en 30 segundos.

<i>x</i>	Frecuencia	Porcentaje	% acumulado
1	1	2	2
5	1	2	4
6	2	4	8
7	5	10	18
8	6	12	30
9	4	8	38
10	7	14	52
11	10	20	72
13	4	8	80
14	4	8	88
15	3	6	94
18	3	6	100

Tabla 5.4 Distribución simple de frecuencias de la prueba de abdominales

Para ejemplificar la aplicación de la fórmula, se determinará el puntaje correspondiente al rango percentil 45 localizando en la columna de % acumulado determinándose que dicho RP se encuentra a nivel del puntaje 10. En primer lugar, el puntaje 10 es el primer puntaje posterior al RP 45. Luego se obtienen los valores:

$$\begin{aligned}
 x &= 10 \\
 lri &= 9,5 \\
 \% \text{ acum. debajo de } x &= 38 \\
 \% \text{ de } x &= 14 \\
 UM &= 1 \\
 RP &= 45
 \end{aligned}$$

Sustituyendo en la fórmula tenemos:

$$P = 9,5 + \left(\frac{45 - 38}{14} \right) (1) = 9,5 + \left(\frac{7}{14} \right) (1) = 9,5 + 0,5 = 10$$

Tomando los datos contenidos en la tabla 2.7 se han establecido los percentiles del 1 al 100, múltiplos de 5, con lo cual se construye la norma para la prueba de salto largo sin impulso como se presenta a continuación en la tabla 5.5:

Percentil	Valor
100	230
95	213
90	208
85	198
80	194
75	193
70	191
65	183
60	179
55	173
50	173
45	170
40	166
35	165
30	159
25	152
20	147
15	142
10	138
5	122
1	117

Tabla 5.5 Percentiles de la prueba de salto longitudinal sin impulso

En el caso de las pruebas donde el menor puntaje es la mejor marca como en las pruebas donde se deben lograr los menores tiempos posible (velocidad, agilidad, la carrera de la milla y similares, porcentaje de grasa, etc.) es erróneo establecer los percentiles de la misma manera como se presentan en la tabla 5.5. Por lo tanto, los percentiles deben ser corregidos de forma tal que, por ejemplo, el percentil 100 se corresponda con el menor valor. Para ello se utiliza la siguiente fórmula:

$$\text{Percentil corregido} = 100 - \text{Percentil calculado}$$

En consecuencia, tomando los datos de la tabla 2.9 de la prueba de velocidad, se procede a calcular los percentiles aplicando la fórmula señalada y luego se procede a corregirlos. Veamos el cálculo del percentil 45 y su corrección.

$$\begin{aligned}
 x &= 7,80 \\
 lri &= 7,795 \\
 \% \text{ acum.debajo de } x &= 44 \\
 \% \text{ de } x &= 2 \\
 UM &= 0,01 \\
 RP &= 45
 \end{aligned}$$

Con estos datos se calcula el percentil correspondiente:

$$P = 7,795 + \left(\frac{45 - 44}{2}\right)(0,01) = 7,795 + \left(\frac{1}{2}\right)(0,01) = 7,795 + 0,005 = 7,80$$

Procediendo a corregir el percentil tenemos lo siguiente:

$$\text{Percentil corregido} = 100 - 45 = 55$$

$$P_{55} = 7,80$$

De acuerdo con los datos de la tabla 2.9, se presentan los percentiles del 1 al 100 calculados de la prueba de velocidad y la consecuente corrección para asignarle al mejor tiempo el percentil más alto (Tabla 5.6).

Percentil calculado	Valor	Percentil corregido	Valor
100	8,97	100	7,09
95	8,57	95	7,14
90	8,37	90	7,21
85	8,25	85	7,27
80	8,13	80	7,34
75	8,00	75	7,40
70	7,98	70	7,49
65	7,95	65	7,77
60	7,94	60	7,77
55	7,85	55	7,80
50	7,84	50	7,84
45	7,80	45	7,85
40	7,77	40	7,94
35	7,77	35	7,95
30	7,49	30	7,98
25	7,40	25	8,00
20	7,34	20	8,13
15	7,27	15	8,25
10	7,21	10	8,37
5	7,14	5	8,57
1	7,09	1	8,97

Tabla 5.6 Percentiles calculados y percentiles corregidos de la prueba de velocidad

Segundo procedimiento. Los percentiles se calculan construyendo la distribución de frecuencias incluyendo las columnas de frecuencia y frecuencia acumulada. Si los datos presentan un rango menor a 20 valores, se procede a construir una distribución simple de frecuencias. En este caso, supongamos que un entrenador de baloncesto evaluó un grupo de 60 alumnos en la prueba de tiro libre, con un máximo de 10 lanzamientos efectivos (Tabla 5.7).

<i>x</i>	<i>f</i>	<i>fa</i>
1	1	1
2	1	2
3	2	4
4	4	8
5	8	16
6	15	31
7	12	43
8	9	52
9	6	58
10	2	60
<i>n</i> =60		

Tabla 5.7 Distribución simple de frecuencias del lanzamiento de tiro libre en baloncesto.

Si se quiere determinar, por ejemplo, el rango percentil de quien haya alcanzado la cantidad de 7 lanzamientos convertidos, es necesario conocer la cantidad de sujetos que lograron 7 o menos cestas. En la columna de frecuencias acumuladas se observa que 43 sujetos hicieron la cantidad de 7 o menos de un total de 60 participantes.

Se divide la cantidad de sujetos con 7 o menos canastas entre el número total de participantes, y este resultado se multiplica por 100 para obtener el rango percentil.

$$\frac{43}{60} = 0,716 * 100 = 71,6 = 72$$

Como conclusión tenemos que quien haya logrado encestar en 7 oportunidades, es igual o mejor que el 72% de los integrantes del grupo.

Para determinar el puntaje correspondiente a un percentil dado, por ejemplo, al percentil 75, primero se establece el número de sujetos que están en el 75%. Para ello se divide 75 entre 100 y el resultado se multiplica por *N*, en este caso 60.

$$\frac{75}{100} = 0,75 * 60 = 45$$

Al ubicar los 45 sujetos en la columna de frecuencia acumulada, tenemos que 43 estudiantes hicieron 7 cestas o menos y 52 estudiantes hicieron 8 cestas o menos, por lo que cada uno de los 9 estudiantes que hicieron 8 cestas se ubican entre los lugares 44 y 52 y el estudiante 45 es uno de esos 9. Ya que los 9 participantes que hicieron 8 cestas no se pueden separar para definir cuál de ellos es el número 45, decimos que el percentil 75 se corresponde con 8 anotaciones; todo aquel que sea efectivo en 8 oportunidades alcanza el percentil 75.

Cuando la cantidad de sujetos y el rango son superiores a 20 se construye una distribución agrupada de frecuencias; veamos un ejemplo en softbol:

<i>x</i>	<i>f</i>	<i>fa</i>
80-89	4	4
90-99	5	9
100-109	6	15
110-119	7	22
120-129	10	32
130-139	9	41
140-149	7	48
150-159	15	63
160-169	10	73
170-179	12	85
180-189	9	94
190-199	7	101
200-209	7	108
210-219	5	113
220-229	2	115
<i>n</i> = 115		

Tabla 5.8 Prueba de lanzamiento en softbol, medido en pies

Para calcular el percentil de un puntaje dado, previamente se debe conocer el grupo en el cual se encuentra ese puntaje. Podemos preguntarnos, ¿cuál es el rango percentil de un sujeto que haya lanzado 146 pies? Según la tabla 5.8 hubo siete sujetos que se encuentran en el grupo 140 – 149 por lo que se asume que los siete están equitativamente entre los límites reales 139,5 – 149,49; para solucionar el problema se aplica la siguiente fórmula:

$$P = \frac{\left(\frac{x - lri}{i}\right) f + fa}{n}$$

donde x es el puntaje original, lri es el límite real inferior del grupo en el que el puntaje cae, i es el tamaño del intervalo, f es la frecuencia del grupo donde cae el puntaje, fa es la frecuencia acumulada hasta el grupo anterior al del puntaje, y n es el número de sujetos. Los datos para solucionar el problema son los siguientes:

$$\begin{aligned}x &= 146 \\lri &= 139,5 \\i &= 10 \\f &= 7 \\fa &= 41 \\n &= 115\end{aligned}$$

Sustituyendo en la fórmula tenemos:

$$P = \frac{\left(\frac{146 - 139,5}{10}\right) 7 + 41}{115}$$

$$P = \frac{\left(\frac{6,5}{10}\right) 7 + 41}{115} = \frac{(0,65)7 + 41}{115} = \frac{4,55 + 41}{115} = \frac{45,55}{115} = 0,396$$

$$P = 0,396 * 100 = 39,6 \%$$

En conclusión, un lanzamiento de 146 pies está aproximadamente en el percentil 40 de la prueba de lanzamiento en softbol.

Ahora bien, para determinar el puntaje correspondiente a un percentil dado, por ejemplo, el percentil 50, el docente o el entrenador puede hacerse esta pregunta: ¿Qué distancia debe lograr un estudiante para ubicarse en la mitad superior del grupo?

Para solucionar este problema primero se divide el percentil señalado entre 100 (50/100) y el resultado se multiplica por el número de sujetos en la prueba:

$$PN = \left(\frac{P}{100}\right) N = \left(\frac{50}{100}\right) 115 = 0,5 * 115 = 57,5$$

Este resultado ubica el percentil 50 en el grupo 150 – 159 de la tabla 5.8, conduciendo al cálculo del puntaje correspondiente al percentil según la siguiente fórmula:

$$X = lri + \left(\frac{PN - fa}{f}\right) i$$

Se procede a determinar el límite real inferior (*lri*), la frecuencia acumulada (*fa*) hasta el grupo anterior, la frecuencia (*f*) del grupo y el tamaño del intervalo (*i*), y se sustituyen en la fórmula:

$$X = 149,5 + \left(\frac{57,5 - 48}{15}\right) 10 = 149,5 + \left(\frac{9,5}{15}\right) 10 = 149,5 + (0,633)10 =$$

$$X = 149,5 + 6,333 = 155,833 = 156$$

En conclusión, el percentil 50 de la prueba de lanzamiento en softbol es de 156 pies, aproximadamente. A continuación, se presentan dos tablas en percentiles de cinco pruebas de condición física.

Tablas de rendimiento para estudiantes universitarios de Educación Física que egresan de la carrera

MASCULINO					
P	1 milla (segundos)	Abdominal (repeticiones)	Flexibilidad (cm)	Agilidad (segundos)	Panículos (mm)
100	318	66	90	10,22	49
95	350	55	79	11,04	52
90	363	55	77	11,22	57
85	373	52	75	11,46	66
80	383	51	73	11,63	69
75	395	50	71	11,93	74
70	407	49	69	12,02	81
65	416	49	68	12,51	87
60	425	47	66	12,98	99
55	434	45	63	13,19	102
50	444	45	61	13,38	113
45	452	44	60	13,50	117
40	459	43	59	13,84	131
35	463	42	57	14,40	144
30	475	41	55	15,14	152
25	487	40	52	15,61	157
20	505	39	49	16,09	164
15	517	35	47	16,36	168
10	539	33	44	16,66	178
5	575	30	39	17,41	195
1	673	29	33	19,52	252

Tabla 5.9 Percentiles de pruebas de condición física para estudiantes universitarios de Educación Física que egresan (Grupo masculino)

FEMENINO					
P	1 milla (segundos)	Abdomi- nal (repe- ticiones 1')	Flexibili- dad (cm)	Agilidad (segun- dos)	Paniculos (mm)
100	441	60	88	10,62	83
95	459	55	80	11,46	93
90	477	50	79	12,11	103
85	488	49	77	12,31	109
80	500	47	74	12,44	119
75	508	45	72	12,50	121
70	515	45	71	13,07	124
65	522	43	70	13,11	125
60	529	40	69	13,35	127
55	536	40	68	13,48	131
50	552	38	67	13,58	138
45	560	37	64	13,78	147
40	574	34	63	14,04	156
35	581	33	61	14,46	168
30	589	31	59	14,87	171
25	594	30	58	15,27	173
20	605	30	56	15,41	178
15	630	28	55	15,46	181
10	655	21	51	16,48	193
5	673	20	47	17,41	213
1	786	14	46	19,60	246

Tabla 5.10 Percentiles de pruebas de condición física para estudiantes universitarios de Educación Física que egresan (Grupo femenino)

La construcción de este tipo de tabla facilita la evaluación de los sujetos que se están estudiando, especialmente cuando la población es numerosa. Cuando se realice la prueba, inmediatamente se le asigna el percentil que se corresponde con el nivel de desempeño. Si una estudiante hace un tiempo de 570 segundos en la prueba de correr una milla, el percentil que le corresponde a ese tiempo es 45, por lo que al puntaje logrado en cada prueba se convierte a percentil y podrá combinarse con los demás percentiles, ya sea sumándolos o promediándolos, para obtener un puntaje general que represente su nivel actual de condición física el cual, finalmente, podrá evaluarse con base a normas o con base en criterios.

Se pueden comparar tipos de puntajes originales al convertirlos en percentiles. Si un sujeto hace un salto alto de 1,68 metros para estar en el percentil 75 y un tiempo de 11,5 segundos al correr 100 metros y queda en el percentil 80, entonces en la prueba de velocidad el sujeto es mejor que en la prueba de salto. Los puntajes originales pueden ahora compararse directamente porque han sido convertidos a la misma base común o estándar.

Puntajes z

Es importante conocer la posición de un puntaje con respecto a otros en un conjunto de datos. Por ejemplo, si un estudiante obtiene 30 de un total de 35 puntos, él pudiese querer saber cómo se compara su calificación con las calificaciones de los otros estudiantes del grupo. La media y desviación estándar de las calificaciones se pueden usar para calcular un puntaje z, que mide la posición relativa de una medición en un conjunto de datos.

El puntaje z muestral es una medida de posición relativa definida por la siguiente fórmula:

$$z = \frac{x - M}{s}$$

donde x es el puntaje obtenido en la prueba, M es la media de los datos, y s es la desviación estándar de esos datos. Un puntaje z mide la distancia entre una observación y la media, medidas en unidades de desviación estándar. Si la media es 25, la desviación estándar es 4 y el puntaje obtenido es 30, tenemos:

$$z = \frac{x - M}{s} = \frac{30 - 25}{4} = \frac{5}{4} = 1,25$$

El puntaje 30 está a 1,25 desviaciones estándar arriba de la media ($30 = M + 1,25s$). En otro ejemplo, si la desviación estándar en un levantamiento de pesas es 11 y $M = 79$, entonces 1 desviación estándar es equivalente a 11 kilogramos en la escala de los puntajes originales. Un puntaje de 90 kilogramos se encuentra a 11 kilogramos, es decir, a 1 desviación estándar por arriba de la media y es equivalente a un puntaje z de +1, mientras que un puntaje original de 68 kilogramos (11 kilogramos o 1 desviación estándar por debajo de la media) está en un puntaje z de -1. Siendo el puntaje z el número de desviaciones estándar que un valor x se encuentra por arriba o por debajo de la media, se puede señalar que un puntaje z es un valor estandarizado.

Cuando la media y la desviación estándar de cualquier conjunto de datos que se distribuyan normalmente son conocidas, para cualquier puntaje original se puede calcular z . Nótese que el puntaje z aumenta a medida que el desempeño se eleva por arriba de la media. Cuando los puntajes más bajos son mejores que los puntajes más altos, el puntaje z debe aumentar a medida que el desempeño cae por debajo de la media. De esta manera en los eventos de velocidad, agilidad tomando el tiempo o en cualquier tipo de prueba donde los puntajes más bajos son mejores que los puntajes más altos, se utiliza la siguiente fórmula para determinar el puntaje z :

$$z = \frac{M - x}{s}$$

En la primera fórmula de z , siempre que un valor sea menor que la media el puntaje z que le corresponde será negativo. En la segunda fórmula, siempre que un valor sea mayor que la media el puntaje z que le corresponde será negativo. En el ejemplo de levantamiento

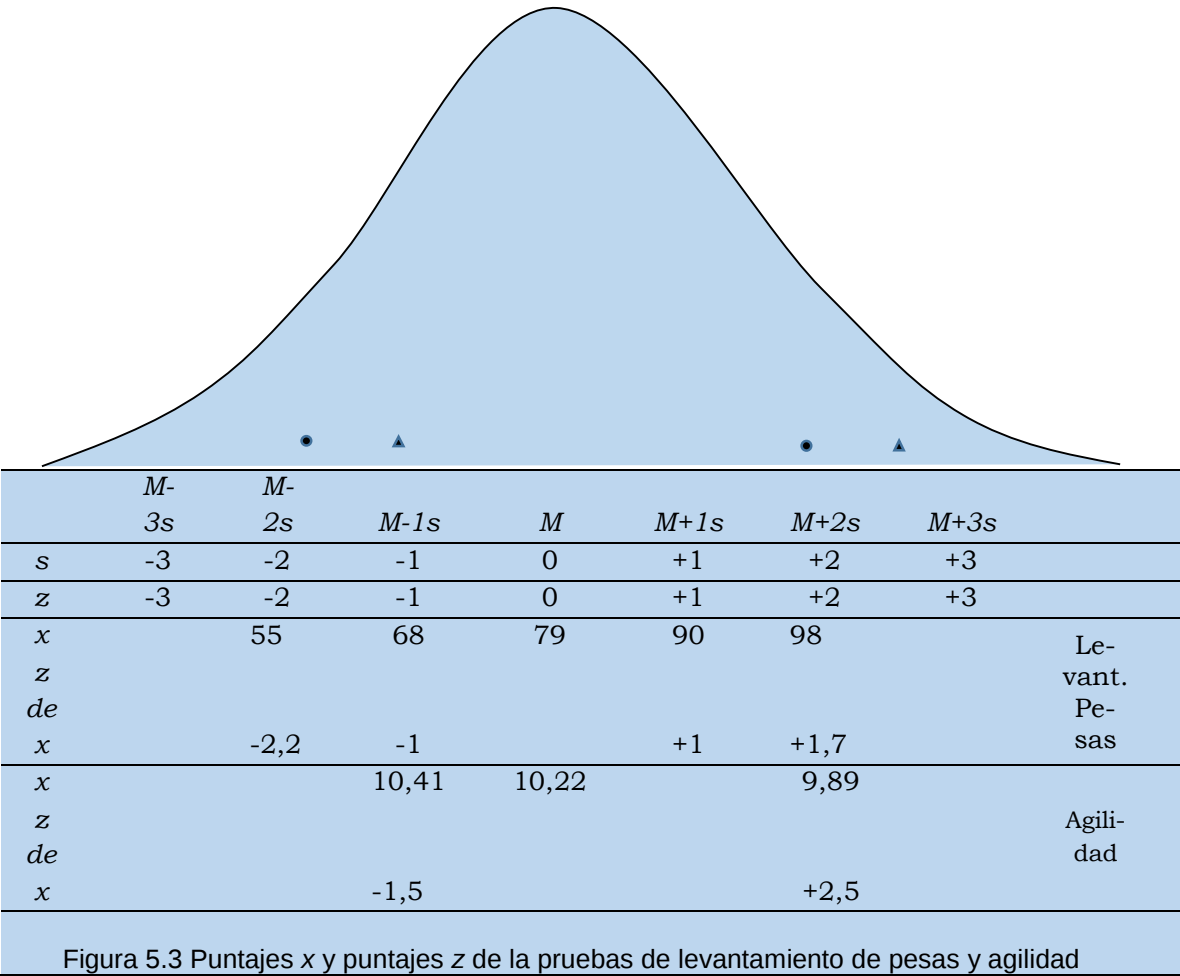
de pesas, tenemos la media $M = 79$, los puntajes $x = 90$ y $x = 68$, y la desviación estándar $s = 11$; los puntajes 68 y 90 coinciden con los puntajes z -1 y $+1$, respectivamente. Ahora, si un sujeto levanta 55 kg y otro 98, el puntaje z para cada uno serán los siguientes:

$$z = \frac{55 - 79}{11} = -2,2 ; z = \frac{98 - 79}{11} = 1,7$$

En una prueba de agilidad, con $M = 10,22$, $s = 0,13$, y los puntajes 10,41 y 9,89, los puntajes z serán:

$$z = \frac{10,22 - 10,41}{0,13} = -1,5 ; z = \frac{10,22 - 9,89}{0,13} = 2,5$$

Obsérvense los puntajes z en la figura 5.3; los puntos en la curva indican la posición de los puntajes z de levantamiento de pesas y los triángulos la ubicación de los puntajes z de la prueba de agilidad.



Para observar mejor la aplicación del puntaje z se trae un ejemplo que compara la estatura, en metros, de dos personajes públicos. Con 1,91 metros Lyndon Johnson fue el presidente de Estados Unidos más alto del siglo XX. Con 2,16 metros Shaquille O'Neal era el jugador más alto del equipo de baloncesto Miami Heat. ¿Quién es relativamente más alto, Lyndon Johnson entre los presidentes del siglo pasado, o Shaquille O'Neal entre los jugadores de su equipo Miami Heat? La estatura media de los presidentes del siglo pasado fue de 1,8161 metros con una desviación estándar de 0,05334. Los jugadores de baloncesto del equipo de Miami Heat tienen una estatura media de 2,032 metros, con una desviación estándar de 0,08382.

Solución: Las estaturas de los presidentes y de los jugadores de baloncesto provienen de poblaciones muy diferentes; para compararlas es necesario estandarizar las estaturas convirtiéndolas en puntajes z .

Lyndon Johnson	$z = \frac{x - M}{s}$	$z = \frac{1,91 - 1,8161}{0,05334}$	$z = 1,67$
Shaquille O'Neal	$z = \frac{x - M}{s}$	$z = \frac{2,16 - 2,032}{0,08382}$	$z = 1,52$

Interpretación: La estatura de Lyndon Johnson está a 1,67 desviaciones estándar por arriba de la media, mientras que la estatura de Shaquille O'Neal está a 1,52 desviaciones estándar por arriba de la media. La estatura de Lyndon Johnson, entre los presidentes del siglo pasado, es relativamente mayor que la estatura de Shaquille O'Neal entre los jugadores de baloncesto del equipo de Miami Heat. Shaquille O'Neal es mucho más alto que Lyndon Johnson, pero este es relativamente más alto cuando lo comparamos con sus colegas.

Puntajes T

Dado que los puntajes z se expresan en fracciones y pueden ser negativos, muchos profesionales prefieren utilizar un tipo de puntaje estándar denominado puntaje T cuando se trata de combinar diferentes tipos de puntajes originales. Los puntajes T se redondean al valor entero más cercano y muy rara vez son negativos. La media de los puntajes T es 50 y su desviación estándar es 10. La fórmula para calcular los puntajes T es la siguiente:

$$T = \frac{10(x - M)}{s} + 50$$

Obsérvese que el término $(x - M) / s$ en la fórmula de puntaje T se utiliza para calcular el puntaje z . Entonces, la fórmula de puntaje T se puede expresar como sigue:

$$T = 10z + 50$$

El puntaje equivalente en puntaje T para un puntaje z de + 1,5 es 65 ($T=10 * 1,5 + 50 = 65$); el puntaje T correspondiente a un puntaje z de -0.50 es 45 ($T=10 * (-0.50) + 50$).

La escala de puntaje T fue creada cuando el público en general tuvo dificultades para comprender el puntaje z, ya que no es común que se piense en un 0 como la media para un conjunto de datos. Mucha gente considera que el 0 no tiene significado y prefiere tener un 50 como el punto medio de un rango que va del 0 al 100. Por ello es que los percentiles son ampliamente usados, porque son más fáciles de comprender.

Una escala en puntaje T tiene una media de 50 y un rango del 0 al 100, pero es muy improbable que un puntaje T vayan más allá de los puntajes z -3 o +3, ya que sólo el 0.26% de los participantes pudiesen estar más allá de los límites de la escala z. En efecto, un puntaje T de 100 correspondería a un puntaje z de +5 o un percentil de 99.99999 lo cual es muy improbable que ocurra, por lo que generalmente se reflejan los puntajes T de 20 a 80.

Supongamos que en una prueba de salto largo sin impulso se establezca una media de 87 centímetros con una desviación estándar de 2,35 y se quiera determinar el puntaje T correspondiente a un salto de 90 centímetros.

Sustituyendo esos valores en la fórmula tendríamos:

$$T = \frac{10(x - M)}{s} + 50 = \frac{10(90 - 87)}{2,35} + 50$$

$$T = \frac{10 * 3}{2,35} + 50 = \frac{30}{2,35} + 50 = 12,76 + 50 = 62.76 = 63$$

Como ya se explicó en el puntaje z, a medida que el puntaje T aumenta el desempeño se eleva por arriba de la media. Cuando los puntajes más bajos son mejores que los puntajes más altos, el puntaje T debe aumentar a medida que el desempeño cae por debajo de la media. De esta manera en los eventos de velocidad o en cualquier tipo de prueba donde los puntajes más bajos son mejores que los puntajes más altos, se utiliza la siguiente fórmula para determinar el puntaje T:

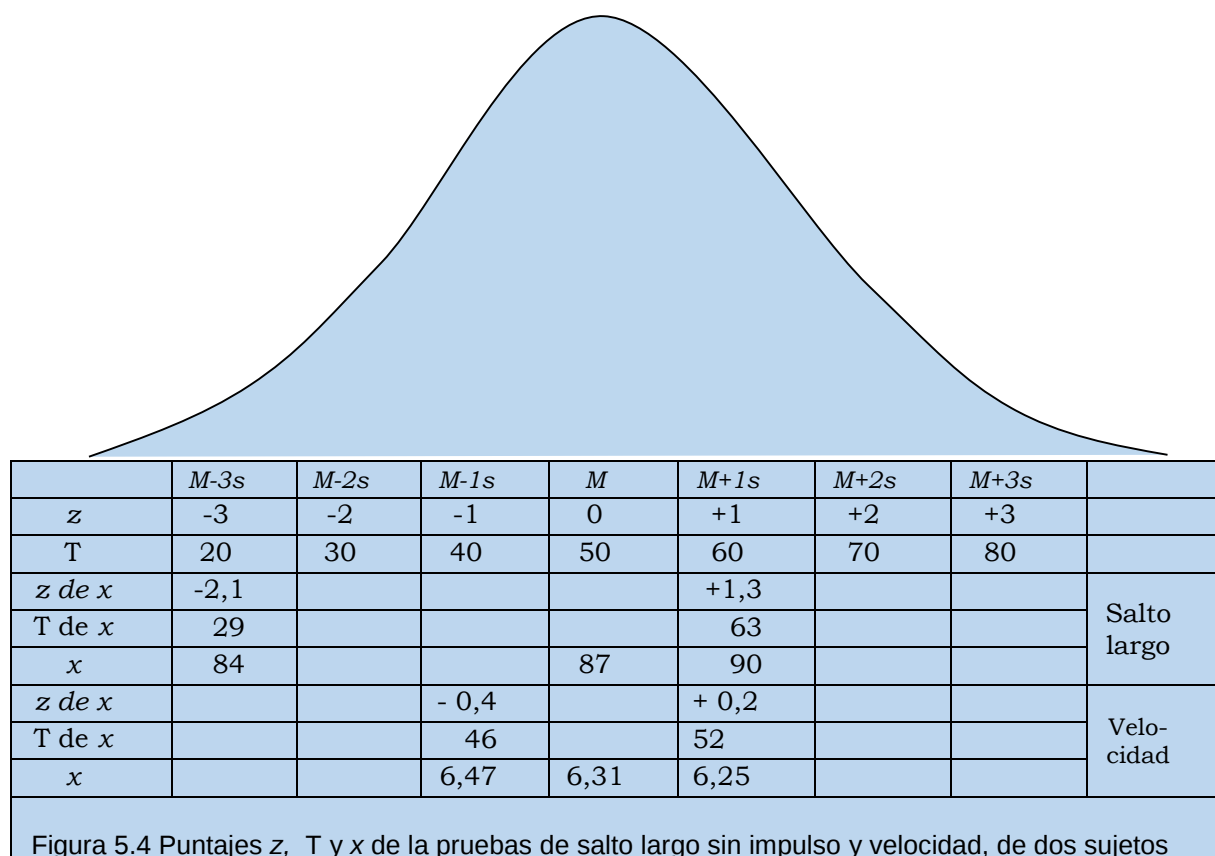
$$T = \frac{10(M - X)}{s} + 50$$

Para ejemplificar, si en una prueba de velocidad en 50 metros se determina una media de 6,31 y una desviación estándar de 0,38 y se quiere calcular el puntaje T de un tiempo de 6,47, tendríamos:

$$T = \frac{10(6,31 - 6,47)}{0,38} + 50$$

$$T = \frac{10(-0,16)}{0,38} + 50 = \frac{-1,6}{0,38} + 50 = -4,211 + 50 = 45,789 = 46$$

En la figura 5.4 se observa la correspondencia entre los puntajes z y los puntajes T tomando como referencia puntajes de las pruebas de salto largo sin impulso y de velocidad en 50 metros. En la prueba de salto largo sin impulso la media es 87 cm, la desviación estándar es 2,35, y los puntajes son 90 ($z = +1,3$ y $T=63$) y 84 ($z = -2,1$ y $T=29$). En la prueba de velocidad en 50 metros la media es 6,31, la desviación estándar es 0,38, y los puntajes son 6,47 ($z = -0,4$ y $T=46$) y 6,25 ($z = +0,2$ y $T=52$). Si se supone que estos puntajes les corresponden a dos sujetos, S_1 saltó 90 cm y corrió en 6,47 segundos, y le corresponden los puntajes T_{63} y T_{46} respectivamente, para un promedio T de 55. S_2 saltó 84 cm y corrió en 6,25, y le corresponden los puntajes T de 29 y 52 respectivamente, para un promedio T de 41. En el momento de la aplicación de las pruebas, en sujeto 1 demostró estar mejor preparado.



También se pueden sumar los puntajes T para determinar y diferenciar la condición física de un grupo de sujetos. A continuación, se presenta la suma de los puntajes T en tres pruebas de condición física (carrera de 1 milla medida en segundos, flexiones de codo en suspensión medidas en repeticiones, abdominales medidos en repeticiones) demostrando que sólo a través de los puntajes estándar que se derivan de los puntajes originales pueden combinarse los resultados de pruebas con diferentes unidades de medida. En este ejemplo

los puntajes T se combinan sumándolos; también se pueden promediar o darle mayor peso a alguna de las pruebas de acuerdo con el criterio de los evaluadores.

N	Ca- rrera 1 mi- lla	T	Flexio- nes de codo	T	Abdomina- les	T	Suma T
1	407	64	0	43	43	52	159
2	511	51	4	64	45	54	169
3	478	55	2	54	50	60	169
4	525	49	1	48	51	61	158
5	480	55	4	64	53	64	183
6	440	60	8	85	51	61	206
7	519	50	1	48	39	47	145
8	488	54	2	54	43	52	160
9	510	51	0	43	30	37	131
10	456	58	3	59	47	57	174
11	603	40	1	48	38	46	134
12	495	53	0	43	40	48	144
13	435	60	1	48	45	54	162
14	588	42	1	48	29	36	126
15	630	37	0	43	29	36	116
16	638	36	0	43	42	51	130
17	511	51	0	43	40	48	142
18	721	26	0	43	28	35	104
19	482	55	0	43	52	62	160
20	630	37	0	43	30	37	117
21	360	69	1	48	54	65	182
22	540	48	0	43	30	37	128

Tabla 5.11 Puntajes originales de pruebas de condición física, su conversión a puntaje T y la suma en puntajes T.

EJERCICIOS

1. De acuerdo con los siguientes puntajes de una prueba de precisión del servicio en tenis de campo y al segundo procedimiento explicado anteriormente, calcule (a) El rango percentil 7 y 10; (b) Los percentiles 45 y 70.

x
4
4
5
7
8
8
9
10
12
12
15

2. La misma prueba fue aplicada en tres secciones de un plantel educativo, a un grupo de 70 participantes. Calcule los percentiles 29, 38, 54 y 71, introduciendo los datos en un programa de computación.

x	f
4	1
5	3
6	3
7	5
8	8
9	7
10	10
11	8
12	6
13	5
14	5
15	3
16	3
17	2
18	1

3. Considerando el primer procedimiento para el cálculo de percentiles, señale el puntaje correspondiente a los percentiles 12, 28, 41, 66, 83 y 97 de los datos que se presentan a continuación.

N	Velocidad (1/100)	Carrera de 1 mi- lla (segundos)	Abdominales (Repeticiones)	Lanzamiento (Centímetros)
1	6,77	584	42	1115
2	6,18	525	58	1243
3	6,41	593	52	1022
4	6,29	577	39	994
5	6,70	498	48	1043
6	6,84	555	41	1122
7	6,44	593	49	999
8	6,17	574	40	1060
9	6,59	499	50	1081
10	6,70	524	57	1210
11	6,48	605	46	1095
12	6,03	537	51	980
13	6,25	571	44	1062
14	6,23	559	50	1058
15	6,55	568	48	1081
16	6,28	542	45	1027
17	6,52	577	40	1091
18	6,64	580	54	1194
19	6,60	516	44	1060
20	6,48	552	52	1159
21	6,09	580	56	1095
22	6,62	555	41	1113
23	6,33	526	48	1073
24	6,52	568	53	1397
25	6,48	570	57	1081
26	6,09	577	53	1066
27	6,28	580	45	1062
28	6,33	559	47	1244
29	6,80	568	52	1073
30	6,52	491	51	1122

4. Dada una media de 25,7 y una desviación estándar de 5,2 de una prueba de flexibilidad, calcular el puntaje z que les corresponde a los siguientes puntajes originales: 21,6 - 28,9 - 24,5; ¿Cuál es la interpretación de los puntajes z ?
5. Calcule el puntaje T de los puntajes originales del problema anterior.
6. En una prueba de velocidad en 100 metros planos el puntaje original es 11,3, la media es 11,6 y la desviación estándar es 0,55; calcule el puntaje T .

7. ¿Cuál de los dos sujetos tiene un mejor puntaje general?

	100 metros planos	Carrera de 600 metros
María	10.50 segundos	2 minutos
Luisa	11.10 segundos	1 minuto, 35 segundos
Media	10.74 segundos	2 minutos
Desviación estándar	1.12 segundos	20 segundos

8. En un equipo de baloncesto infantil la media de los jugadores es de 170 cm y la desviación estándar es de 5 cm. López, uno de los jugadores, mide 180 cm.

- ¿Cuál es la diferencia entre el jugador López y la media?
- La diferencia hallada, ¿A cuántas desviaciones estándar se encuentra?
- Convierta la estatura de López en puntaje z.
- Si se considera que por lo general los puntajes se hallan entre las puntuaciones $z - 2$ y $+2$, ¿Cómo se interpreta el puntaje z de López?

Capítulo 6. Correlación y Regresión

Hay muchas situaciones en las que el docente de educación física, el especialista en ejercicio o el fisioterapeuta desearían conocer la relación entre los puntajes de dos pruebas diferentes. Por ejemplo, si se determina que la velocidad y el rendimiento en una habilidad deportiva están relacionados, el docente de educación física podría tratar de mejorar la velocidad de los estudiantes de menor rendimiento. O, si se descubriera que el peso y la fuerza están relacionados, el especialista en ejercicio o el fisioterapeuta pueden querer utilizar un estándar de evaluación diferente para cada categoría de peso. Conocer la relación entre los puntajes también puede conducir a una mayor eficiencia del programa de medición. Por ejemplo, si hay siete pruebas en una batería de condición física y dos de ellas están muy relacionadas, la batería podría reducirse a seis pruebas sin pérdida de información. Utilizamos dos técnicas diferentes para determinar las relaciones de puntuación: una técnica gráfica y una técnica matemática llamada correlación.

Asimismo, si en una clase los estudiantes aprenden que el tiempo que toma correr 1.5 millas está directamente relacionado con el consumo de oxígeno medido en un laboratorio por $VO_2\text{max}$ (ml/kg/min), y el docente les explica que los sujetos con los valores más altos de $VO_2\text{max}$ tienden a tener el mejor tiempo en la carrera de 1.5 millas, entonces y en términos estadísticos, el docente les está diciendo que estas dos variables están correlacionadas. Saber qué tan rápido alguien puede completar una carrera de 1.5 millas permitirá el cálculo (usando una técnica llamada regresión lineal bivariada) de una estimación del $VO_2\text{máx}$ de la persona. Esto es más útil cuando una variable (carrera de 1.5 millas) es mucho más fácil de medir que una variable relacionada ($VO_2\text{max}$). Adelante veremos cómo hacer esta predicción.

Correlación

La correlación se define como el coeficiente numérico que indica el grado en el cual dos variables están relacionadas o asociadas. Técnicamente hablando, la correlación es el grado en el cual la dirección y tamaño de las desviaciones de la media en una variable se relacionan con la dirección y tamaño de la media de otra variable. Esta técnica se denomina coeficiente de correlación producto momento de Pearson. La correlación es bastante utilizada en los estudios de la actividad física, por ejemplo, al calcular la correlación entre estatura y peso, la correlación entre los panículos de grasa y el porcentaje de grasa corporal, o la correlación entre la cantidad de remates de tenis en voleibol practicados y la cantidad de remates efectivos en competencia.

El coeficiente o número que representa la correlación siempre estará entre +1.00 y -1.00; una correlación perfecta positiva (+1.00) existiría si los resultados de cada sujeto varían a igual distancia de la media en la misma dirección (medida en puntajes z) en dos variables diferentes. Por ejemplo, si cada sujeto que obtuvo un puntaje de $1z$ sobre la media en la variable X también obtuvo un puntaje sobre la media de $1z$ en la variable Y , el

resultado de la correlación debería ser +1.00; igualmente, si se obtienen puntajes por arriba o por debajo de la media en la variable X, pero a igual distancia en sentido opuesto en la variable Y, el resultado de la correlación debería ser -1.00.

Si el resultado de la correlación es similar pero no perfecta, el coeficiente es menor a +1.00, tal como 0.90 o 0.80 en el caso de que sea positiva, o mayor de -1.00, tal como -0.90 o -0.80 en el caso de que sea negativa. Un coeficiente de correlación de 0.00 significa que no hay relación entre las variables. Las correlaciones positivas se obtienen cuando los sujetos reciben altos puntajes numéricos en las dos variables. Por ejemplo, hay una correlación positiva entre el número de tiros libres realizados en el entrenamiento y el porcentaje de tiros libres convertidos en una temporada, por lo que se debe considerar que los jugadores que tienen mucha práctica tienden a tener un alto porcentaje de tiros libres convertidos en juego y deben ser los primeros en ser seleccionados. Por otra parte, las correlaciones negativas se obtienen cuando el puntaje en una variable tiende a ser alto mientras el puntaje de la otra variable tiende a ser bajo.

Negativa				Positiva				
-1.0	>	-0.7	0.5	0.00	0.5	0.7 a	>0.9	1.0
	-	a	a		a	0.8		
	0.9	-0.8	-		0.7			
		0.7						
Per- fecta	Alta	Mo- de- rada	Baja	Cero	Baja	Mo- de- rada	Alta	Per- fecta

Tabla 6.1 Rango de la correlación

¿Qué significa una correlación de 0.9, o de 0.85, o de 0.30? Uno de los principales objetivos de la correlación es predecir una variable a partir de los valores de otra variable para un sujeto determinado. Para propósitos predictivos una correlación más baja de 0.7 puede producir grandes errores inaceptables para una predicción individual, especialmente si la desviación estándar de cualquiera de las variables es muy amplia. Como regla general una correlación entre 0.5 y 0.7 se considera baja, de 0.7 a 0.8 es moderada, y de 0.9 en adelante es alta, cuando se quiere predecir la variable Y a partir de la variable X. Los valores menores de 0.5, si son estadísticamente significativos, pudiesen ser útiles para identificar la falta de relación entre las variables, pero probablemente no sean lo suficientemente altos como para ser útiles en la predicción de los puntajes individuales.

Aunque el coeficiente de correlación indica el grado de relación entre dos variables, ello no significa que determinen la causa de la relación o que los cambios en una variable causarán cambios en la otra. Otras variables pueden estar actuando en una o en ambas variables afectándolas en la misma dirección. Si se considera la relación entre el número de tiros libres realizados en el entrenamiento y el porcentaje de tiros libres en una temporada, la efectividad del jugador puede verse afectada por situaciones como la motivación, la presión de la posible clasificación a una final o de una eliminación, el tiempo

dedicado al entrenamiento o las propias condiciones físicas. Los cambios en una variable no incidirán automáticamente en otra variable; si se toman dos variables como la longitud de las piernas y la longitud del pantalón indudablemente que ambas se relacionan, pero el alargamiento del pantalón no hará que las piernas se alarguen más.

Gráfico de dispersión. La técnica para graficar los puntajes es más rápida que la técnica matemática pero no es tan precisa. El coeficiente de correlación se puede representar a modo de descripción visual en un gráfico de dispersión en el cual los puntajes obtenidos por cada sujeto se ubican en el eje correspondiente a cada variable, proyectándose hacia el punto del gráfico en que ambos coinciden. Se desarrolla un sistema de coordenadas de acuerdo con los valores o puntajes de una prueba que se ubican en el eje horizontal y los puntajes de la otra prueba que se ubican en el eje vertical; se dibuja un punto para señalar la ubicación de cada sujeto donde coinciden los puntajes de ambas pruebas. Al construir el gráfico se puede encontrar una relación entre las dos variables estudiando el patrón general de los puntos graficados. Si existe un patrón es necesario observar su dirección. Si tenemos dos pruebas, la prueba de flexiones de codo en apoyo facial (X) y la prueba de flexiones de codo en suspensión (Y) con los puntajes que se señalan a continuación, se puede observar la forma que adquiere el gráfico de dispersión.

N	X (Flexiones de codo en apoyo facial)	Y (Flexiones de codo en suspensión)
A	15	10
B	5	2
C	11	5
D	10	6
E	14	7
F	3	1
G	16	9
H	8	5

Tabla 6.2 Puntajes originales

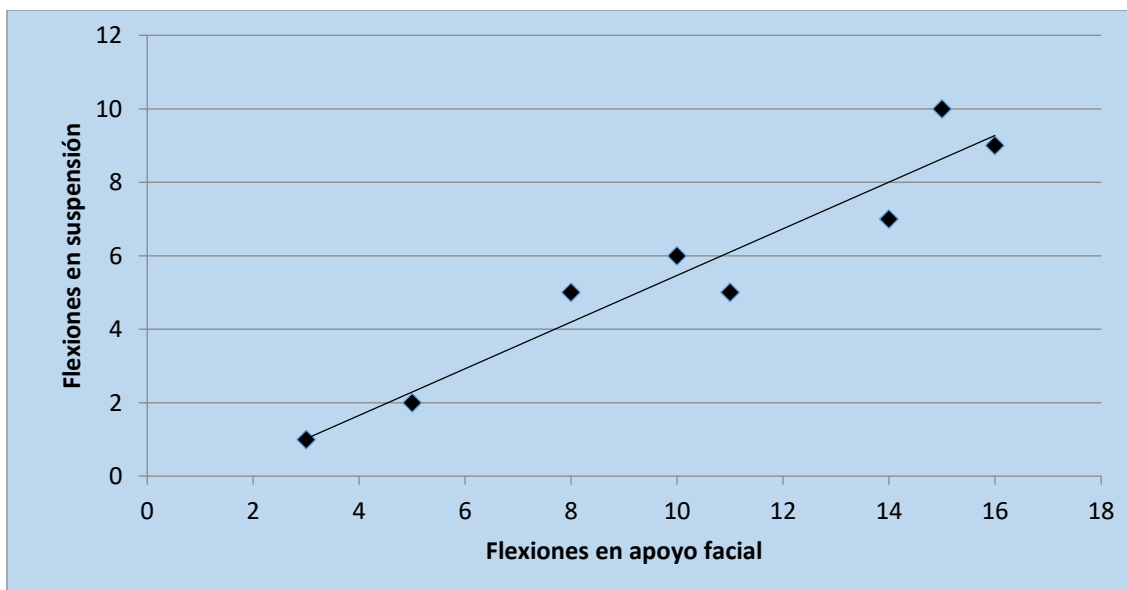


Figura 6.1 Relación positiva de los puntajes

En este caso se observa que el sujeto A obtuvo 15 flexiones de codo en apoyo facial y 10 flexiones de codo en suspensión, mientras el sujeto B obtuvo 5 y 2 respectivamente, y así con cada sujeto. Al ubicar a todos los sujetos se va a presentar una línea recta denominada línea de mejor ajuste o línea de regresión que expresa la tendencia o patrón de los puntajes graficados, indicando en este caso que los sujetos con mayor cantidad de flexiones en apoyo facial también tienen una mayor cantidad de flexiones en suspensión y viceversa, por lo que se ha producido una relación positiva entre ambas pruebas. Cuando los puntajes altos de una prueba se relacionan con los puntajes bajos de otra prueba, la relación es negativa (Figura 6.2). Mientras más cercanos estén los puntos trazados en el gráfico a la línea de tendencia, más fuerte es la relación, y la máxima relación se establece cuando todos los puntajes trazados en el gráfico se ubican en la línea señalada. Si los puntos trazados adquieren una forma de círculo haciendo imposible el trazado de la línea de mejor ajuste (Figura 6.3), se dice que no hay una relación lineal entre las pruebas que se están graficando.

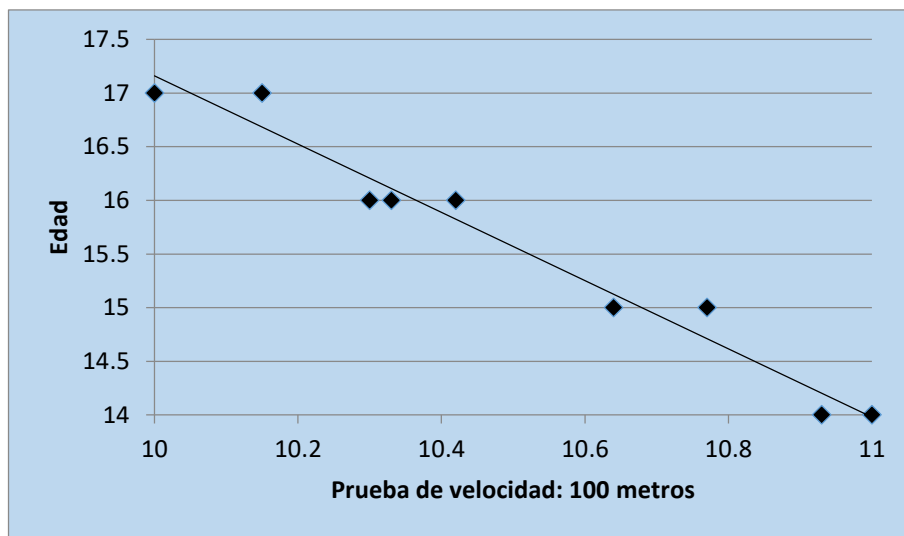


Figura 6.2 Relación negativa de los puntajes

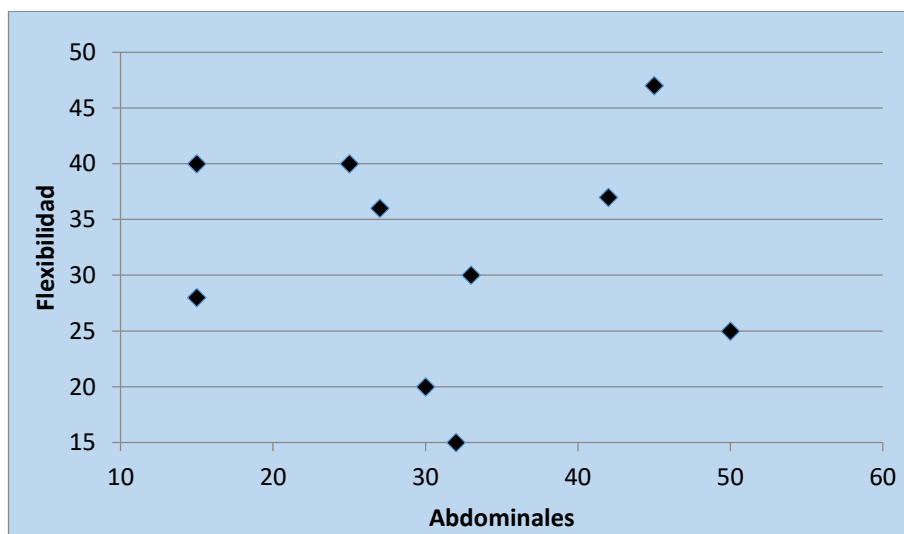


Figura 6.3 Ausencia de relación entre los puntajes

En general, al examinar un diagrama de dispersión se debe estudiar el patrón general de los puntos graficados. Si los datos van hacia arriba, esto sugiere que cuando una variable aumenta la otra también lo hace. Si los datos van hacia abajo, esto sugiere que cuando una variable aumenta la otra disminuye.

Cálculo del coeficiente de correlación. Para calcular el coeficiente de correlación se utiliza la fórmula desarrollada por Karl Pearson que busca determinar el grado de relación entre dos grupos de puntajes. Aunque el coeficiente de correlación puede calcularse a través del procesamiento de la media y la desviación estándar, es más utilizada la fórmula de Pearson porque no es necesario redondear los valores hallados lo que permite una precisión ligeramente superior; se requiere sólo la sumatoria de las columnas y las operaciones correspondientes como se observa en la fórmula:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

Donde r es el coeficiente de correlación producto – momento de Pearson, n es el número de sujetos, $\sum XY$ es la suma del producto de la multiplicación de cada X por Y , $\sum X$ es la suma de los puntajes de una variable, $\sum Y$ es la suma de los puntajes de la otra variable, $\sum X^2$ es la suma del cuadrado de los puntajes X , y $\sum Y^2$ es la suma del cuadrado de los puntajes Y . Si tenemos 8 sujetos que realizan dos pruebas, la prueba de flexiones de codo en apoyo facial y la prueba de flexiones de codo en suspensión, se hallan los valores correspondientes a cada suma y luego se calcula el coeficiente de correlación.

n	X (Flexiones de codo en apoyo fa- cial)	Y (Flexiones de codo en suspensión)	XY	X ²	Y ²
1	10	15	150	100	225
2	2	5	10	4	25
3	5	11	55	25	121
4	6	10	60	36	100
5	7	14	98	49	196
6	1	3	3	1	9
7	9	16	144	81	256
8	5	8	40	5	64
n=8	$\sum X = 45$	$\sum Y = 82$	$\sum XY = 560$	$\sum X^2 = 321$	$\sum Y^2 = 996$

Tabla 6.3 Puntajes para el cálculo del coeficiente de correlación

Sustituyendo los puntajes tenemos:

$$r = \frac{8(560) - (45)(82)}{\sqrt{[8(321) - 45^2][8(996) - 82^2]}}$$

$$r = \frac{4480 - 3690}{\sqrt{(2568 - 2025)(7698 - 6724)}} = \frac{790}{\sqrt{(543)(1244)}} = \frac{790}{\sqrt{675492}}$$

$$r = \frac{790}{821.88} = 0.96$$

Dado que la relación entre dos conjuntos de puntajes rara vez es perfecta, en la mayoría de las ocasiones son fracciones como 0.96, 0.87, etc. Mientras más cerca se halle de +1 o de -1, la relación entre las variables es más fuerte. Cuando la relación no es perfecta, los puntajes en una variable tienden a cambiar con los puntajes de la otra variable. Pero una alta correlación entre dos variables por lo general no indica una relación de causa – efecto. Una relación perfecta entre estatura y peso no indica que al incrementarse el peso también se incrementará la altura.

El verdadero indicador del grado de relación entre dos variables es el coeficiente de determinación: la cantidad de variabilidad en una medida que es explicada por la otra medida. El coeficiente de determinación es el cuadrado del coeficiente de correlación (r^2). En el ejercicio anterior el cuadrado del coeficiente de correlación es 0.92 (0.96^2) lo cual significa que el 92% de la variabilidad de los puntajes de flexiones de codo en apoyo facial se debe a los individuos que obtuvieron diferentes puntajes en las flexiones en suspensión. El 8% de la variabilidad de los puntajes de flexiones de codo en apoyo facial se explica por otros factores y no por las flexiones en suspensión.

Regresión

Los términos correlación, regresión y predicción están muy vinculados. La correlación se refiere a la relación entre dos variables; cuando dos variables se correlacionan es posible hacer una predicción. La regresión es un modelo estadístico utilizado para explicar la variabilidad (varianza) basada en una o más variables diferentes como, por ejemplo, al determinar la cantidad de variabilidad en la condición física con base en el porcentaje de grasa. Los docentes, entrenadores e investigadores están muy interesados en predecir unos puntajes que podían ser difíciles de obtener por otros medios. La predicción es la estimación del puntaje de una persona en una prueba con base en el puntaje de esa persona en otra u otras pruebas, como podría ser la estimación del porcentaje de grasa al tomar en cuenta las mediciones de los pliegues de grasa.

Predicción simple o regresión bivariada. La predicción simple es la predicción de un puntaje desconocido (Y) para un individuo, que se estima utilizando el puntaje que obtuvo en una prueba conocida (X). Para desarrollar una predicción simple o regresión, una gran cantidad de sujetos deben ser medidos (al menos 50 y mucho mejor si es mayor a 100), siendo un puntaje la variable independiente (X) o predictor, y el otro puntaje la variable dependiente (Y), de respuesta o criterio. Cuando la correlación entre dos variables es suficientemente alta, se puede predecir cuál será el puntaje de un sujeto en la variable Y si se conoce su puntaje en la variable X. Esto es particularmente útil si la variable X es relativamente fácil de medir, pero el puntaje en la variable Y no es fácil de obtener. Por ejemplo,

es fácil medir el tiempo que le toma a una persona correr 1.5 millas, pero es difícil medir el VO_2 máx en una cinta o una bicicleta. Por supuesto, esta predicción no es perfecta ya que contiene cierta cantidad de error que se debe aceptar para evitar la dificultad, la carencia o el costo de medir directamente el consumo de oxígeno. Supongamos que una prueba de abdominales en un minuto a 15 estudiantes en dos ocasiones en días continuos. Con los valores obtenidos por cada sujeto en las pruebas se elabora un gráfico XY, el gráfico de dispersión, donde los puntajes del primer día se ubican en el eje X y los del segundo día en el eje Y. En el gráfico de dispersión los datos se agrupan alrededor de la recta de mejor ajuste o recta de regresión. Una vez que se conoce esta recta, se observa que los puntajes X se proyectan verticalmente hacia la recta y los puntajes Y se proyectan horizontalmente hacia la misma, de forma tal que los puntajes del mismo sujeto coinciden en un punto del gráfico y de forma cercana a la recta.

Utilizando la recta de mejor ajuste se puede predecir el número de abdominales que se estima que se ejecutarían en segundo día si alguien no asiste o por alguna razón no puede presentar la prueba, con base en el puntaje del primer día. Por ejemplo, si uno de los sujetos logra el primer día 58 abdominales y no asiste el segundo día, se proyecta el puntaje desde el eje X en forma vertical hasta la recta de mejor ajuste, y desde allí se proyecta ese punto hacia el eje Y para determinar que el puntaje estaría comprendido entre 50 y 60 abdominales el segundo día, por lo que se puede estimar la cantidad de 55 abdominales como el puntaje de ese sujeto que estuvo ausente. Esta respuesta no es exacta ya que se debe localizar visualmente o por medio del gráfico con algún programa informático el punto que coincida con el eje Y. Por ello, para determinar con mayor precisión la estimación de Y, se procede a utilizar la ecuación de regresión correspondiente.

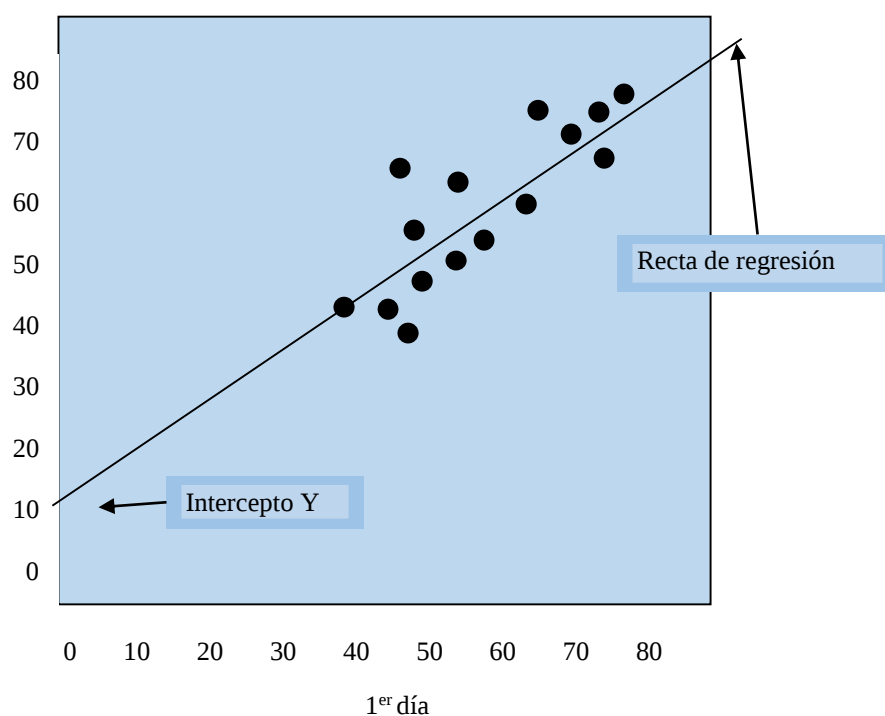


Figura 6.4 Predicción de Y a partir de X en un gráfico de dispersión (predicción aproximada)

Siendo la correlación entre las dos variables lo suficientemente alta, se grafica la recta de regresión de una forma más precisa al utilizar una ecuación que relacione las dos variables matemáticamente, donde la variable dependiente Y está determinada por la variable independiente X ; la ecuación de regresión se expresa así:

$$Y' = a + bX$$

donde Y' es el puntaje Y que se predice al considerar los puntajes X , a es el intercepto en el eje Y de la recta de mejor ajuste, b es la pendiente de la recta, X es el puntaje que se conoce. Para calcular b se utiliza la siguiente fórmula:

$$b = \left[r \left(\frac{s_y}{s_x} \right) \right]$$

donde r es el coeficiente de correlación entre los puntajes X y Y , s_y es la desviación estándar de los puntajes Y , y s_x es la desviación estándar de los puntajes X .

Para calcular a se utiliza la siguiente fórmula:

$$a = M_y - bM_x$$

donde M_y es la media de los puntajes Y , y M_x es la media de los puntajes X ; b es la pendiente de la recta. Si tenemos los siguientes valores del grupo de sujetos que presentaron la prueba de abdominales, se puede predecir la cantidad de abdominales que uno de ellos podría haber alcanzado el segundo día, al que no asistió:

1 ^{er} día (X)	r	2 ^{do} día (Y)
$M = 52,33$ $s_x = 12,31$	0.845	$M = 50,87$ $s_y = 10,56$
Tabla 6.4 Ejemplo de valores para el cálculo de la ecuación de regresión		

Si el sujeto obtuvo el primer día la cantidad de 58 abdominales, se sustituye primero la fórmula de b obteniendo el siguiente resultado:

$$b = \left[0,845 \left(\frac{10,56}{12,31} \right) \right] = (0,845 * 0,858) = 0,73$$

Luego se sustituye a :

$$a = [50,87 - (0,73)52,33] = 50,87 - 38,2 = 12,67$$

Una vez que se obtienen estos valores, se procede a calcular Y' (el puntaje a predecir con base en el puntaje 58 del primer día) aplicando la ecuación de regresión:

$$Y' = 12,67 + 0,73(58) = 12,67 + 42,34 = 55,01$$

Como conclusión tenemos que el sujeto que hizo 58 abdominales el primer día, podría haber obtenido 55 abdominales el segundo día. Ahora que se conoce la ecuación $Y' = 12,67 + 0,73(X)$, se puede predecir para cualquier valor X , en este caso abdominales, el puntaje Y' correspondiente.

Las ecuaciones de regresión se utilizan para predecir el valor de una variable, dado algún valor particular de la otra variable. Si la recta de mejor ajuste o recta de regresión se ajusta adecuadamente a los puntajes, entonces se puede utilizar la ecuación para hacer predicciones sobre todo si el coeficiente de correlación es significativamente alto.

Error estándar de la predicción o estimación. Cuando la correlación es menor a ± 1.0 , habrá una diferencia entre la medición realizada y el valor estimado de Y . Esa diferencia o distancia representa un puntaje residual, que es la diferencia entre el valor Y estimado y el valor Y real; el error residual representa el error en la predicción de cada valor Y . Por ejemplo, si el sujeto que hizo 58 abdominales el primer día hubiese hecho 51 abdominales el segundo día, el error residual es -4 , que es la diferencia entre el puntaje real ($Y = 51$) y el puntaje estimado ($Y' = 55$). Todo puntaje Y' que haya sido estimado tiene algún grado de error.

Para determinar el error de la predicción se utiliza la ecuación de regresión para predecir el valor de Y' de cada sujeto y luego comparar el valor Y estimado (Y') con el valor Y real, para determinar la cantidad de error de la predicción para cada sujeto ($Y - Y'$). Los valores que resultan de efectuar la operación $Y - Y'$ representan todos los errores o residuales alrededor de la recta de mejor ajuste o recta de regresión. Si tenemos 15 sujetos que hayan hecho la prueba de abdominales se presenta el siguiente análisis aplicando la ecuación de regresión $Y' = 12,67 + 0,73(X)$.

1er día (X)	2do día (Y)	Y'	$Y - Y'$	$(Y - Y')^2$
72	68	65,23	2,77	7,67
70	64	63,77	0,23	0,05
69	58	63,04	-5,04	25,4
67	61	61,58	-0,58	0,34
61	62	57,2	4,8	23,04
58	51	55,01	-4,01	16,08
51	55	49,9	5,1	26,01
49	44	48,44	-4,44	19,71
47	43	46,98	-3,98	15,84
43	60	44,06	15,94	254,08
42	46	43,33	2,67	7,13
41	38	42,6	-4,6	21,16
40	36	41,87	-5,87	34,46
40	37	41,87	-4,87	23,72
35	40	38,22	1,78	3,17
$\Sigma(Y - Y')^2 =$				477,86

Tabla 6.5 Cálculo del error estándar de la predicción

¿Qué significa la diferencia $Y - Y'$? Como se estableció anteriormente, la ecuación de regresión representa la recta que se ajusta mejor a los puntajes, por lo que para una muestra de puntajes apareados XY , un residual es la diferencia entre el puntaje original (Y) y el valor estimado (Y'). Por ejemplo, obsérvense en la tabla 6.5 los puntajes apareados ($X=51$, $Y=55$; $Y'=49,9$) y ($X=41$, $Y=38$; $Y'=42,6$) con residuales de 5,1 y -4,6, respectivamente.

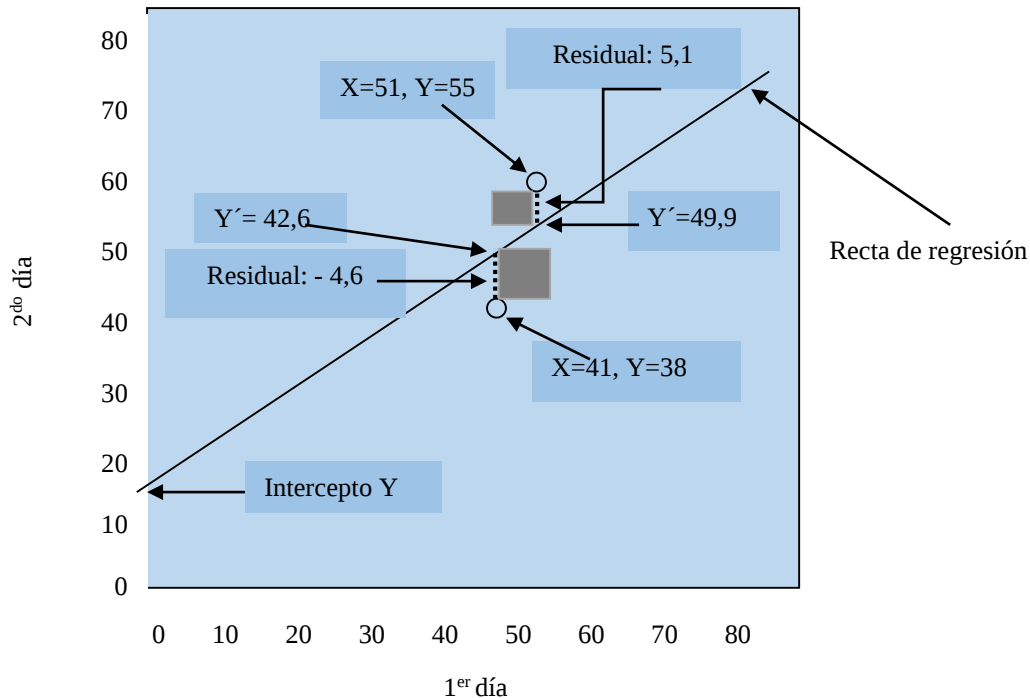


Figura 6.5 Residuales y cuadrados de los residuales

Los puntajes estimados $Y' = 49$ y $Y' = 42,6$ se ubican en la recta de regresión, apuntando hacia el puntaje X correspondiente en forma vertical. Los residuales se representan con la línea punteada. La zona sombreada es el cuadrado de cada residual; si todos los puntajes Y' con sus residuales se señalaran en el gráfico, la sumatoria de los cuadrados de los residuales sería 477,86 en el ejemplo dado (Tabla 6.5).

La desviación estándar de los residuales (el error estándar de la predicción) se calcula con la siguiente fórmula:

$$s_{res} = \sqrt{\frac{\sum(Y - Y')^2}{N}} = \sqrt{\frac{477,86}{15}} = 5,64$$

Otra forma de calcular el error estándar de la estimación, menos tediosa y más fácil, es la siguiente:

$$SE_E = s_y \sqrt{1 - r^2}$$

donde SE_E es el error estándar de la estimación o predicción, s_y es la desviación estándar de Y , y r^2 es el cuadrado del coeficiente de correlación (el coeficiente de determinación). Tomando en cuenta los valores ya reseñados tenemos:

$$SE_E = 10,56\sqrt{1 - (0,845)^2} = 5,64$$

El error estándar de la predicción se interpreta como la desviación estándar de todos los errores, o residuales, que se originan cuando se predice Y a partir de X . Dado que es la desviación estándar de unos residuales que se distribuyen normalmente, el SE_E puede interpretarse como un puntaje z de ± 1.0 ; se determina así que el 68% de todos los errores de la predicción se encuentran entre $\pm 1*SE_E$, el 90% se encuentran entre $\pm 1.65*SE_E$, el 95% se hallan $\pm 1.96*SE_E$, y el 99% se hallan entre $\pm 2.58*SE_E$.

§

Una vez que se conoce el error estimado para un valor estimado de Y , se puede ejemplificar siguiendo con los datos anteriores. Un sujeto que ejecutó 60 abdominales el primer día y no asistió el segundo presenta la siguiente predicción:

$$Y' = 12,67 + 0,73(60) = 56,47$$

El SE_E de esta predicción es $\pm 5,64$ a un nivel de confianza del 68%, lo que significa que la predicción de $56,47 \pm 1.0*5,64$ abdominales (entre 50 y 62 abdominales) tiene un 32% de probabilidades de ser incorrecta. Cuando el puntaje z es 1.96, el error de la predicción cae al 5% ($1.96*5,64 = 11,1$); la predicción se expresa así:

$$Y' = 56,47 \pm 1.96 (5,64) = 56,47 \pm 11,1, p = 0,05$$

En otras palabras, las probabilidades son 95 a 5 de que el valor estimado de Y se encuentre entre 45,4 y 67,6, o en números enteros, entre 45 y 68. La forma definitiva de la ecuación de regresión se representa así:

$$Y' = a + b(X) \pm z(SE_E)$$

donde z es un valor que genera el nivel de confianza deseado. Cuando la desviación estándar de Y es baja y la correlación XY es alta, el error de predicción es bajo. Si el coeficiente de correlación cae por debajo de 0.7 los errores se incrementan y la predicción no es útil.

EJERCICIOS

1. En el departamento de educación física se propuso realizar una prueba de fuerza de los miembros superiores que fuese fácil de aplicar. La prueba de flexiones de codo en las barras paralelas y la prueba de flexiones de codo en la barra horizontal fueron consideradas las más apropiadas. Uno de los docentes pensó que ambas pruebas no eran necesarias por cuanto la correlación entre ellas probablemente fuesen altas. Para evaluar este supuesto, a 141 estudiantes se les aplicaron ambas pruebas. Se estableció que la prueba en las barras paralelas fuera X y en la barra horizontal fuese Y , llegándose a los siguientes valores:

ΣX	3416
ΣY	1899
ΣX^2	93810
ΣY^2	28697
ΣXY	50509

Calcule la correlación entre ambas variables según la fórmula señalada.

2. Se quiso estudiar la correlación entre el tiempo realizado al correr una milla (1609 metros) y los valores de consumo de oxígeno ($VO_2\text{máx}$) hallados en el laboratorio de fisiología del ejercicio, aplicadas a un grupo de 18 estudiantes, con un tiempo de separación entre ambas pruebas de dos días. Los puntajes fueron los siguientes:

N	Carrera de 1 milla (seg.)	$VO_2\text{máx}$
1	250	60,3
2	315	57,2
3	420	55,4
4	410	51,4
5	436	52,5
6	511	45,6
7	460	38,4
8	510	41,5
9	530	39,6
10	586	33,2
11	591	37,7
12	600	40,1
13	626	32,0
14	643	35,4
15	650	33,7
16	675	35,9
17	710	27,4
18	720	25,3

- a) Establezca el coeficiente de correlación utilizando una calculadora
- b) Interprete el resultado

- c) Introduzca los datos en un programa informático (hoja de cálculo) y compare ambos resultados
3. Se registraron las edades y los tiempos de 150 corredores de un maratón elegidos al azar y se calculó el coeficiente de correlación lineal $r = 0,144$.
- Señale el tipo de correlación calculado e interprete ese resultado.
4. Leonardo da Vinci (1452-1519) dibujó la figura de un hombre, indicando que la distancia entre los brazos extendidos de una persona (midiendo por la espalda con los brazos extendidos para formar una “T”) es casi igual a la estatura de una persona. Para probar lo dicho por él, medimos ocho personas con los siguientes resultados:

	Sujetos							
	1	2	3	4	5	6	7	8
Dis-	1,7	1,58	1,6	1,76	1,73	1,75	1,57	1,53
tan	3		5	5			5	
cia								
bra								
zos								
Es-	1,7	1,57	1,6	1,77	1,70	1,70	1,60	1,57
ta-	5	5	5	8	2	2		5
tur								
a								

- Trace una gráfica de dispersión para distancia entre los brazos extendidos y estatura. Use la misma escala en los ejes horizontal y vertical. Describa la relación entre las dos variables.
 - Si da Vinci estaba en lo correcto y la distancia entre los brazos extendidos de una persona es casi igual a la estatura de esa persona, ¿cuál debe ser la pendiente de la recta de regresión?
 - Calcule la recta de regresión para predecir la estatura con base en la distancia entre los brazos extendidos de una persona. ¿El valor de la pendiente b confirma las conclusiones de usted del inciso b)?
 - Si una persona tiene una distancia de 1,575 cm entre los brazos extendidos, ¿cuál sería el pronóstico de usted respecto a la estatura de la persona?
5. ¿El promedio de bateo de un equipo depende en alguna forma del número de cuadrangulares conectados por el equipo? Los datos de la tabla siguiente muestran el número de cuadrangulares del equipo y el promedio general de bateo del equipo para ocho equipos de ligas mayores seleccionados para la temporada de 2006.

Equipo	Total cuadrangulares	Promedio de bateo del equipo
Bravos de Atlanta	222	.270
Orioles de Baltimore	164	.227
Medias Rojas de Boston	192	.269
Medias Blancas de Chicago	236	.280
Astros de Houston	174	.255
Filis de Filadelfia	216	.267
Yanquis de Nueva York	163	.259
Marineros de Seattle	172	.272

- a. Grafique los puntos usando una gráfica de dispersión. ¿Le parece que hay alguna relación entre el total de cuadrangulares y el promedio de bateo del equipo?

Capítulo 7. Fiabilidad y objetividad

Hay unas características esenciales en toda medición: sin ellas, no se puede tener fe en las mediciones y prácticamente no se les puede dar algún tipo de uso. La característica más importante de una medición es la validez (ver Capítulo 8). Luego le siguen la fiabilidad y la objetividad.

Una medida es fiable cuando se mantiene sin cambios en forma consistente durante un corto período de tiempo. Esto significa que si un individuo cuyas destrezas no han cambiado es medido en dos oportunidades con una diferencia no mayor a uno o dos días entre cada medición, los dos puntajes deben ser idénticos. Para que una medición sea válida, debe ser fiable. La objetividad también es llamada fiabilidad del evaluador, ya que se ha definido como el acuerdo entre jueces competentes sobre el valor de una medición. Esto es, si dos jueces anotadores que miden al mismo individuo en la misma prueba no están de acuerdo en el puntaje, la medición pierde objetividad y, por tanto, ningún puntaje tiene fiabilidad ni validez. Una falta de objetividad reduce, entonces, la fiabilidad y la validez.

La objetividad es necesaria para que haya fiabilidad, y la fiabilidad es necesaria para que haya validez. De modo que, para determinar si estas tres importantes características de una medición están presentes, la objetividad y la confiabilidad son calculadas y la validez es estimada. Primero se expone la fiabilidad y objetividad de las pruebas referidas a normas, y posteriormente la fiabilidad de las pruebas referidas a criterio.

Fiabilidad a partir de la selección de un puntaje de criterio

Un puntaje de criterio es la medida utilizada para indicar la capacidad de una persona. Al menos que una medida tenga una fiabilidad perfecta, el puntaje de criterio es el mejor indicador cuando se desarrolla a partir de más de una prueba. Las pruebas múltiples son comunes en educación física cuando se miden pliegues cutáneos, flexibilidad, saltos, lanzamientos y pruebas de carrera corta, es decir, todas las pruebas cuyo rendimiento no se ve afectado negativamente por la fatiga. Además, hay múltiples pruebas como pliegues cutáneos, flexibilidad y pruebas de fuerza, que son comunes en ciencias del ejercicio.

El puntaje medio versus la mejor puntuación. En una prueba con intentos múltiples, el puntaje de criterio puede ser, o el mejor puntaje de la persona, o su puntaje promedio. El mejor puntaje es el puntaje óptimo que recibe una persona en cualquier intento, el puntaje promedio es el promedio de todos los intentos en la prueba en cuestión. La elección entre el mejor puntaje y el promedio debe basarse en la cantidad de tiempo disponible para obtener el puntaje de criterio y el uso que se le dé. Es un proceso mucho más lento calcular una puntuación media que elaborar una lista con la mejor puntuación de cada persona. Esta es probablemente la razón por la cual las instrucciones estandarizadas para la mayoría de las pruebas de intentos múltiples sugieren usar la mejor puntuación. Ciertamente,

cuando el puntaje de criterio se va a usar como un indicador del máximo rendimiento posible, se debe usar el mejor puntaje. También con certeza, para los participantes en una competencia atlética donde se pueden realizar múltiples saltos o lanzamientos, el mejor puntaje es su puntaje personal. Sin embargo, el mejor puntaje en varios intentos podría no ser lo acostumbrado en esa persona, ya que es poco probable que se logre nuevamente durante algún tiempo y podría verse incrementado por un error de medición.

La fiabilidad de un puntaje de criterio aumenta a medida que aumenta el número de intentos en los que se basa. Por lo tanto, el puntaje de criterio más confiable y el mejor indicador de la capacidad real (rendimiento típico) puede ser el puntaje promedio. En teoría, la suma de los errores de medición es 0 cuando se suman todos los puntajes en la prueba de un individuo. Debido a que, en la mayoría de las situaciones, uno quiere una indicación del estado o rendimiento típico, la puntuación media puede ser la puntuación de criterio preferida para una prueba de ensayos múltiples. Por lo tanto, si se tomaran múltiples medidas de pliegues cutáneos en cada sitio, la media de las mediciones en un sitio sería el puntaje de criterio para ese sitio. Cuando los puntajes cambian de un día a otro, como puede ser cierto para las personas con discapacidades, la media de los puntajes recopilados en varios días puede ser el único indicador de la verdadera capacidad.

En general, el puntaje de criterio de un individuo con datos de varios intentos podría determinarse de cualquiera de las siguientes maneras, dependiendo de la situación y de lo que el administrador de la prueba considere más apropiado: (1) promedio de todos los puntajes de los intentos, (2) mejor puntaje de todos los puntajes de los intentos, (3) promedio de los puntajes de los intentos seleccionados en función de los intentos en los que el grupo obtuvo el mejor puntaje, o (4) promedio de los puntajes de los intentos seleccionados en función de los intentos en el que el individuo obtuvo la mejor puntuación. En términos del cuarto método, si se administran entre cuatro y seis intentos, parece haber una tendencia a que una persona tenga un puntaje en un intento que es notablemente peor que el resto de los puntajes del intento. Por lo tanto, el peor puntaje del intento podría descartarse y la media de las puntuaciones restantes podría ser la puntuación de criterio. Es interesante notar aquí que descartar la peor prueba para una persona hace que la distribución de la puntuación de prueba restante para la persona sea considerablemente más parecida a una distribución normal (los investigadores consideran que esa puntuación en la prueba se descartó como "atípica"). Por lo tanto, la forma en que se determina el puntaje de criterio de un individuo influye en el puntaje del individuo y es importante.

Tipos de fiabilidad

La fiabilidad de las mediciones del desempeño físico tradicionalmente ha sido estimada por una de las dos vías: el método de estabilidad test – retest, o por el método de consistencia interna. Debido a que cada uno produce un coeficiente de fiabilidad diferente, es importante usar el método más apropiado para un instrumento de medición dado. También es importante notar los métodos que otros han usado para calcular sus coeficientes de fiabilidad. Recuerde también que una prueba puede arrojar puntajes fiables para un grupo de individuos y no para otro. Por ejemplo, una prueba que arroja puntajes altamente fiables

para estudiantes universitarios puede arrojar solo puntajes moderadamente fiables para estudiantes de secundaria o participantes en un programa de acondicionamiento físico para adultos. Un principio implícito en el cálculo del coeficiente de fiabilidad consiste en determinar si este coeficiente indica estabilidad o consistencia interna.

Fiabilidad por estabilidad. Cuando los puntajes de los sujetos cambian muy poco de un día a otro, los puntajes son estables, y cuando los puntajes permanecen estables, se considera que son fiables. Se utiliza el método de test – retest para obtener el coeficiente de fiabilidad por estabilidad. Por medio de este método se mide cada sujeto con la misma prueba en varias oportunidades, por lo general en dos días diferentes. La correlación entre los dos conjuntos de puntajes constituye el coeficiente de fiabilidad por estabilidad. Mientras más cercano se encuentre este coeficiente a uno positivo (+ 1), más estables y fiables son los puntajes.

Un bajo coeficiente de fiabilidad por estabilidad puede deberse a tres factores: (a) Las personas medidas ejecutan la prueba de manera diferente en cada oportunidad; (b) Los instrumentos de medición no son aplicados en igualdad de condiciones; (c) El personal que aplica la prueba puede ser cambiado. La falta de sueño, las lesiones menores y la ansiedad tienden a disminuir el nivel de rendimiento. Además, si el instrumento no es consistente día a día, por ejemplo, si los dispositivos de medición se descalibran o funcionan mal, o si los procedimientos utilizados para recopilar las medidas cambian, la fiabilidad por estabilidad disminuye. Finalmente, la fiabilidad disminuye si el administrador cambia la manera de anotar o cómo percibe los cambios en el rendimiento.

Por lo general el método de test – retest se aplica por segunda vez entre uno y tres días después de la primera aplicación. Sin embargo, en una prueba de máximo esfuerzo se recomienda aplicar la segunda prueba siete días después ya que la fatiga y el dolor puede afectar los puntajes. Si el intervalo entre las dos mediciones es muy largo, por ejemplo, 15 días, los puntajes pueden cambiar debido a factores como la maduración o el incremento del tiempo de práctica, factores que se podrían pasar por alto cuando se consideran las fuentes del error de la medición.

Algunas personas se oponen al método test-retest debido al tiempo requerido para administrar un instrumento de medición al menos dos veces. Además, los puntajes del día 1 se usan como medidas de desempeño; los puntajes posteriores se usan únicamente para determinar la fiabilidad. Sin embargo, el método es probablemente el más apropiado de los procedimientos para determinar la fiabilidad de las medidas de rendimiento físico. Sin la coherencia test-retest, carecemos de un indicador verdadero, no solo de la capacidad de cada participante, sino de la fe que podemos tenerle a la medición.

Fiabilidad por consistencia interna. Muchos docentes y entrenadores personas utilizan un método de consistencia interna para estimar la fiabilidad de las mediciones, cuya ventaja es que todas las mediciones se realizan el mismo día. La consistencia interna se refiere a una tasa constante de puntuación por parte de los individuos que se evalúan a lo largo de una prueba o, cuando se administran varios intentos, de un intento a otro.

Quienes administran pruebas regularmente utilizan aquellas que requieren varios intentos cuando se trata de medir cualidades físicas, por ejemplo, las mediciones de los panículos de grasa, las pruebas de fuerza de un músculo o grupo muscular, o varios intentos en una prueba de rendimiento físico. Para obtener el coeficiente de fiabilidad por consistencia interna, el evaluador debe darle al participante al menos dos intentos el mismo día. Un cambio en el resultado entre los dos intentos de la persona que se está evaluando indica una falta de fiabilidad de la prueba. La correlación entre los puntajes obtenidos en los dos intentos es el coeficiente de fiabilidad por consistencia interna. Por supuesto, esta técnica no debe usarse en una prueba de máximo esfuerzo (por ejemplo, correr una milla) porque la fatiga seguramente afectará el resultado en el segundo de los intentos.

Estabilidad versus consistencia interna. El coeficiente de fiabilidad por consistencia interna no es comparable al coeficiente de fiabilidad por estabilidad. El primero no se ve afectado por los cambios diarios en el rendimiento, una fuente importante del error de medición en el segundo. Un coeficiente de fiabilidad por consistencia interna es casi siempre más alto que su correspondiente coeficiente de fiabilidad por estabilidad.

La educación, la psicología y otras disciplinas que dependen en gran medida de las pruebas con papel y lápiz rara vez (si es que alguna vez) usan el método test-retest, utilizando en su lugar el método de consistencia interna. Recuerde que el coeficiente de estabilidad asume que la capacidad real no cambia de un día para otro, una suposición a menudo injustificable con las pruebas de papel y lápiz, porque el aprendizaje cognitivo podría ocurrir entre la aplicación de ambas pruebas. El aprendizaje psicomotor es menos apto para variar en un lapso de uno o dos días, lo que hace que el coeficiente de estabilidad sea un mejor indicador de la fiabilidad de los datos de rendimiento físico.

Aspectos teóricos de la fiabilidad

Podemos entender mejor la fiabilidad de las pruebas referidas a normas si comprendemos la teoría matemática de su base. La fiabilidad puede explicarse en términos de "puntajes observados", "puntajes verdaderos" y "puntajes de error". La teoría de la fiabilidad supone que cualquier medición en una escala continua contiene un componente de error inherente, el error de medición. Cualquiera o más de los siguientes factores pueden ser una fuente de error de medición:

1. Falta de acuerdo entre los anotadores (es decir, objetividad);
2. Falta de consistencia en el rendimiento de los sujetos medidos;
3. Fallas de un instrumento o equipo de medición, para medir consistentemente;
4. Fracaso del examinador en seguir los procedimientos estandarizados de medición.

Supongamos que se va a medir la estatura de cinco personas, conociendo que todas miden 1,73 metros. Si se informa que hay un puntaje diferente a 1,73 metros, ha ocurrido un error en la medición. Por lo tanto, en este caso la varianza de las estaturas es un buen indicador de la magnitud del error de la medición. La varianza de la población es el cuadrado de la desviación estándar y se simboliza σ^2 , mientras que la varianza de la muestra

se simboliza s^2 . Si todas las estaturas fuesen 1,73, no hay error en la medición y la varianza sería 0. Pero si las cinco personas no tienen la misma estatura, la varianza de las estaturas puede deberse a una diferencia en las estaturas verdaderas o a un error de la medición. En cualquier caso, la varianza no puede ser utilizada como un indicador del error de medición.

En teoría, el puntaje observado X (lo que se anotó) es producto de la suma entre el puntaje verdadero y el error de la medición:

$$X = t + e$$

donde t es el puntaje verdadero y e es el error de la medición.

Si a un sujeto con una estatura de 1,78 se le anota una talla de 1,79, el error de medición es de 0,01:

$$1,79 = 1,78 + 0,01$$

Si el mismo sujeto es medido nuevamente y se le anota 1,77, el error de medición es de -0.01 :

$$1,77 + -0,01$$

La varianza del conjunto de puntajes observados es igual a la varianza de los puntajes verdaderos más la varianza del error de los puntajes:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

donde σ_x^2 es la varianza de los puntajes observados, σ_t^2 es la varianza de los puntajes verdaderos, y σ_e^2 es la varianza del error de los puntajes.

La fiabilidad, entonces, es la relación entre la varianza del puntaje verdadero y la varianza del puntaje observado:

$$Fiabilidad = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

De acuerdo con esta fórmula, cuando el error de medición no existe, esto es, cuando σ_e^2 es igual a 0, la fiabilidad es 1. Si el error de medición se incrementa, σ_e^2 también se incrementa y la fiabilidad disminuye, concluyéndose que la fiabilidad es un indicador de la magnitud del error de medición en un conjunto de puntajes.

La fiabilidad depende de dos factores básicos:

1. La reducción de la variación atribuible al error de medición, y
2. La detección de las diferencias individuales (es decir, la variación de los puntajes verdaderos) dentro del grupo medido.

La fiabilidad de un conjunto de puntajes debe ser vista en términos de la medición del error (varianza del error) y cuán bien los puntajes discriminan entre los diferentes niveles de destreza dentro del grupo medido (varianza del puntaje verdadero).

Estimación de la fiabilidad: correlación intraclase. Como hemos señalado, una puntuación observada (X) se compone teóricamente de una puntuación verdadera (t) y una puntuación de error (e). Además, la varianza de los puntajes observados (σ_x^2) es igual a la varianza de los puntajes verdaderos (σ_t^2) más la varianza de los puntajes de error (σ_e^2). La confiabilidad es igual a la varianza de puntaje real dividida por la varianza de puntaje observado. Así como la varianza de puntaje observado se puede dividir en varias partes, la variabilidad total (s^2) para un conjunto de puntajes se puede dividir en varias partes. Para dividir o dividir la varianza, utilizamos la técnica de análisis de varianza (ANOVA). Entonces podemos usar estas partes en la varianza total para calcular un coeficiente de confiabilidad intraclase.

Antes de analizar dos modelos de análisis de varianza para el cálculo del coeficiente de fiabilidad intraclase, hay dos cosas que se deben reconocer. Primero, la fiabilidad del puntaje de una prueba debe estimarse antes de aplicar una prueba para recopilar una gran cantidad de datos en un programa de medición o una investigación. Por lo tanto, la prueba se administra a un pequeño grupo de personas de características similares a las del grupo al que finalmente se aplicará la prueba, y se procede a estimar la fiabilidad de los puntajes del grupo pequeño. En la investigación, esta evaluación de un grupo pequeño se llama estudio piloto. A menudo, el número de puntajes de prueba recopilados (puntajes de intentos múltiples o de días diferentes) se necesitan en cada uno de estos procedimientos del estudio piloto para poder calcular un coeficiente de fiabilidad intraclase. En el capítulo correspondiente se presentarán los procedimientos de ANOVA.

Error estándar de la medición. En muchas ocasiones resulta útil estimar el error de medición de cada prueba. A veces nos quedamos con el aura de la recopilación de datos científicos y no tomamos en cuenta que la posibilidad del error de la medición siempre existe. Es el caso de una prueba de máximo consumo de oxígeno (VO_{2max}) como una prueba válida para medir la condición cardiorrespiratoria. Las pruebas de campo por lo general se validan a través de la correlación con una VO_{2max} . Sin embargo, se debe ser cuidadoso y no considerar que esta prueba esté libre de error. Todo rendimiento en una prueba son puntajes observados y solo se puede obtener una estimación del puntaje verdadero de un sujeto. Es mucho mejor que los puntajes de una prueba caigan dentro de un rango que contiene el puntaje verdadero.

El error estándar de la medición (s_e) se define como la desviación estándar de la distribución de las puntuaciones de error. Como la desviación estándar de cualquier distribución de puntuaciones es la raíz cuadrada de la varianza, es posible despejar la ecuación y deducir, a partir de la relación entre la fiabilidad y la varianza del error, el error estándar de la medición.

Si cada sujeto tiene varios intentos en una prueba, se puede calcular la desviación estándar de los puntajes, y esta desviación estándar es el error de medición de ese individuo. Si cada persona tiene un solo intento, el promedio del error de medición de la prueba se estima para todo el grupo calculando el error estándar de la medición de acuerdo con la siguiente fórmula:

$$s_e = s_x \sqrt{1 - r}$$

donde s_e es el error estándar de la medición, s_x es la desviación estándar de los puntajes de la prueba, y r es el coeficiente de fiabilidad de las pruebas.

Al medir, por ejemplo, el porcentaje de grasa corporal de un grupo de mujeres adultas supongamos que la desviación estándar es 5,6% y la correlación test – retest es 0,83. El error estándar de la medición sería:

$$s_e = 5,6\sqrt{1 - 0,83} = 2,3\%$$

Supongamos también que una persona en particular tiene un 22,4% de grasa. Entonces podemos utilizar el error estándar de la medición para estimar el rango dentro del cual su porcentaje de grasa probablemente se encuentre.

En una distribución normal, el error estándar de la medición se interpreta de igual manera que la desviación estándar. Cerca de dos tercios (68,26%) de todos los puntajes de la prueba caerán dentro ± 1 error estándar de la medición de los puntajes verdaderos. En otras palabras, hay alrededor de un 68% de oportunidades de que el puntaje verdadero de una persona se encuentre dentro de un rango de ± 1 error estándar de la medición con respecto al puntaje que haya obtenido. Considerando la mujer que obtuvo un puntaje de 22,4% de grasa, hay 2 de 3 oportunidades (dos tercios) de que su porcentaje de grasa verdadero sea $22,4\% \pm 2,3\%$ (el error estándar de la medición), o de otra forma, entre el 20,1% y el 24,7%.

Si se quiere obtener un mayor nivel de confianza, se multiplica el error estándar de medición por 2 de manera que el 95,44% de las veces el puntaje verdadero se encuentre en un rango del puntaje observado ± 2 error estándar de la medición. En el ejemplo, se puede asegurar que el verdadero % de grasa es $22,4\% \pm (2,3 \times 2)$, es decir, entre el 17,8% y el 27%. De acuerdo con la fórmula se puede deducir que el error estándar de la medición está condicionado por la variabilidad de los puntajes y la fiabilidad de la prueba. Si se obtiene un alto grado de fiabilidad, el error estándar de la medición disminuye. El error estándar de la medición, entonces, especifica los límites dentro de los cuales se puede esperar que los puntajes varíen debido al error de la medición.

Objetividad

La objetividad, o fiabilidad del evaluador, es una característica importante en una medición. Para que una medición sea válida y fiable, debe poseer objetividad. Se puede definir

objetividad como del acuerdo más cercano posible entre los puntajes asignados a cada persona por dos o más jueces. Estos jueces actúan en deportes como gimnasia o en pruebas de condición física como la prueba de abdominales. En muchas ocasiones se prefieren las mediciones objetivas sobre las subjetivas, por cuanto se depende esencialmente de cuán válidas y confiables sean dichas mediciones, más que del criterio subjetivo del juez o del docente. Las mediciones objetivas no son automáticamente mejores que las subjetivas, sino que el rendimiento expresado en puntajes cuantitativos es más visible, palpable, y pueden ser procesados estadísticamente con mayor facilidad.

Factores que afectan la objetividad. La objetividad depende de dos factores relacionados: (a) La claridad del sistema de puntuación, y (b) El grado con el que el juez asigna los puntajes con precisión. Algunas pruebas tienen claramente definida la forma de anotación: la carrera de una milla se registra con un cronómetro, y la prueba de abdominales se registra según el número de repeticiones. En este tipo de pruebas el evaluador puede determinar fácilmente los puntajes de las personas. En contraste, la ejecución en danzas o en nado sincronizado no tiene un mecanismo tan preciso de medición, recayendo en el juicio del evaluador la asignación del puntaje.

El segundo factor es más obvio. Si un juez no sabe cómo asignar un puntaje, no puede hacerlo con precisión. Por ejemplo, un anotador que no esté familiarizado con los cronómetros probablemente no le asigne un puntaje preciso a un corredor de velocidad. Por supuesto, es una simple cuestión de entrenar a los jueces en el manejo de los procedimientos y equipamiento de medición.

Es importante que exista un alto grado de objetividad cuando dos o más personas administran una prueba. Esto sucede, por ejemplo, cuando un grupo de 50 personas es dividido en dos grupos y cada grupo es medido por jueces diferentes. Si no es evidente un alto grado de objetividad porque los dos grupos de evaluadores emplean procedimientos administrativos o estándares de anotación diferentes, el puntaje de los participantes va a depender de la perspectiva de ese evaluador. Asimismo, si un evaluador es más indulgente que otro, el participante que sea evaluado por ese evaluador tendrá una ventaja.

También se requiere un alto grado de objetividad si el evaluador actúa en diversas oportunidades. Por ejemplo, un juez puede medir un tercio del grupo en cada uno de los tres días, o al grupo completo al comienzo y al final de una unidad instruccional o de un período de entrenamiento. Desde luego, en el primer factor es esencial que se utilicen los mismos procedimientos administrativos y estándares de anotación. Esto también es cierto en el segundo factor, donde cualquier variación en el puntaje de un participante debe ser expresión de un cambio en el desempeño del sujeto y no del cambio en el procedimiento o en los estándares.

Se ha discutido la objetividad en dos categorías: intra-evaluador o inter-evaluador. La objetividad intra-evaluador es el grado de acuerdo en el puntaje asignado a cada persona por un juez cuando observa el desempeño de cada persona en dos ocasiones diferentes. Por ejemplo, si las personas fueron grabadas en video mientras realizaban la actividad física o la prueba, un juez podría ver el video grabado y calificar a cada persona en cada uno de los dos días diferentes. La objetividad inter-evaluador es el grado de acuerdo entre

los puntajes asignados a cada persona por dos o más jueces. La objetividad, tal como se define y discute en esta guía, es inter-evaluador. La objetividad intra-evaluador es vital para la objetividad inter-evaluador, pero una alta objetividad intra-evaluador no garantiza una alta objetividad inter-evaluador. En la mayoría de las situaciones de medición, se asume la alta objetividad intra-evaluador. Por lo general, la objetividad intra-evaluador será mayor que la objetividad inter-evaluador.

Estimación de la objetividad. Para estimar el grado de objetividad de un puntaje en una prueba de rendimiento físico en un estudio piloto, dos o más jueces califican cada persona a medida que ejecuta la prueba. Luego se calcula un coeficiente de correlación intraclase para cada persona sobre la base de los puntajes de los jueces. Probablemente en el programa de medición real o en el estudio de investigación se utilizaría un solo juez. Por lo tanto, se debe utilizar la fórmula para estimar la fiabilidad de un puntaje de criterio que es un puntaje único.

Para calcular el coeficiente de objetividad, los puntajes de los jueces serán los intentos, insertando sus puntajes individuales en los términos de la prueba en la fórmula de fiabilidad. Si se supone que todos los jueces están usando los mismos estándares, podríamos considerar una diferencia entre los jueces como un error de medición y estimaríamos la objetividad usando la fórmula ANOVA unidireccional o de una vía.

Si no se espera que todos los jueces usen los mismos estándares, estimaríamos la objetividad usando la fórmula ANOVA bidireccional o de dos vías. Muy rara vez se descartarían los datos de los jueces al considerar los múltiples intentos. Dado que la objetividad es un tipo de confiabilidad (confiabilidad del evaluador), la información sobre tamaño de muestra, confiabilidad aceptable, etc., generalmente se aplica al calcular e interpretar los coeficientes de objetividad.

Fiabilidad de los puntajes en las pruebas referidas a criterios

Los estándares referidos a criterios y la configuración de los criterios referidos se discutieron en un capítulo anterior. Con base en un estándar referido a criterios, una persona se clasifica como competente o no competente, ya sea aprobado o no. En un ejemplo, en un curso de certificación de la Cruz Roja, una persona cumple con los requisitos mínimos y está certificada o no cumple con los requisitos mínimos y no está certificada. En otro ejemplo, el estándar referido a criterio en un curso de acondicionamiento físico para adultos podría ser la capacidad de correr continuamente durante 30 minutos. Según esta norma, una persona se clasifica como competente o no competente. Una característica muy interesante de un estándar referido a criterios es que no hay cuotas predeterminadas sobre cuántas personas se clasifican como competentes.

La fiabilidad referida a criterios se define de manera diferente a la fiabilidad referida a normas. En el caso de la referida a criterios, la fiabilidad se define como la consistencia de la clasificación. Por lo tanto, si un puntaje en una prueba referida a criterios es

fiable, una persona se clasificará de la misma manera en cada una de las dos ocasiones. Esto sería en pruebas dentro del mismo día.

Para estimar la fiabilidad de un puntaje en una prueba referida a criterios, se forma una tabla de doble clasificación como se presenta en la tabla 7.1. En el cuadro A está el número de personas que pasaron los dos días y en el cuadro D está el número de personas que fallaron en ambos días. Observe que el cuadro B y el cuadro C es la cantidad de personas que no fueron clasificadas de la misma manera en ambas ocasiones. Obviamente, la mayor cantidad en los cuadros A y D, y la menor cantidad en los cuadros B y C son deseables, porque la fiabilidad es la consistencia de la clasificación en ambas ocasiones. Todos los paquetes de programas estadísticos de computadora tienen un programa de tabulación cruzada que proporcionará la tabla de clasificación doble (Tabla 7.1) necesaria para estimar la fiabilidad de una prueba referida a criterios.

		Día 2	
		Aprueba	No aprueba
Día 1	Aprueba	A	B
	No aprueba	C	D

Tabla 7.1 Tabla para la estimación de la fiabilidad de una prueba referida a criterios

La forma más popular de estimar la confiabilidad de esta tabla de clasificación doble es calcular la proporción del coeficiente de acuerdo (P) donde

$$P = \frac{A + D}{A + B + C + D}$$

Por ejemplo, al usar los datos de la tabla 7.2 se determina la proporción del acuerdo (P) en una prueba referida a criterios.

		Día 2	
		Aprueba	No aprueba
Día 1	Aprueba	84	21
	No aprueba	5	40

Tabla 7.2 Datos para la determinación de la fiabilidad de una prueba referida a criterios

La suma de las cajas A y D es 124, y la suma de las cuatro cajas es 150, por lo que $P = 0,83$.

$$P = \frac{84 + 40}{84 + 21 + 5 + 40} = \frac{124}{150} = 0,83$$

La proporción de acuerdo (P) no se consigue por el hecho de que algunas clasificaciones, en ambas ocasiones, pueden haber sucedido totalmente por azar. El coeficiente kappa (k) corrige los acuerdos de azar. La fórmula para kappa es:

$$k = \frac{Pa - Pc}{1 - Pc}$$

Donde

Pa = proporción del acuerdo

Pc = proporción del acuerdo esperado por azar

$$Pc = \frac{(A + B)(A + C) + (C + D)(B + D)}{(A + B + C + D)^2}$$

en una tabla de doble clasificación. Por ejemplo, utilizando con los datos de la tabla 7.2 se determina el coeficiente kappa (k) en una prueba referida a criterios; para solucionarlo se siguen tres pasos:

Paso 1. Calcular Pa

$$P = \frac{A + D}{A + B + C + D} = \frac{84 + 40}{84 + 21 + 5 + 40} = \frac{124}{150} = 0,83$$

Paso 2. Calcular Pc

$$\begin{aligned} Pc &= \frac{(A + B)(A + C) + (C + D)(B + D)}{(A + B + C + D)^2} = \frac{(105)(89) + (45)(61)}{(84 + 21 + 5 + 40)^2} \\ &= \frac{9345 + 2745}{150^2} \\ &= \frac{12090}{22500} = 0,537 = 0,54 \end{aligned}$$

Paso 3. Calcular kappa

$$k = \frac{Pa - Pc}{1 - Pc} = \frac{0,83 - 0,54}{1 - 0,54} = \frac{0,29}{0,46} = 0,63$$

Obsérvese que para los datos en la tabla 7.2 hay una diferencia definitiva entre la proporción de acuerdo (P) y el coeficiente kappa (k). Como P puede verse afectado por las clasificaciones por azar, los valores de P inferiores a 0.50 se interpretan como inaceptables. Por lo tanto, los valores de P deben estar más cerca de 1.0 que de 0.50 para ser bastante aceptables. Los valores de k también deberían estar más cerca de 1.0 que 0.00 para ser bastante aceptable.

Fiabilidad de los puntajes de diferencia

Los puntajes de diferencia a veces se llaman puntajes de cambio o puntajes de mejora. Muchas personas todavía usan puntajes de diferencia, y K y L han presentado una técnica mejorada para calcular puntajes de diferencia. Entonces, los puntajes de diferencia se discuten brevemente.

Si las personas comienzan un programa de instrucción o un programa de capacitación con puntajes marcadamente diferentes, la evaluación de las personas en términos de sus puntajes al final del programa pone a las personas que comienzan el programa con puntajes bajos en desventaja para las personas que comienzan el programa con puntajes altos. Entonces, los puntajes de diferencia a veces se calculan para determinar el grado en que el puntaje de cada persona ha cambiado con el tiempo, desde el principio hasta el final de un programa de instrucción o programa de capacitación. Los puntajes de diferencia se pueden calcular rápida y fácilmente como

$$\text{Puntajes de diferencia} = \text{Puntaje final} - \text{Puntaje inicial}$$

donde el puntaje inicial es al comienzo y el puntaje final es al final de un programa de instrucción o programa de capacitación.

El uso de puntajes de diferencia para la comparación entre grupos y el desarrollo de estándares de desempeño plantean dos problemas reales. Primero, las personas que se desempeñan bien inicialmente no tienen la misma oportunidad de lograr puntajes de grandes diferencias que las personas que comienzan mal. Por ejemplo, la persona que inicialmente corre una milla de 6 minutos tiene menos oportunidades de mejora que la persona que inicialmente corre una milla de 9 minutos. Segundo, los puntajes de diferencia son muy poco fiables. La fórmula para estimar la fiabilidad de los puntajes de diferencia ($Y-X$) es la siguiente:

$$R_{dd} = \frac{R_{xx}s_x^2 + R_{yy}s_y^2 - 2R_{xy}s_x s_y}{s_x^2 + s_y^2 - 2R_{xy}s_x s_y}$$

donde

X =	Puntaje inicial
Y =	Puntaje final
$R_{xx} - R_{yy}$ =	Coeficientes de fiabilidad de las pruebas X – Y
R_{xy} =	Correlación entre las pruebas X – Y
R_{dd} =	Fiabilidad de la diferencia de los puntajes
$s_x - s_y$ =	Desviaciones estándar de los puntajes inicial y final

El siguiente es un ejemplo de estimación de la fiabilidad de los puntajes de diferencia. Se administró una prueba con fiabilidad de .90 al comienzo y al final de un programa de capacitación. Como suele ser el caso, la fiabilidad de la prueba no cambió entre

las administraciones. La correlación entre los dos conjuntos de puntuaciones fue de 0.75. La desviación estándar fue 2.5 para los puntajes iniciales y 2.0 para los puntajes finales.

El siguiente es un ejemplo de estimación de la fiabilidad de los puntajes de diferencia. Se administró una prueba que arrojó una fiabilidad de .90 al comienzo y al final de un programa de entrenamiento. Como suele ser este caso, la fiabilidad de la prueba no cambió entre las dos aplicaciones. La correlación entre los dos conjuntos de puntuaciones fue de 0,75. La desviación estándar fue 2.5 para los puntajes iniciales y 2.0 para los puntajes finales.

$$R_{dd} = \frac{(0,90)(2,5)^2 + (0,90)(2,0)^2 - 2(0,75)(2,5)(2,0)}{(2,5)^2 + (2,0)^2 - 2(0,75)(2,5)(2,0)} = 0,63$$

De la fórmula se desprende que las medidas altamente fiables entre las cuales la correlación es baja son necesarias si el puntaje de diferencia va a ser fiable. Dado que los puntajes de diferencia generalmente se calculan a partir de los puntajes de una prueba aplicada al comienzo y al final de un programa de instrucción o entrenamiento, y la correlación entre estos dos conjuntos de puntajes es normalmente de .70 o más, los puntajes de diferencia altamente fiables parecen imposibles.

Se han presentado otras técnicas para calcular puntajes de diferencia donde se discuten los problemas al usar estas técnicas y se presenta una técnica de puntaje de diferencia que no tiene esos problemas. Con la nueva técnica, (1) cada persona tiene un puntaje inicial (I) y un puntaje final (F); (2) se calcula una puntuación de cambio (C) para cada persona donde $C = F - I$; (3) usando la predicción simple como se presenta en el capítulo 2 y disponible en un paquete como el SSPS, se predice una puntuación de cambio (C') para cada persona con la variable independiente o estimada de la puntuación inicial (I) para que $C' = (b)(I) + a$; y (4) se calcula un puntaje de diferencia (D) para cada persona, donde $D = C - C'$. Un puntaje de diferencia positiva es mejor que un puntaje de distancia de diferencia negativa, solo el puntaje de cambio (C) es mayor de lo que fue estimado.

EJERCICIOS

1. Dos características importantes de todas las mediciones son que sean fiables y objetivas. Describa la naturaleza básica de cada una de estas características.
2. Se utilizan dos métodos básicos para estimar la fiabilidad de una medición: fiabilidad de estabilidad y fiabilidad de consistencia interna. Describa brevemente cada método y el procedimiento que utiliza para estimar la fiabilidad.
3. En teoría, la fiabilidad es la relación entre la varianza del puntaje verdadero y la varianza del puntaje observado. La varianza de puntaje observado consiste en la varianza de puntaje de error y la varianza de puntaje verdadero. Describa estas fuentes básicas de varianza y la combinación que produce la estimación más alta de fiabilidad.
4. La objetividad, o fiabilidad del evaluador, se logra cuando los puntajes de dos o más jueces diferentes están de acuerdo. La correlación entre los puntajes es un coeficiente de objetividad. Resuma el procedimiento que produce alta objetividad.
5. La fiabilidad aceptable es esencial para todas las mediciones. Muchos factores afectan la fiabilidad de una medición, y algunos de estos se enumeran a continuación. Identifique las condiciones bajo las cuales se puede esperar la mayor fiabilidad para los siguientes factores:
 - a. Personas evaluadas
 - b. Longitud de la prueba
 - c. Entorno de la prueba
 - d. Administrador de la prueba
6. La fiabilidad aceptable es esencial para todas las mediciones. Muchos factores afectan la fiabilidad de una medición, y algunos de estos se enumeran a continuación. Identifique cómo cada uno de estos factores afecta el coeficiente de fiabilidad estimado obtenido para los puntajes de las pruebas con criterios de referencia.
 - a. Coeficiente de fiabilidad calculado
 - b. Azar
 - c. Composición el grupo

Capítulo 8. Validez

Se considera que la validez consiste en la interpretación del puntaje obtenido en una prueba y, por lo tanto, las interpretaciones que se hagan de los puntajes de las pruebas serían válidas o correctas. Como ejemplo, las interpretaciones en cuanto a la cantidad de fuerza de la cintura escapular y los brazos con base en los puntajes de la prueba de flexión-extensión de codos, son las que son válidas. Para que exista validez, el puntaje de la prueba debe ser fiable, pero la fiabilidad no garantiza interpretaciones válidas. Obviamente, para que las interpretaciones de los puntajes de las pruebas sean válidas, la prueba o el instrumento deben medir de manera consistente. Por ejemplo, si las personas saben que van a correr la prueba de una milla y han tenido experiencia previa en esa prueba, la carrera de la milla arroja puntajes fiables (estables de día a día), a partir de los cuales se pueden hacer interpretaciones válidas de la capacidad aeróbica, pero las interpretaciones con respecto a la velocidad de la carrera no serían válidas. Pero, ¿qué pasa si los puntajes en la prueba de la milla no son fiables (cambian de un día a otro)? Si, por ejemplo, un día una persona corre la milla en 13 minutos y dos días después la corre en 10 minutos. La interpretación de los 13 minutos sería muy diferente de las interpretaciones de los 10 minutos, y ninguna de las interpretaciones que se hagan pueden ser correctas. Las interpretaciones de las puntuaciones en la carrera de la milla no serán válidas en esta situación. Si las interpretaciones basadas en los datos de una prueba o instrumento de medición no son válidas, los datos son inútiles. Cualquier decisión tomada sobre las personas en función de esos datos será un error.

Validez de las pruebas referidas a normas

De lo que se ha presentado hasta ahora, debería ser evidente que no se validan las pruebas, sino que se recopilan evidencias para validar las interpretaciones hechas a partir de los puntajes de las pruebas. Por ejemplo, se puede interpretar la medida del porcentaje de grasa corporal de una persona de varias maneras y, según las interpretaciones, se debe recopilar evidencia de validez para respaldar una interpretación en particular. Al suponer que la suma de medidas de pliegues cutáneos de una mujer se traduce en 40% de grasa corporal, una interpretación de este porcentaje de grasa corporal es que el 40% del peso corporal de la mujer es tejido graso. Una segunda interpretación del 40% de grasa corporal podría ser que es demasiado alta y que la mujer está en riesgo de tener futuros problemas de salud. Para la primera interpretación, se debe recopilar evidencia de que existe una alta correlación entre la suma de las medidas de pliegues cutáneos y el porcentaje de grasa corporal medido, utilizando un criterio o medida estándar de oro, como la hidrodensitometría. La interpretación dos requiere evidencia de que una suma alta de medidas de pliegues cutáneos es un indicador de alto porcentaje de grasa corporal asociado con ciertos problemas de salud.

Cuando una prueba mide lo que pretende medir, son posibles interpretaciones válidas de los puntajes. Para tener validez, entonces, una prueba debe ser relevante para el rasgo que se mide y los puntajes de la prueba deben ser fiables.

Tipos y estimación de la validez

En las pruebas referidas a normas, por lo general se señalan tres tipos básicos de evidencia de validez: evidencia de validez de contenido, evidencia de validez de criterio y evidencia de validez de constructo. Sin embargo, en numerosas fuentes se señala que en realidad solo hay un tipo de validez: la validez de constructo. Es decir, todos los enfoques para recopilar evidencia de validez realmente constituyen diferentes formas de investigar la validez de constructo.

Basado en la definición antigua de validez, se solía hablar sobre tipos de validez. Ahora, usando una nueva definición de validez, se habla sobre los tipos de evidencia de validez. Además, se recomienda informar varios enfoques o varios tipos de evidencia de validez. Sin embargo, un indicador fuerte es mejor que varios indicadores débiles de validez; por ejemplo, la evidencia relacionada con el contenido solo es una evidencia débil de validez.

La validez se puede estimar de forma lógica o en forma estadística. Recientemente, la tendencia en educación y psicología ha sido alejarse del enfoque estadístico, mientras que la ciencia del ejercicio depende más de los enfoques estadísticos. Ello no significa necesariamente que la educación física y la ciencia del ejercicio deban abandonar el enfoque estadístico. La educación y la psicología trabajan predominantemente con pruebas de papel y lápiz, mientras que la educación física y la ciencia del ejercicio trabajan predominantemente con pruebas de rendimiento físico. El método para estimar la validez debe elegirse en función de la situación, no de la tendencia.

Evidencia de validez de contenido. Generalmente, a este enfoque se le ha llamado un enfoque lógico o un tipo lógico de evidencia. En educación física y ciencias del ejercicio, se denomina evidencia de validez lógica. Se establece mediante un examen las capacidades para medir y determinar si con la prueba o el instrumento se están, de hecho, obteniendo los puntajes que miden esas capacidades. Por ejemplo, una prueba de conocimiento escrita arroja puntajes sobre los cuales se pueden hacer interpretaciones válidas cuando sus preguntas son fiables, se basan en el material enseñado y muestran los objetivos educativos establecidos de la unidad. Aquí es donde se origina el término "evidencia de validez de contenido". El análisis de contenido de la prueba es un enfoque lógico. Las personas bien capacitadas en educación física y ciencias del ejercicio usan una lógica como esta en pruebas y medidas de rendimiento físico. Podrían indicar que el puntaje en 100 metros planos permite interpretaciones válidas de la velocidad de carrera porque, obviamente, los puntajes indican la velocidad con la que una persona puede correr. Para una habilidad multicomponente complicada, como un deporte de equipo, un consenso de jueces expertos (por ejemplo, entrenadores) puede proporcionar el único medio para determinar el grado de contenido o validez lógica.

Con seguridad, es cierto que el enfoque de evidencia de contenido se ha utilizado con éxito y a veces, mal utilizado. Idealmente, varios tipos diferentes de evidencia de validez (relacionada con el contenido, relacionada con el criterio, relacionada con la construcción), se proporcionan en una prueba, pero una validez sólida es mejor que varias débiles. Si es posible, la evidencia de validez de contenido debe usarse en conjunción con otra evidencia de validez.

La evidencia de validez de contenido parece haberse utilizado con bastante éxito en las pruebas de condición física y las baterías de pruebas de condición física. Por ejemplo, los puntajes en abdominales arrojan interpretaciones válidas de la fuerza abdominal porque la fuerza y la resistencia muscular abdominal son necesarias para hacer un abdominal. En un segundo ejemplo, los componentes del estado físico relacionado con la salud se seleccionaron lógicamente en función de cuáles son los problemas de salud más graves en la actualidad. Sin embargo, no se ha encontrado que el puntaje de la prueba de flexibilidad arroje evidencia válida de flexibilidad en los isquiotibiales y la espalda baja. Entonces, a veces lo que parece ser evidencia de validez de contenido realmente no lo es.

La evidencia de validez de contenido parece haber sido utilizada con menos éxito en pruebas complejas de destrezas deportivas. Algunas afirmaciones de evidencia de validez de contenido para interpretar los puntajes de ciertas pruebas de destrezas deportivas y de las baterías de pruebas de destrezas deportivas han sido cuestionadas ocasionalmente. El voleo en la pared en voleibol es un ejemplo. Otro ejemplo es una batería integral de pruebas de destrezas para softbol, baloncesto o voleibol, construidas con base en los componentes básicos del deporte.

La evidencia de validez de contenido, entonces, se basa únicamente en la toma de decisiones subjetivas, lo que ha llevado a la crítica de este tipo de evidencia de validez. Sin embargo, esta evidencia de validez no es la única que descansa en una decisión subjetiva; La evidencia de validez relacionada con los criterios a menudo se determina utilizando las decisiones subjetivas de un juez o de unos jueces expertos.

Evidencia de validez de criterio. La evidencia de validez de criterio se obtiene determinando la correlación entre los puntajes de una prueba y algún criterio específico, lo que constituye un procedimiento tradicional para obtener la evidencia de validez. Ello implica el cálculo del coeficiente de correlación producto momento de Pearson o algún tipo de correlación similar entre los puntajes de la prueba y los obtenidos en una medida de criterio. La correlación resultante o coeficiente de validez es un estimado de la validez de las interpretaciones basadas en los puntajes de la prueba. Cuando el coeficiente de validez está cerca de 1, la prueba mide en forma similar a la medida de criterio y las interpretaciones basadas en los puntajes de la prueba se consideran válidas; si el coeficiente está cerca de 0, la prueba y el criterio miden aspectos diferentes y las interpretaciones con base en los puntajes de la prueba tienen poca validez.

Un paso crucial para la obtención de la evidencia de validez de criterio es la selección de una medida de criterio apropiada. La calificación de expertos y la posición en un

torneo son usadas en las pruebas de destrezas deportivas. Asimismo, criterios predeterminados como el $VO_2\text{máx}$ y la composición corporal son criterios utilizados en las ciencias del ejercicio.

- *Calificación de expertos.* Una de las medidas de criterio es la calificación subjetiva de uno o más expertos. Como un juez puede pasar por alto alguno de los factores del desempeño de un sujeto o puede tener algún prejuicio que afecta sus calificaciones, la suma de las calificaciones de varios jueces capacitados es la más confiable medida de criterio. Para ello es vital que haya un alto grado de objetividad entre los jueces; posteriormente se correlacionan los puntajes de las pruebas con las calificaciones subjetivas del juez o los jueces, utilizando el coeficiente de correlación producto-momento de Pearson, u otro coeficiente de correlación apropiado.

Las calificaciones subjetivas generalmente toman la forma de rangos grupales o una escala de puntos. Utilizando los rangos grupales, las personas en el grupo se ordenan de mejor a peor en función del rendimiento. Usando una escala de puntos, las personas en el grupo se puntúan individualmente en una escala de puntos numerada, como lo harían en un evento de gimnasia. Los rangos grupales se vuelven inviables a medida que aumenta el tamaño del grupo. Además, es preferible un sistema de escala de puntos, ya que no requiere que una persona se clasifique mejor y otra peor. Por lo general, se obtiene una mayor validez con un sistema de puntaje que con los rangos de grupo, siempre que el sistema de puntaje esté bien definido para que sea posible una alta objetividad.

Las calificaciones de expertos se han usado comúnmente para estimar la validez de las interpretaciones basadas en los puntajes de las pruebas de destrezas. Una prueba de destrezas tiene ciertas ventajas sobre las pruebas que deben ser juzgadas por expertos. Si los jueces estén calificados, proporcionan un buen indicador del desempeño en las destrezas. Sin embargo, en la práctica es difícil encontrar incluso un experto, y mucho menos varios, que tenga tiempo para evaluar subjetivamente un grupo de estudiantes. Además, las pruebas de destrezas suelen ser más fáciles de administrar y requieren menos tiempo que las evaluaciones subjetivas.

- *Clasificación en torneos.* Una segunda medida de criterio es la clasificación en torneos. Los puntajes de una prueba se correlacionan (utilizando el coeficiente de correlación producto-momento de Pearson, u otro coeficiente de correlación apropiado) con la clasificación que obtengan en un torneo los estudiantes que están siendo medidos. Este método asume que la clasificación en el torneo es un buen indicador de la habilidad general en la destreza que se examina, especialmente en disciplinas individuales o en pareja como el tenis o el bádminton. Muchos docentes pueden estar tratando de encontrar una prueba que suministre la misma información que un torneo; tal prueba hace innecesario organizar torneos round-robin para evaluar los niveles de destreza de los estudiantes.

- *Criterio predeterminado.* La medida de criterio final es la puntuación de un instrumento conocido y bien aceptado. La evidencia de la validez de las interpretaciones basadas en los puntajes de las pruebas entonces se proporciona determinando el coeficiente de correlación momento-producto de Pearson entre los puntajes de la prueba que se propone y los

puntajes del instrumento aceptado. El método generalmente se usa para desarrollar una prueba que sea más rápida y / o más fácil de administrar que el instrumento aceptado.

Idealmente, la medida del criterio es aquella que todos reconocen como el "estándar de oro" (el último o mejor estándar). Si no hay un estándar de oro, y a menudo no lo hay, se debe seleccionar una de las medidas de criterio aceptables disponibles. Es posible que no todas las personas están de acuerdo con qué medida de criterio utilizar, por cuanto la medida de criterio influye en el valor del coeficiente de validez. Es por eso que es difícil realizar una validación utilizando un criterio predeterminado, ya que el coeficiente de validez podría no ser alto.

Por ejemplo, el método de criterio predeterminado se ha utilizado para validar las interpretaciones de la composición corporal basadas en pliegues cutáneos. Se pueden obtener medidas aceptadas del porcentaje de grasa corporal en un laboratorio al pesar a una persona bajo el agua. Una persona con un alto porcentaje de grasa corporal pesa menos bajo el agua porque el tejido graso es menos denso que los músculos y los huesos. Las mediciones de pliegues cutáneos son fáciles de obtener y se correlacionan altamente (r es al menos 0,80) con el porcentaje de grasa determinado por el método más complicado de pesaje en inmersión. Además, las mediciones de pliegues cutáneos se pueden usar en personas que temen ser pesadas bajo el agua. Al obtener medidas seleccionadas de los pliegues cutáneos, podemos estimar el porcentaje de grasa corporal de la persona usando ecuaciones de regresión que a menudo se usan para obtener coeficientes de validez. En estas situaciones, Y es el puntaje de criterio (variable dependiente), y los diversos puntajes utilizados para estimar Y son los puntajes X (variables independientes). En el ejemplo anterior, Y sería el porcentaje de grasa corporal determinado por laboratorio con el peso bajo el agua, y las puntuaciones X serían las medidas del pliegue cutáneo.

La prueba que se examina en el proceso de validación es una prueba práctica o de campo. El criterio predeterminado es una puntuación de un instrumento de laboratorio u otro instrumento que es muy costoso, difícil y / o requiere mucho tiempo para administrar, y puede requerir una considerable capacitación o experiencia de la persona que aplica el instrumento. Hay pocos estándares de oro en educación física y ciencias del ejercicio. A menudo no hay un criterio predeterminado aceptable, o hay varios criterios predeterminados posibles, pero las personas no se ponen de acuerdo sobre cuál usar. Nuevamente, esta es la razón por la cual es difícil realizar una validación utilizando un criterio predeterminado. Se supone que una prueba de abdominales mide la fuerza y la resistencia abdominal. Pero no existe una medida de criterio que sea estándar de oro para la fuerza y resistencia abdominal, y no es probable que se tengan expertos en el campo para acordar una medida de criterio aceptable.

Una mirada a la literatura sobre la medición de la actividad física, reflejará muchas medidas de criterio diferentes utilizadas para recopilar evidencia de validez de criterio. No hay acuerdo sobre cuál medida de criterio utilizar. Quizás esto se deba a diferentes interpretaciones de los puntajes de las pruebas de actividad física y los cuestionarios. Los puntajes pueden interpretarse como la cantidad de actividad, la cantidad de calorías gastadas, el nivel de condición física, etc. De esto se trata la validez: la validez de las interpretaciones basadas en los puntajes de las pruebas.

Evidencia de validez de constructo. Los enfoques de validez de contenido y de validez de criterio son procedimientos establecidos para la determinación de la validez. El enfoque de evidencia de validez de constructo es el más nuevo y más complejo procedimiento de validación, siendo utilizado cada vez más en educación física, las ciencias del ejercicio y demás campos relacionados. El enfoque de validez de constructo es más utilizado en pruebas abstractas que en pruebas concretas; una prueba abstracta mide algo que no es directamente observable. Las actitudes hacia la actividad física son rasgos humanos que no son fácilmente observables ni comprensibles para alguien sin entrenamiento. Por otro lado, el número de tiros libres convertidos en baloncesto en un total de 100 intentos es una prueba concreta de la destreza del lanzamiento de tiro libre.

El enfoque de validez de constructo se basa en el método científico. Primero, existe la presunción de que una prueba o pruebas pueden medir algún rasgo abstracto. En segundo lugar, se desarrolla una teoría para explicar tanto el constructo como las pruebas que lo miden. Finalmente, se aplican varios procedimientos estadísticos para confirmar o rechazar la teoría.

La validez de constructo se determina juzgando el grado en que la información teórica y estadística respalda los supuestos constructos. Este procedimiento les da a los docentes de educación física y a los especialistas del ejercicio una manera de evaluar sus pruebas. Por ejemplo, al considerar una prueba de destrezas en natación se plantea la siguiente pregunta, ¿es cierto que mientras mayor sea el puntaje en una prueba se poseen mejores destrezas en natación? Luego se aplica la prueba a nadadores avanzados e intermedios y, como se sabe que los nadadores avanzados son mejores nadadores, se tiene la evidencia de validez de constructo siempre que la media de los nadadores avanzados sea superior a la media de los nadadores intermedios. Probablemente se debe realizar una prueba estadística de la significancia de la diferencia entre las dos medias en alfa igual a 0.01.

Un segundo ejemplo de este enfoque es la comparación del desempeño de un grupo antes y después de la instrucción o del entrenamiento. Esto implica evaluar la habilidad inicial de un grupo, aplicar algunas técnicas de instrucción o entrenamiento durante varias semanas y luego volver a evaluar la habilidad final del grupo. Dado que el grupo debería haber mejorado su capacidad como resultado de la instrucción o del entrenamiento, tenemos evidencia de validez de constructo si la media de la prueba final es significativamente mejor que la media de la prueba inicial (ver la prueba *t* de medidas repetidas).

En todos los casos, se recolecta evidencia de validez para una población definida de personas. Todas las personas utilizadas para recopilar la evidencia de validez deben pertenecer a la población o tener habilidades similares a las de las personas que pertenecen a la población. Esto es particularmente cierto cuando se comparan grupos que difieren en habilidad al igual que los nadadores que hemos mencionado. Si se utilizan grupos con capacidades muy diferentes para recopilar evidencia de validez, siempre habrá una diferencia significativa entre los grupos en las puntuaciones medias de las pruebas. Por ejemplo, si se obtuvieron evidencias de validez de una prueba que se utilizará en individuos de 18 años y se comparan individuos entre 6 y 18 años en términos de casi cualquier prueba

de habilidad física, los individuos de 18 años serían superiores. Es difícil creer que cualquier persona de 18 años tenga puntajes en los exámenes tan pobres como los de las personas de 6 años. Aunque se obtuviesen diferencias significativas entre los grupos, la evidencia no es generalizable para cualquier intento de usar la prueba.

Si, por ejemplo, la prueba de natación que se realiza a los nadadores intermedios y avanzados se va a utilizar para la selección del equipo de natación, se obtendría información útil al comparar los puntajes de desempeño de los miembros del primer equipo con los puntajes de los miembros del segundo equipo. Las diferencias significativas entre los grupos en la puntuación media proporcionarían evidencia de validez para la interpretación de las puntuaciones de la prueba.

Como información adicional, el análisis factorial es un procedimiento estadístico que puede usarse para identificar constructos y las pruebas que producen puntajes que conducen a interpretaciones válidas de la capacidad de los individuos en términos de constructos aislados. En este contexto, el factor aislado del análisis es un constructo abstracto. El análisis factorial a menudo se ha utilizado para identificar los componentes básicos de la aptitud física o de una habilidad deportiva.

- *Procedimientos de validación.* La validación del constructo se realiza a través de tres etapas que incluyen el tipo de evidencia que se necesita recolectar en cada una de ellas. Las tres etapas son (a) la etapa de evidencia de la definición, la etapa de la evidencia confirmatoria, y la etapa de revisión de la teoría.

La etapa de la evidencia de la definición examina el grado en que los aspectos operacionales de la prueba representan los aspectos teóricos del constructo, empleándose el enfoque lógico o de contenido para recolectar las evidencias de validez; en este paso se revisan las propiedades de la prueba más que sus puntajes o resultados. Debe desarrollarse la definición del constructo antes de encontrar o desarrollar una prueba que mida ese constructo.

En la segunda etapa, la etapa de evidencia confirmatoria, el investigador debe confirmar la definición del constructo de la primera etapa. Aquí se utiliza cualquiera de los dos enfoques de recolección de evidencia de validez, la evidencia de validez de criterio o la evidencia de validez de constructo.

La tercera etapa consiste en la revisión de la teoría. En este punto la teoría del constructo que se quiere medir se examina utilizando el método científico. En situaciones reales los constructos no existen en solitario por lo que muchos factores influyen en el constructo que interesa, factores que son llamados determinantes del constructo. Asimismo, el constructo que se estudia influye en otros constructos, lo que se llama efectos del constructo. Los determinantes y los efectos de un constructo deben ser investigados en esta etapa y es probable que tengan que estudiarse simultáneamente. Para un investigador esto es un proyecto de gran envergadura. Un profesional debe ser consciente de que la investigación debe realizarse con relación a los principales determinantes y efectos de un constructo, sin que sea de gran profundidad. De esta manera el establecimiento de la validez es un proceso continuo e interminable.

Los puntajes de una prueba tienen que ser fiables para que sean válidos. El coeficiente de validez se relaciona directamente con la fiabilidad entre la prueba que se investiga y la medida de criterio. El coeficiente de validez máximo posible entre dos pruebas se estima como sigue:

$$r_{x,y} = \sqrt{(r_{x,x})(r_{y,y})}$$

donde $r_{x,y}$ es la correlación entre las pruebas X , Y , (X es una nueva prueba, Y es el estándar de oro, el criterio de medida), $r_{x,x}$ es el coeficiente de correlación de la prueba X , y $r_{y,y}$ es el coeficiente de fiabilidad de la prueba Y . Por ejemplo, si la fiabilidad en ambas pruebas es 0.90, el coeficiente de validez máximo posible sería 0.90:

$$r_{x,y} = \sqrt{(0,90)(0,90)} = \sqrt{0,81} = 0,90$$

Si la fiabilidad de la prueba X es 0.90 y de la prueba Y es 0.40, el coeficiente de validez máximo posible sería mucho más bajo, sólo 0.60:

$$r_{x,y} = \sqrt{(0,90)(0,40)} = \sqrt{0,36} = 0,60$$

Validez y Puntaje de Criterio

Selección del puntaje de criterio. Un puntaje de criterio es la medida utilizada para indicar la capacidad de una persona. A menos que una prueba tenga una fiabilidad perfecta, es el mejor indicador cuando se desarrolla a partir de más de un intento. Las pruebas con intentos múltiples son comunes en educación física al medir pliegues cutáneos, flexibilidad, saltos, lanzamientos y carreras cortas, pruebas cuyo rendimiento no se ve afectado negativamente por la fatiga. Múltiples intentos en pliegues cutáneos, flexibilidad y pruebas de resistencia son comunes en las ciencias del ejercicio. La selección de la puntuación del criterio cuando se administran múltiples intentos de la prueba se discute en el capítulo de fiabilidad. Una revisión de esta información podría ser útil. Recuerde que la fiabilidad es esencial para la validez, pero la fiabilidad no garantiza la validez. Por lo tanto, la puntuación del criterio debe seleccionarse teniendo en cuenta tanto la fiabilidad como la validez. Desde 1990, algunos desarrolladores de pruebas han descubierto que, para las pruebas de habilidad deportiva con intentos múltiples, la mejor puntuación y la puntuación media son similares en fiabilidad; así que han usado la mejor puntuación como puntaje de criterio porque es más rápido y fácil de obtener. Esto es preocupante si la mejor puntuación es mucho menos válida que alguna otra puntuación de criterio.

El rol de la preparación. La preparación es un determinante importante de la validez de las interpretaciones basadas en un puntaje de criterio. Un puntaje de criterio es el mejor indicador de la capacidad real si las personas se miden solo después de que entienden completamente cómo ejecutar una prueba. Por ejemplo, supongamos que un instructor le anuncia un día que medirá su capacidad aeróbica al hacer que corra una milla para deter-

minar su tiempo. Si nunca ha participado en carreras de distancia, no sabría cómo mantener el ritmo, cómo correr de manera relajada, cómo evitar ser encerrado por otros corredores o cómo correr cerca de la curva interior. Su puntaje ese día no sería un indicador real de su capacidad aeróbica. Si te hicieran la prueba nuevamente al día siguiente, tu puntaje mejoraría, no porque tu capacidad aeróbica haya mejorado, sino porque habrías aprendido a abordar la prueba. Y en un tercer día, podrías hacerlo mejor aún.

La investigación indica que las personas a menudo necesitan un día de práctica antes de las pruebas de rendimiento físico para familiarizarse con una prueba, incluso si han tenido experiencia con ella. Esta práctica permite un período de reaprendizaje y refuerzo que es necesario porque la retención de habilidades nunca es perfecta. Las investigaciones han encontrado que una puntuación media de los intentos en una prueba de intentos múltiples será mayor si varias pruebas de práctica (llamadas prácticas de calentamiento para la prueba) se hacen justo antes de la prueba.

Validez de las pruebas referidas a criterio

Generalmente, la misma definición de validez de una prueba referida a normas se aplica a una prueba referida a criterios. Las técnicas para estimar la validez de una prueba referida a criterios son diferentes de las técnicas de una prueba referida a normas. Recuerde, una prueba basada en criterios se utiliza para clasificar a las personas como competentes o no competentes, o como aprobadas o reprobadas.

Una técnica para estimar la validez de una prueba referida a criterios se llama validez referida a dominio. El término dominio se refiere a la habilidad o comportamiento del criterio que se supone que mide la prueba. Por lo tanto, las interpretaciones basadas en los puntajes de una prueba referida a criterios se validan lógicamente al mostrar que la prueba mide la capacidad o el comportamiento del criterio. Esta técnica es básicamente la misma que la técnica de contenido o validez lógica discutida anteriormente.

Otra técnica para estimar la validez con una prueba referida a criterios se llama validez de decisión. Esto trata con la precisión de la prueba al clasificar a las personas como competentes o no competentes. Esencial para esta técnica es la capacidad de clasificar con precisión a las personas como competentes o no competentes, independientemente de la prueba. Esto es a menudo muy difícil de hacer, y se debe desarrollar la capacidad de hacerlo. Entonces, independientemente de la prueba y también como resultado de la prueba, se clasifican las personas como competentes o no competentes. Estas clasificaciones permiten generar una tabla de clasificación de doble entrada como la tabla 8.1. Todos los paquetes de programas informáticos estadísticos tienen un programa de tabulación cruzada, por ejemplo, el SPSS, que proporciona una tabla similar.

A partir de esta tabla de clasificación de doble entrada, la clasificación de las probabilidades de resultado (C) se puede calcular como una estimación de validez.

$$C = \frac{A + D}{A + B + C + D}$$

		Clasificación verdadera	
		Competente	No competente
Prueba de clasificación	Competente	A	B
	No competente	C	D

Tabla 8.1 Tabla de estimación de la validez de una prueba referida a criterios

Ahora, se presenta un ejemplo del cálculo de la estimación de la calidez de una prueba referida a criterios con los datos de la tabla 8.2:

		Clasificación verdadera	
		Competente	No competente
Prueba de clasificación	Competente	40	5
	No competente	10	20

Tabla 8.2 Tabla de estimación de la validez de una prueba referida a criterios

$$C = \frac{40 + 20}{40 + 5 + 10 + 20} = \frac{60}{75} = 0,80$$

Si C es igual a 0.50, indica que la clasificación por la prueba no fue mejor que el azar, y las interpretaciones en cuanto a la competencia basadas en los puntajes de las pruebas no serían muy válidas. Por esta razón, un valor de C más cercano a 1.0 que a 0.50 es lo más deseable.

Otra estimación de validez que se puede calcular a partir de una tabla de clasificación de doble entrada es phi (ϕ). Tomando nuevamente los datos de la tabla 8.2 para el cálculo correspondiente, tenemos:

$$\phi = \frac{(AD) - (BC)}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$$

Sustituyendo:

$$\phi = \frac{(40)(20) - (5)(10)}{\sqrt{(40 + 5)(10 + 20)(40 + 10)(5 + 20)}} = \frac{800 - 50}{\sqrt{1687500}} = \frac{750}{1299,04} = 0,58$$

Los valores de ϕ pueden variar de -1.0 a 1.0, por lo que los coeficientes de ϕ pueden interpretarse como cualquier otro coeficiente de correlación. Sin embargo, debido a que ϕ es la correlación entre dos variables dicotómicas, no se esperan valores altos de ϕ . Se requiere un valor positivo de ϕ para la validez. Un valor negativo de ϕ indica que la prueba está clasificando individuos en oposición a la forma en que el criterio clasifica a los individuos.

Podría surgir un problema al clasificar a las personas como competentes o no competentes, independientemente de la prueba (la clasificación real en la tabla de doble entrada). En algunas situaciones de instrucción, en la clasificación competente podrían estar personas que hayan recibido instrucción, y en la clasificación no competente podrían estar personas que no hayan recibido instrucción. Además, en una situación de instrucción, el no competente podría estar en la clasificación de las personas antes de recibir la instrucción, y el competente podría estar en la clasificación de las mismas personas después de recibir la instrucción. En algunas situaciones de ciencias del ejercicio, puede existir una medida estándar o de laboratorio aceptada que se puede utilizar para clasificar a las personas en términos de pertenencia a un grupo real.

La validez de los estándares referidos a criterios de las pruebas de condición física relacionadas con la salud de los jóvenes, podría investigarse para determinar si las clasificaciones son precisas para ellos. Esto es validez concurrente, "por el momento". También, la validez de los estándares referidos a criterios de las pruebas de condición física relacionadas con la salud de los jóvenes podría ser investigada para determinar si las clasificaciones son precisas cuando los jóvenes se hagan adultos. Esto es validez predictiva, o "para el futuro". Entonces, como hemos dicho muchas veces, la validez se aplica a las interpretaciones de los puntajes de los exámenes, no a los puntajes en sí. La validez predictiva estaría determinada por la clasificación de la prueba en la tabla de doble entrada que es la clasificación del sujeto (apto o no apto) como un joven y la clasificación verdadera en la tabla de doble entrada es la clasificación del sujeto (apto o no apto) como un adulto. Desafortunadamente, existen pocos datos como este.

EJERCICIOS

1. Defina validez y los métodos utilizados para estimarla
2. Para que una prueba sea válida, las características más importantes de todo tipo de medición es que deben ser fiables, objetivas y que las interpretaciones se basen en las mediciones.
3. Describa la naturaleza básica de la validez.
4. Describa brevemente los tipos de evidencia de validez.
5. Un principio básico de la teoría de la medición es que los puntajes de la prueba deben ser primero fiables para que su interpretación sea válida. ¿A qué se debe esto?
6. ¿Qué efecto tiene la objetividad sobre la validez?
7. La evidencia de la validez de criterio implica la determinación de la correlación entre una prueba propuesta y un criterio de medida. Señale dos medidas de criterio que pueda ser utilizada en educación física y las ciencias del ejercicio.
8. Muchas pruebas de destrezas de condición física y de destrezas deportivas requieren de varios intentos. ¿Cómo puede una persona estar segura de que las interpretaciones del puntaje criterio son tan válidos como sea posible?
9. Establezca la validez de una prueba referida a criterios de condición física según los procedimientos de C y ϕ , de acuerdo con los siguientes datos: A=55; B= 35; C= 25; D= 35.

Capítulo 9. Nociones estadísticas para la investigación de la actividad física

Investigación y análisis estadístico

El diseño de la investigación y el análisis estadístico se basan en principios de medición. Están tan entrelazados que es difícil discutir uno sin referirse al otro. Hay técnicas estadísticas diseñadas para ayudar en la evaluación de datos. Cómo se recopilan los datos, qué instrumentos se usan y cómo se controlan las variables son parte del diseño de la investigación.

La investigación es una técnica especial para resolver problemas. Identificar el problema es una parte crítica de la investigación. Un problema generalmente comienza con preguntas como "Me pregunto si ...", "Me pregunto por qué ..." o "Me pregunto qué ...". Si no se encuentra la respuesta a la pregunta, entonces hay un problema que puede ser investigado. Primero, podríamos preguntarle a un experto en el tema sobre la pregunta. Si el experto sabe la respuesta, el problema está resuelto. Si el experto no lo sabe, se puede visitar la biblioteca, hablar con otros colegas o consultar por internet. Si se encuentra la respuesta en estas consultas, el problema está resuelto. Pero cuando no se encuentra la respuesta en la biblioteca, por internet o hablando con otras personas, entonces la única forma de resolver el problema es realizar una investigación.

Muchas personas piensan en un experimento cuando escuchan la palabra investigación, pero la investigación no se limita a explorar problemas de diseño científicos y experimentales. El problema puede resolverse de varias formas, entre ellas la investigación histórica, la investigación descriptiva o la investigación experimental. Como ejemplo, se presentan algunas características de la investigación experimental.

El diseño experimental es el proceso de investigación que implica manipular y controlar eventos o variables para resolver un problema. Para comenzar una investigación, se necesita un plan. El diseño de la investigación es el plan que establece la manera en que se recopilarán y analizarán los datos. A menudo, el plan exige la resolución de sub-problemas o el estudio piloto (recopilación de datos preliminares) para determinar la viabilidad del diseño de la investigación. Es un error común de los investigadores principiantes saltar a la recopilación de datos antes de hacer una planificación adecuada y un estudio piloto.

Una vez que se ha identificado el problema y se ha completado la revisión bibliográfica, si el problema no se ha resuelto se debe estar listo para formular una hipótesis. Una hipótesis es una suposición que se basa en investigaciones previas o hechos conocidos y que puede ser probada por el diseño experimental. La hipótesis debe establecerse de tal

manera que se pueda realizar un análisis estadístico de los datos para determinar la probabilidad (o probabilidades) de obtener los resultados previstos si la hipótesis es verdadera.

Hay momentos en que el objetivo es determinar si varios grupos son diferentes, es decir, hombres versus mujeres, o no atletas versus atletas. El objetivo pudiese ser determinar si un grupo ha mejorado su capacidad debido a un tratamiento como un programa de ejercicios. Los investigadores por lo general utilizan pruebas t (que no deben confundirse con las puntuaciones T) y el análisis de varianza (ANOVA) para determinar si existe una diferencia lo suficientemente grande entre dos o más grupos en el rendimiento medio para concluir que los grupos no son iguales a ese nivel. Además, las pruebas t y ANOVA se pueden usar para determinar si un grupo ha mejorado su capacidad.

Teoría e hipótesis

Una teoría es una creencia con respecto a un concepto o una serie de conceptos relacionados. Las teorías no son necesariamente verdaderas o falsas. Solo son productivas o no productivas en la producción de hipótesis. Las teorías fructíferas producen muchas hipótesis que pueden ser probadas mediante el proceso científico de experimentación. Una hipótesis es más específica que una teoría y se puede probar en un experimento. Cuando se completa el experimento, se pueden establecer las probabilidades de que la hipótesis sea correcta o incorrecta. Se determinan estas probabilidades mediante el uso de procedimientos estadísticos.

Cuando una teoría produce muchas hipótesis, la mayoría de las cuales sino todas se consideran correctas (por estudios originales y de confirmación), la teoría se acepta como verdadera. Por lo general, este proceso pasa por muchas etapas, con revisiones de la teoría a medida que avanza el proceso. El proceso puede llevar años, décadas o incluso siglos. Finalmente, la teoría es aceptada como la interpretación más correcta de los datos que se han observado. Cuando muchas o la mayoría de las hipótesis producidas por cierta teoría son rechazadas o no pueden ser confirmadas por otros científicos imparciales, la teoría es revisada o abandonada.

Un ejemplo de este proceso en el campo del comportamiento motor se encuentra en el concepto de práctica mental o visualización. Durante muchos años, los científicos y los atletas creyeron, pero no pudieron probar, que el aprendizaje de las destrezas motoras se lograba mejor mediante una combinación de procesos mentales y físicos. Los atletas parecían saber intuitivamente que podían mejorar el rendimiento al visualizarse ejecutando la habilidad. Esta teoría produjo muchas hipótesis, que fueron probadas. Una hipótesis popular era que, si una persona solo pensaba en una habilidad física, sin haberla realizado físicamente, su desempeño en la habilidad mejoraría.

Algunos estudios a mediados del siglo XX produjeron algunas pistas de que esto pudiera ser cierto. En esta área se publicó en 1949 un estudio que demostró diferencias significativas en el rendimiento físico entre los sujetos de control y los sujetos que se visualizaban realizando una habilidad novedosa. Pronto, otros estudios con hipótesis similares, pero modificadas, se agregaron a la base de conocimiento.

En la década de 1960 se publicaron muchos estudios basados en la teoría de la visualización, muchos de los cuales confirmaron la conclusión de que los grupos de visualización funcionan significativamente mejor que los grupos de control, pero no tan bien como los grupos de práctica física. La teoría, o creencia general, produjo hipótesis comprobables que fueron respaldadas por experimentos. Actualmente, la teoría de que la visualización mejora el rendimiento de las habilidades físicas está bien aceptada.

¿Qué es una hipótesis? Es una afirmación acerca de un parámetro de la población. En el análisis estadístico se hace una afirmación, es decir, se establece una hipótesis y luego se sigue con una prueba para verificar la afirmación, o para determinar que no es cierto. La hipótesis estadística se define así:

Hipótesis: Enunciado acerca de un parámetro de la población, que se desarrolla con el propósito de realizar pruebas.

Prueba de hipótesis

La hipótesis que impulsa la investigación se llama hipótesis de investigación y está simbolizada por H_1 . Por lo general, predice relaciones o diferencias entre los sujetos de los grupos. Después de todo, si no se creyera que existieran relaciones o diferencias, entonces probablemente no se realizaría el experimento en primer lugar. Por otra parte, H_1 no es la hipótesis que normalmente se prueba mediante análisis estadístico.

La hipótesis a ser probada estadísticamente se llama hipótesis nula, o H_0 . La hipótesis nula no predice ninguna relación o diferencia entre los grupos. Si H_0 es verdadero, entonces H_1 es falso y viceversa. Son proposiciones mutuamente excluyentes. Esencialmente, la hipótesis nula establece que cualquier relación o diferencia que pueda observarse mediante la medición es el resultado de ocurrencias aleatorias.

El análisis estadístico reporta la probabilidad de que los resultados ocurran tal como los observamos si la hipótesis nula es cierta. Por ejemplo, las estadísticas pueden indicar que la probabilidad de que la hipótesis nula sea verdadera es menor que el 1 por ciento. Si la probabilidad de que H_0 sea verdadera es pequeña, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa o de investigación (H_1). Si la probabilidad de que H_0 sea verdadera es grande, aceptamos la hipótesis nula.

La hipótesis nula se usa normalmente para evaluar los resultados porque los investigadores generalmente no están seguros de cuáles serán los resultados del experimento. Si se conocieran los resultados del experimento antes de comenzar, no sería necesario hacerlo. Pero los investigadores comienzan un experimento con una idea en mente, y esta idea está representada por la hipótesis de la investigación. Por lo general, la hipótesis de investigación impulsa el experimento, pero la hipótesis nula es la que se prueba.

El investigador debe probar la hipótesis nula a menos que una evidencia sólida de otra investigación, de literatura relacionada, de la opinión de expertos o de otras fuentes confiables, sugiera que se espera una diferencia significativa. La opinión (tal vez sesgada)

del investigador no es razón suficiente por sí sola para probar la hipótesis de la investigación.

Se rechaza H_0 y se acepta H_1 cuando las diferencias o relaciones entre las variables se establecen más allá de una duda razonable, o en un nivel aceptable de confianza. Antes de recopilar datos, se debe establecer un nivel de confianza. Por ejemplo, se puede decidir rechazar la hipótesis nula si las probabilidades en su contra son mejores de 95 a 5. Dicho de otra manera, p (la probabilidad de que H_0 sea verdadero) es ≤ 0.05 ($p \leq 0.05$). Esto significa que las probabilidades de que H_0 sea verdadera son menores o iguales al 5 por ciento.

Pasos de la prueba de hipótesis

La prueba de hipótesis se inicia con una afirmación o suposición sobre un parámetro de la población, como la media (μ) de la población. Esta declaración se conoce como hipótesis.

Una hipótesis podría ser que el sueldo medio mensual de los docentes de educación física de un país sea de 100 dólares. Prácticamente, es imposible preguntarle a cada docente para verificar que el sueldo medio sea de 100 dólares, ya que el costo logístico y financiero de encuestar a todos los docentes de educación física es altísimo.

Para comprobar la validez de la afirmación ($\mu = 100$ dólares), es necesario seleccionar una muestra de la población, calcular la media de la muestra y aceptar o rechazar la hipótesis, con base en una serie de reglas para decidir. Por ejemplo, un sueldo medio de 50 dólares generaría el rechazo de la hipótesis. Ahora, si el sueldo medio de la muestra es de 96 dólares, ¿está lo suficientemente cerca de 100 dólares como para aceptar el supuesto de que la media de la población es 100 dólares? ¿Se puede atribuir la diferencia entre ambas medias a un error de muestreo, o bien esa diferencia es significativa desde el punto de vista estadístico?

La prueba de hipótesis es un procedimiento que se basa en la evidencia de las muestras y en la teoría de probabilidades para determinar si la hipótesis es un enunciado razonable.

Procedimiento para probar una hipótesis. La prueba de hipótesis proporciona una especie de prueba más allá de una duda razonable, por lo que se siguen procedimientos específicos de evidencia. El primer paso consiste en plantear la hipótesis que se prueba, y esta es siempre la hipótesis nula.

La hipótesis nula se designa como H_0 , leyéndose “H subíndice cero”; la letra mayúscula H significa hipótesis, y el subíndice cero supone “sin diferencia”. Por lo general, la hipótesis nula incluye el término “no”, y significa que “no hay cambio” o diferencia. La hipótesis nula se plantea para el propósito de la prueba: se rechaza o se acepta.

La hipótesis nula no se rechaza a menos que los datos de prueba proporcionen evidencias convincentes de que es falsa. La imposibilidad de rechazar la hipótesis nula no

demuestra que H_0 sea verdadera, significa que no fue posible rechazar H_0 . Para demostrar que la hipótesis nula es verdadera, habría que contar toda la población, lo cual no es posible. La alternativa consiste en tomar una muestra de la población. En el ejemplo del sueldo de los docentes de educación física, la hipótesis nula se escribiría $H_0: \mu = 100$, y se intentaría probar que el sueldo medio de la población (\$ 100) no presenta diferencias significativas con el sueldo medio de la muestra (\$ 96).

Con frecuencia, la hipótesis nula se inicia con la afirmación “No existe una diferencia significativa entre...” o “El sueldo medio de los docentes de educación física no tiene una diferencia significativa de...”. Para aclarar si existe o no una diferencia, se realiza una prueba de significación o prueba de hipótesis. La pregunta clave es: ¿Se trata de una diferencia significativa, o la diferencia se debe al azar (muestreo)?

Para realizar la prueba se requiere de la hipótesis alternativa, que describe la conclusión a la que se llegase si se rechaza la hipótesis nula. La hipótesis alternativa se acepta si los datos de la muestra proporcionan suficiente evidencia estadística de que la hipótesis nula es falsa.

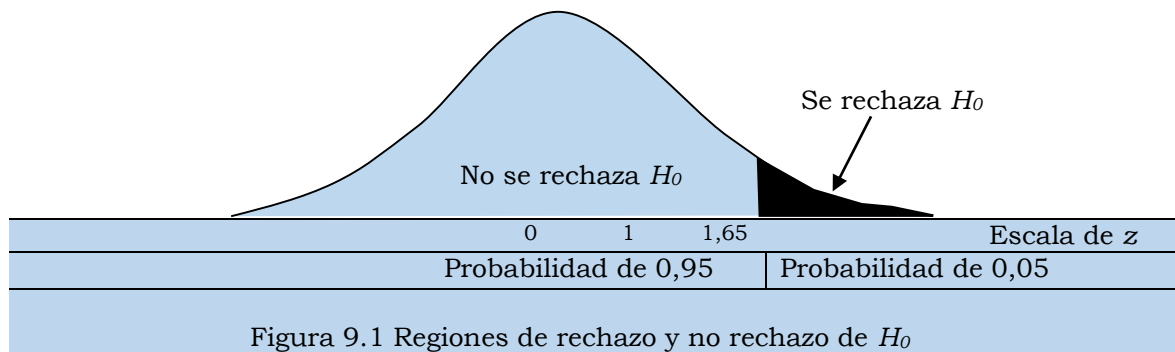
Así, el primer paso de la prueba de hipótesis consiste en determinar las hipótesis nula y alternativa. La hipótesis nula representa la condición actual o declarada, y se escribe, por ejemplo, $H_0: \mu = 100$. La hipótesis alternativa es que la afirmación no es verdad, es decir, $H_1: \mu \neq 100$. Sin importar el problema, la hipótesis nula siempre contendrá el signo $=$, que nunca aparecerá en la hipótesis alternativa.

El segundo paso consiste en seleccionar un nivel de significancia señala el riesgo o probabilidad de rechazar o no la hipótesis nula (cuando es verdadera), y se designa con la letra griega alfa (α). La decisión de seleccionar el nivel de significación depende del contexto de la investigación, aunque tradicionalmente se selecciona el nivel 0,05 o nivel del 5 por ciento.

El tercer paso es calcular el estadístico de prueba, es decir, un valor que se calcula con base en la información de la muestra y que se utiliza para determinar si se rechaza la hipótesis nula. El estadístico de prueba puede ser z , t , F , o chi-cuadrado (χ^2). Por ejemplo, en la prueba de hipótesis para la media de la población (μ) se calcula con la fórmula de z :

$$z = \frac{M - \mu}{\sigma/\sqrt{n}}$$

El cuarto paso consiste en la formulación de una regla de decisión, la cual es una afirmación de las condiciones bajo las cuales se rechaza la hipótesis nula y bajo las que no se rechaza. Por ejemplo, en el caso de una prueba de una cola, se selecciona un nivel de significancia de 0,05 donde el valor de 1,65 separa las regiones de rechazo o no de la hipótesis nula, y en el que ese valor de 1,65 constituye el valor crítico (Figura 9.1), la prueba de hipótesis hacia la izquierda de 1,65 indicaría que no se rechaza la hipótesis nula, y hacia la derecha de 1,65 se rechazaría la hipótesis nula.



El quinto paso consiste en tomar una decisión de rechazar o no la hipótesis nula. Para rechazar la hipótesis nula como en la figura 9.1 tendría que ser mayor a 1,65 y viceversa, si es igual o menor a 1,65 no se rechaza la hipótesis nula. Con la prueba de hipótesis sólo es posible una de las dos decisiones, ya sea aceptar o rechazar la hipótesis nula. Al redactar la decisión, muchos investigadores prefieren expresiones como “No rechazar H_0 ”, “Se deja de rechazar H_0 ”, o “Los resultados de la muestra no permiten rechazar H_0 ”.

Probabilidad y prueba de hipótesis

La estadística ha sido llamada la ciencia de hacer estimaciones apropiadas. A menos que se mida toda la población, la única afirmación que se puede hacer sobre una población con base en una muestra es una estimación apropiada acompañada de una declaración de probabilidad. Las estadísticas nos permiten hacer una declaración y luego citar las probabilidades de que sea correcta. Por ejemplo, si quisiéramos saber la altura de las mujeres de 15 años en una ciudad, podría afirmarse que, en base a una muestra aleatoria de 200 mujeres, la altura promedio de la población se estima en 1,59 metros con un factor de error de $\pm 0,05$ cm y las probabilidades de que esta estimación sea correcta son 95 a 5. En términos estadísticos, esto se llama un nivel de confianza del 95% o una probabilidad de error del 5%. Este factor de error generalmente se representa como $p < 0,05$ (p casi nunca es exacto, por lo que generalmente decimos “la probabilidad de error es inferior al 5%”). Además, $p < 0,05$ también indica que la probabilidad de que la hipótesis nula sea verdadera es inferior al 5%.

Otro enfoque podría ser hipotetizar (estimar o predecir) cuál es la altura promedio de la población y luego usar una muestra y un enunciado de probabilidad para probar la hipótesis. Supongamos una hipótesis de que la altura promedio de la población de mujeres de 15 años en una ciudad es de 1,61 cm. Usando los datos de la muestra de 200 descrita anteriormente, se infiere que hay un 95% de posibilidades de que la altura promedio de la población es 1,59 cm con un factor de error de ± 5 cm (Revisar cómo predecir los parámetros de la población usando inferencias estadísticas). Esto significa que las probabilidades son de 95 a 5 que la altura promedio real de la población de mujeres está entre 1,54 cm y 1,64 cm. En estas condiciones, se acepta como cierta la hipótesis de que la altura es de 1,61 cm porque se encuentra dentro de los límites de los valores estimados a partir de la muestra. Sin embargo, una hipótesis de que la altura media o media es de 1,67 cm sería rechazada porque ese valor no cae dentro de los límites del valor estimado de la población.

La técnica de hipotetizar y luego probar la hipótesis mediante experimentación proporciona respuestas a preguntas que no pueden responderse directamente de ninguna otra manera. Aquí radica el valor real de las estadísticas inferenciales: la capacidad de hacer una declaración basada en la investigación y estimar la probabilidad de que la afirmación sea correcta.

Variables independiente y dependiente

Una variable independiente es aquella que es totalmente libre de variar por sí misma y no covaria con otra variable. Los cambios que ocurren en la primera variable no están necesariamente relacionados con los cambios en la segunda variable. Por ejemplo, si una persona mejora su equilibrio en la gimnasia, eso no tiene relación con la capacidad de esa persona para consumir oxígeno (VO_2) ni tiene ningún efecto sobre ella. Del mismo modo, un VO_2 mejorado a través del acondicionamiento aeróbico no tiene efecto en la capacidad de una persona para realizar una parada de manos. Cada variable (equilibrio y consumo de oxígeno) es independiente y no se ve afectada por los cambios en la otra.

Una variable dependiente no es libre de variar. Depende de los efectos de una o más variables diferentes. Por ejemplo, el peso depende parcialmente de la altura. Durante los años de crecimiento, a medida que las personas crecen, también se vuelven más pesadas. En adultos, generalmente se observa que las personas altas son más pesadas que las personas bajas. Pero la altura no depende del peso. Una persona adulta puede aumentar o perder peso sin afectar la altura.

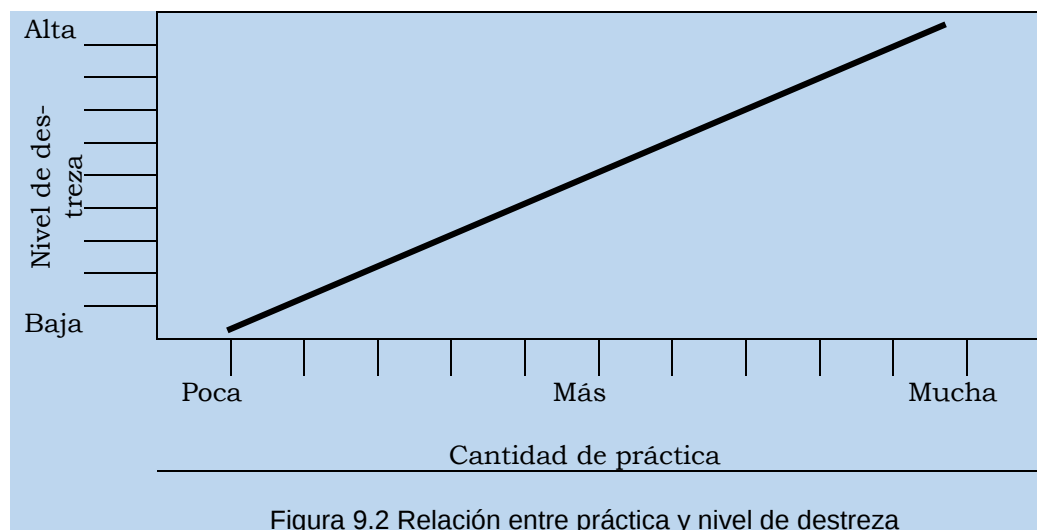
La destreza depende de la práctica. Por lo general, cuanto más se practica, mejor se realiza. Si se construyera una gráfica del tiempo de práctica (en el eje X) y de la destreza (en el eje Y), la gráfica de la relación progresaría de la esquina inferior izquierda a la superior derecha (ver figura 9.2). A medida que aumenta el tiempo de práctica en el eje X, la habilidad también se incrementaría en el eje Y.

En el diseño experimental, es común referirse a la variable independiente, que es controlada por la investigación, como la variable predictora y llamar a la variable dependiente, que se está estudiando, la variable de criterio. Si medimos el efecto de una dieta sobre el aumento o la pérdida de peso, la dieta sería la variable independiente o predictiva, y el peso sería la variable dependiente o criterio. El aumento o pérdida de peso depende de los efectos de la dieta. Manipulamos la variable predictora y medimos sus efectos en la variable criterio.

Estadística inferencial

Gran parte del trabajo realizado por un estadístico implica hacer predicciones sobre un grupo grande (generalmente un grupo de personas, pero podría ser un grupo de animales, rocas, peces, estrellas o casi cualquier objeto) en función de los datos recopilados de una pequeña porción del grupo. Una población es cualquier grupo de personas, lugares o cosas que tienen al menos una característica común. Estas características están especificadas por

la definición de la población. Todas las niñas de séptimo grado en una escuela determinada constituyen una población. Los hombres entre las edades de 31 y 40 años que no participan en el ejercicio regular son otra población. En resumen, una población puede ser cualquier grupo, siempre que se definan los criterios de inclusión en el grupo para que quede claro quién califica como miembro.



Por lo general, la población de interés es bastante grande, tan grande que sería prácticamente imposible o financieramente irracional medir a todos los miembros. Si es imposible o poco práctico medir a todos los miembros de una población, entonces se mide una parte, o fracción, de la población, que se le llama la muestra. Se supone que los sujetos de la muestra representan o tienen las mismas características que la población. Por lo tanto, los datos recopilados en la muestra pueden generalizarse para estimar las características de la población.

Cuanto más grande sea la muestra, representará con mayor exactitud a la población. Si se quiere saber la altura promedio de todas las mujeres de 15 años en una ciudad, es más probable que se obtenga una imagen precisa de la población midiendo 1000 sujetos que midiendo 10 sujetos. El error en la predicción está inversamente relacionado con el tamaño de la muestra. Cuanto mayor es la muestra, menor es el error y cuanto menor es la muestra, mayor es el error.

Una muestra aleatoria es una muestra en la que cada miembro de la población tiene la misma oportunidad de ser seleccionado en la muestra. Cualquier inferencia sobre la población basada en una muestra no aleatoria es de dudoso valor. De hecho, todas las fórmulas y cálculos que se utilizan para hacer inferencias sobre las poblaciones se basan en el supuesto de un muestreo aleatorio. Se pueden cometer errores importantes en la estimación de la población si la muestra no es aleatoria. Por ejemplo, supongamos que necesitamos estimar el porcentaje promedio de grasa corporal de todos los hombres en una

universidad. Una muestra aleatoria requeriría que todos los estudiantes varones tengan las mismas posibilidades de ser seleccionados. Si seleccionamos solo hombres que ingresen al gimnasio entre la 1:00 p.m. hasta las 3:00 p.m., es probable que la muestra represente una mayor porción de atletas que en la población, porque muchos atletas ingresan al gimnasio en este momento para prepararse para la práctica. Los atletas tienden a tener poca grasa corporal, por lo que una muestra de este tipo probablemente subestimaría el porcentaje promedio de grasa corporal de toda la población y se consideraría sesgada.

Posiblemente no haya un lugar en el campus en un momento dado donde todos tengan las mismas posibilidades de ser elegidos. Si la muestra se toma durante el día, los estudiantes nocturnos quedan excluidos. Si se toma el martes, entonces los estudiantes que no están en el campus el martes están excluidos. Si se toma cerca del edificio de ingeniería, es como si se eligiera una cantidad desproporcionado principalmente de ingeniería. Probablemente el único lugar donde se puede tomar una verdadera muestra aleatoria en una universidad es en los archivos del edificio administrativo. Hacer que una computadora seleccione números de identificación aleatorios de una lista de todos los estudiantes podría garantizar que cada estudiante tenga las mismas posibilidades de ser seleccionado en la muestra.

Cuando se desea una selección aleatoria en una población grande e interesan las subcategorías de la misma, se puede tomar una muestra estratificada. Para hacer esto, se selecciona una cierta parte de la muestra por cada subgrupo de la población. En el ejemplo anterior, se debe asegurar que se incluyan estudiantes varones de cada especialidad, algunos de las clases diurnas, algunos de las clases nocturnas, algunos por nivel académico (estudiantes de primer año, estudiantes de segundo año, avanzados, estudiantes de posgrado) y algunos de cada grupo étnico. La proporción de estudiantes por cada subgrupo de la muestra debe ser la misma que la proporción de ese subgrupo en toda la población. Las estadísticas pueden compilarse para toda la muestra o para cualquier subgrupo.

Una muestra no puede representar con precisión una población a menos que se extraiga sin sesgo. El sesgo significa que hay factores que operan en la muestra para que no sea representativa de la población. Estos factores a veces son muy sutiles, y es posible que no haya conciencia de ello. Pero si la muestra es totalmente aleatoria y suficientemente grande, incluso los factores desconocidos que pueden sesgar los resultados se eliminarán o distribuirán dentro de la muestra de la misma manera que se distribuyen en la población.

Parámetros y estadísticas

La única forma de conocer las características exactas de una población es medir a todos sus miembros. Un parámetro es una característica de toda la población. Una estadística es una característica de una muestra que se utiliza para estimar el valor del parámetro de población.

Cualquier estimación de un parámetro de población basada en estadísticas de muestra contiene algún error. La cantidad de error nunca se conoce exactamente, pero se puede estimar en función del tamaño y la variabilidad de la muestra. Por ejemplo, si se

mide a todos los hombres en una universidad y se determina que el porcentaje promedio de grasa corporal en la población es del 21%, este valor se llama parámetro. Si luego tomamos una muestra y encontramos que el porcentaje de grasa en la muestra es 18%, este valor se llama estadística. La diferencia del 3% entre el parámetro y la estadística es el resultado de un error en una muestra.

Antecedentes de las pruebas estadísticas. A menudo, los investigadores están interesados en las características de un grupo grande. Por ejemplo, ¿cuál es el porcentaje medio de grasa corporal para los 40000 hombres que trabajan para una industria en particular, o cuál de las tres técnicas de entrenamiento físico es la mejor para los 40000 trabajadores? Para responder a la primera pregunta, un investigador podría medir a todos los 40000 hombres por porcentaje de grasa corporal y luego calcular la media. Esto llevaría mucho tiempo y sería costoso, y probablemente imposible.

En lugar de examinar a todo el gran grupo de su interés, llamado población, los investigadores prueban una parte de la población, llamada muestra. Se calcula la media de una muestra y luego el investigador asume que la muestra y la media de la población son iguales. En otras palabras, si la media de la muestra es 17, el investigador infiere que la media de la población es 17. En el ejemplo anterior sobre cuáles técnicas de entrenamiento físico son las mejores, los investigadores obtendrían tres muestras y utilizarían una técnica diferente de entrenamiento físico en cada muestra. Después de que la técnica de entrenamiento se haya utilizado durante un tiempo considerable (por ejemplo, 6 semanas o más), el investigador evaluará a todos los participantes y determinará si el rendimiento promedio de los grupos fue igual. Cualquiera sea el resultado, el investigador inferiría el hallazgo de la muestra a la población.

Para hacer esta inferencia, la muestra debe ser representativa de la población y seleccionada al azar. Para que una muestra sea representativa, debe tomarse de una población identificada y debe ser lo suficientemente grande (es decir, 30 o más) para que estén representadas todas las diferentes habilidades de la población. La selección aleatoria de los participantes garantiza que todos los miembros de la población tengan las mismas posibilidades de ser seleccionados. En un entorno de investigación, esto se hace con una tabla especial llamada tabla de números aleatorios, pero en un entorno práctico se puede lograr colocando el nombre de cada persona en una hoja de papel en un contenedor, extrayendo 50 hojas de papel y designando a esas personas como muestra.

Una investigación comienza con una declaración sobre lo que se piensa que puede ser el caso a nivel de población. Esta afirmación se llama hipótesis nula. Por ejemplo, "Dieciocho por ciento es el porcentaje medio de grasa corporal para hombres adultos", o "Las tres técnicas de entrenamiento físico son igualmente efectivas, por lo que las medias de los tres grupos son iguales". Por lo general, la media o medias obtenidas de la (s) muestra (s) no serán iguales a los valores en la hipótesis nula. Entonces, la pregunta es si la diferencia entre lo que se hipotetizó a nivel de población y se encontró a nivel de muestra es lo suficientemente grande como para sugerir que la hipótesis nula es falsa o si la diferencia se debe a un error de muestreo.

El error de muestreo se debe a que las muestras no son 100% representativas de la población. Por ejemplo, si se supuso que el porcentaje medio de grasa corporal era 18 pero era 17 en la muestra, ¿la hipótesis es falsa o esta diferencia se debe a un error de muestreo? Si las tres muestras no tienen la misma media después de recibir su respectiva técnica de entrenamiento, ¿esto indica que la hipótesis es falsa? Para poder tomar una decisión objetiva, se realiza una prueba estadística. Esta prueba estadística es muy parecida a las declaraciones de probabilidad ya mencionadas. Basado en la prueba estadística, se puede determinar la probabilidad de que ocurra un hallazgo en la muestra, si la hipótesis nula es verdadera. Si la probabilidad es bastante pequeña, la hipótesis nula se rechaza y la diferencia se considera real. De lo contrario, se acepta y se considera que la diferencia se debe al error de la muestra. Por lo tanto, un investigador debe seleccionar algún nivel de probabilidad que justifique el rechazo de la hipótesis nula.

Este nivel de probabilidad se llama nivel alfa, y generalmente se especifica como 0.05 o 0.01. Lo que hacen la mayoría de los investigadores es buscar en la tabla estadística adecuada el valor de la prueba estadística necesaria para rechazar la hipótesis nula en el nivel alfa seleccionado. Si el valor de la prueba estadística es mayor o igual que el valor de la tabla, el investigador rechaza la hipótesis nula. Muchos programas de computadora de escritorio proporcionan un nivel de probabilidad (p) para el valor de la prueba estadística. En este caso, el investigador rechaza la hipótesis nula si el valor p es menor o igual que el nivel alfa. Todo esto se aclarará con los ejemplos proporcionados, cada uno con un diseño de investigación y prueba estadística diferentes.

Prueba t

Supongamos que una entrenadora de gimnasia quería saber si las gimnastas universitarias tenían mayor flexibilidad de cadera y espalda baja que los gimnastas. Para responder a esta pregunta, midió la flexibilidad de los atletas de los equipos masculino y femenino, usando un cajón de flexibilidad. Al final de la sesión, la distancia promedio alcanzada más allá de la punta de los pies para los hombres fue de 12 centímetros, mientras que las mujeres promediaron 15 centímetros. En esta universidad, las mujeres eran claramente más flexibles que los hombres. No fue necesario más análisis estadístico.

Luego, la entrenadora quiso comparar a todos los gimnastas universitarios en todo el estado. Esto requería que ella midiera a todos los atletas en el estado, lo que no era factible. En cambio, tomó muestras aleatorias de gimnastas masculinos y femeninos de varias universidades de todo el estado. Luego realizó una prueba t para determinar si las medias en las dos muestras reflejaban con precisión las medias de las poblaciones de las que provenían. La predicción de parámetros de población a partir de estadísticas de muestra se denomina estadística inferencial.

Recuérdese que una media muestral puede usarse como una estimación de una media poblacional, y también que se puede determinar la probabilidad de que la media de la población se encuentre dentro de ciertos límites numéricos utilizando la media de la muestra como predictor. Esta misma técnica se puede utilizar a la inversa, para determinar

si es probable que una muestra dada haya sido seleccionada al azar a partir de una población específica.

Si se sabe o se supone que la media de la población tiene cierto valor, y si la media de la muestra no está lo suficientemente cerca de la media de la población como para caer dentro de los límites establecidos por un nivel de confianza seleccionado, entonces una de las siguientes conclusiones debe ser cierta: (a) La muestra no se extrajo al azar de la población, o (b) la muestra se extrajo de la población, pero se ha modificado para que ya no sea representativa de la población de la que se extrajo originalmente.

Usando una lógica similar, se pueden sacar conclusiones sobre dos conjuntos de datos. Si extraemos dos muestras de la misma población, y las medias de estas muestras difieren en cantidades mayores de las esperadas en función de las distribuciones normales, una de las siguientes conclusiones debe ser cierta: (a) Una o ambas muestras de la población no fueron extraídas al azar, o (b) algún factor ha afectado a una o ambas muestras, haciendo que se desvíen de la población de la que se extrajeron originalmente.

Pruebas t. Cuando se extrae una muestra de una población con una media (μ) y una desviación estándar (σ) conocida o estimada, la probabilidad (o probabilidades) de que la media de una muestra extraída al azar (M) se encuentre dentro de ciertos límites de μ puede ser determinado. Para determinar la probabilidad de que una muestra dada provenga de una determinada población, el valor del error estándar de la media debe calcularse mediante una de las siguientes fórmulas. Si se conoce la desviación estándar de la población (σ), la fórmula utilizada es

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

donde σ_M es el símbolo del actual error estándar de la media para una población con una media y una desviación estándar conocidas. Si no se conoce σ , la fórmula utilizada es

$$SE_M = \frac{SD}{\sqrt{N}}$$

donde SE_M es el error estándar de la media estimada de una muestra.

Usando σ_M o SE_M , podemos determinar las probabilidades de que una muestra sea representativa de la población de la que se extrajo, haciendo una prueba Z si se conoce la población,

$$Z = \frac{M - \mu}{\sigma_M}$$

o una prueba t,

$$t = \frac{M - \mu}{SE_M}$$

si se estima la población.

Evaluando Z. Cuando se conoce σ_M , Z indica la importancia de la diferencia entre M y μ . En estas condiciones, podemos determinar la importancia de Z comparándolo con relaciones críticas de 1.65 (en $p = 0.10$), 1.96 (en $p = 0.05$) y 2.58 (en $p = 0.01$). Si Z excede uno de estos niveles podemos concluir, en el nivel de confianza (LOC) dado, que la muestra no se extrajo al azar de la población, o que se ha modificado de alguna manera para que ya no represente a la población de la cual fue extraída.

Evaluación de t a partir de una muestra simple. La aproximación de σ_M por SEM no es precisa en muestras pequeñas (N menor que 60). Student (un seudónimo) desarrolló una serie de aproximaciones de la curva normal para explicar el sesgo en la estimación de σ_M llamada distribución t de Student. La tabla 9.1 enumera los valores para la distribución t de Student. Si no hubiera error en la estimación, la tabla del área bajo la curva normal podría usarse en todos los casos. Pero cuando las muestras se usan para estimar los parámetros de la población, especialmente cuando las muestras son pequeñas, la distribución t (tabla 9.1) debe usarse para evaluar el estadístico t . Los valores de la tabla 9.1 en los niveles de p dados se denominan proporciones críticas. Representan la relación t que se debe alcanzar para rechazar el azar.

La prueba t para una muestra produce la relación de la diferencia media real entre la muestra y la población con respecto a la diferencia media esperada, (esa cantidad de diferencia entre M y μ que se puede esperar que ocurra solo por casualidad). La diferencia media esperada se estima mediante la ecuación

$$SE_M = \frac{SD}{\sqrt{N}}$$

y se denomina error estándar de la media. Para interpretar t para una sola muestra, primero debemos encontrar los grados de libertad, que pueden calcularse mediante la fórmula $gl = N - 1$. La relación se compara con los valores para una prueba de dos colas (las pruebas de una y dos colas se explicarán más adelante en este capítulo) de la distribución t en la tabla 9.1 para el gl apropiado.

Cuando t excede el valor de la tabla 9.1 para un nivel p dado, podemos concluir que M no se extrajo de μ . Cuando t es menor que la razón crítica en la tabla 9.1, se acepta la hipótesis nula (H_0); no hay diferencia fiable entre M y μ . Cuando t excede la relación crítica, se rechaza H_0 y se acepta H_1 ; algún factor diferente al azar está operando en la media muestral. Observe que en la tabla 9.1 cuando los grados de libertad son grandes ($gl > 120$), los valores de una prueba t de dos colas en un valor p dado son los mismos que los valores de la tabla del área bajo la curva normal para una prueba Z (1.65, 1.96 y 2.58).

Esta técnica es útil para determinar si los cambios introducidos por un experimento tienen un efecto en los sujetos. Si se conocen o estiman los parámetros de la población y luego se extrae una muestra aleatoria y se le trata de una manera que se espera que altere su valor medio, se puede determinar las probabilidades de que el tratamiento tenga un efecto mediante el uso de una prueba t . Si t excede las proporciones críticas en la tabla 9.1, se puede concluir que el tratamiento fue efectivo porque hay altas probabilidades de que la muestra ya no sea representativa de la población de la que se extrajo. El tratamiento ha provocado que la muestra cambie para que no coincida con las características de la población de origen.

Suposiciones para la prueba t . Se deben cumplir varios supuestos para que la prueba t se aplique correctamente. Si no se cumplen estos supuestos, los resultados pueden no ser válidos. Cuando el investigador sabe que uno o más de estos criterios no se cumplen, se debe seleccionar un nivel más conservador (es decir, $p = 0.01$ en lugar de $p = 0.05$) para evitar errores. Esto permite tener confianza en las conclusiones y ayuda a compensar el hecho de que no se cumplieron todos los supuestos. La prueba t es bastante robusta; produce resultados razonablemente confiables, incluso si los supuestos no se cumplen por completo. La prueba t se basa en los siguientes supuestos:

- La población de la que se extraen las muestras se distribuye normalmente (consulte los métodos para determinar la cantidad de asimetría y curtosis en un conjunto de datos).
- La muestra o muestras se seleccionan aleatoriamente de la población. Si las muestras no se seleccionan al azar, no se puede hacer una generalización de la muestra a la población.
- Cuando se extraen dos muestras, las muestras tienen una varianza aproximadamente igual. La varianza de un grupo no debe ser más del doble de la varianza del otro. Esto se llama homogeneidad de varianza.
- Los datos deben ser paramétricos, es decir, basados en una escala de medición de intervalo o relación.

Un ejemplo en educación física. La coordinación de un departamento de educación física estaba preocupada por la aparente falta de destrezas que los estudiantes desarrollaban en una unidad de voleibol de 5 semanas. Los estudiantes se asignaron a los instructores de manera aleatoria, alfabéticamente, según el apellido. Por lo general, la clase solo juega durante todo el período y recibe poca o ninguna práctica en destrezas específicas. Una prueba estandarizada sobre la destreza en el servicio de voleibol (50 puntos, máximo) se les ha aplicado a todos los estudiantes durante varios años y se han guardado los datos. La media para más de 1000 estudiantes (la población) en esta prueba es de 31 puntos (μ) con una desviación estándar de 7,5 (σ).

Un docente decidió probar un enfoque diferente. En una clase (las muestras), la mitad del período se dedicó a la enseñanza de las destrezas, especialmente a las destrezas

del servicio, y los estudiantes practicaron bajo la dirección del docente durante 20 minutos. Hubo juegos solo al final del período, y la última semana del bloque de voleibol se organizaron equipos y se celebró un torneo.

A los estudiantes de esta clase ($N = 30$) también se les aplicó la prueba estandarizada de servicio. Su puntaje promedio (M) fue de 35 puntos de los 50 posibles con una desviación estándar (SD) de 8,3. ¿Fue efectivo el maestro en mejorar las destrezas de servicio? En otras palabras, si asumimos que la clase de 30 estudiantes es una muestra aleatoria de la población de más de 1000 estudiantes, ¿es probable que el puntaje promedio para esa clase ($M = 35$) sea representativo de la media de la población ($\mu = 31$)? Para probar la hipótesis de que la clase de muestra representa a la población, calculamos el error estándar de la media para la población:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{7,5}{\sqrt{30}} = 1,37$$

Luego se realiza una prueba Z (porque μ y σ son conocidas) para determinar las probabilidades de que la media de una muestra extraída al azar de la población difiera de la media de la población en hasta 4 puntos:

$$z = \frac{35 - 31}{1,37} = 2,92$$

¿Cuál es la probabilidad de que se obtenga una puntuación z de 2.92, si la muestra se extrajera de la población y no se tratara? Debido a que z es mayor que 2,58, la probabilidad de que M no provenga de μ es mayor que 99 a 1, $p = <0,01$.

Deben considerarse dos posibilidades:

- La clase no era una muestra aleatoria de la población y, por lo tanto, no representa a la población. Quizás por suerte, o por diseño, estos 30 estudiantes fueron mejores al comienzo del bloque de 5 semanas que los estudiantes típicos asignados a las anteriores clases.
- La muestra fue aleatoria al comienzo del bloque de 5 semanas, pero el tratamiento (instrucción y práctica) ha cambiado a los estudiantes de tal manera que ya no son representativos de la población. En otras palabras, los estudiantes han cambiado de modo que ahora representan otra población, una que tiene instrucción y práctica en lugar de juego libre.

Si se puede demostrar la asignación aleatoria a los instructores, de modo que se garantice que la clase fue representativa al comienzo del bloque de 5 semanas, solo queda una conclusión. Una diferencia de 4 puntos entre la media de la muestra y la media de la población ocurriría menos de 1 vez en 100 por azar. En otras palabras, las probabilidades de que la instrucción sea efectiva son mejores que 99 a 1 ($LOC = 99\%$). Rechazamos H_0 y concluimos que la instrucción fue efectiva en $p = <0.01$.

En el ejemplo anterior, se usó σ porque se conocía su valor. Pero en la mayoría de las investigaciones, μ y σ no se conocen. Si no se conoce σ , se estima SE_M , el error estándar de la media, usando la ecuación

$$SE_M = \frac{SD}{\sqrt{N}}$$

y la desviación estándar de la muestra. Luego, se usa la prueba t en lugar de Z para determinar la importancia de la diferencia entre la muestra y la media poblacional estimada. Primero, se calcula SE_M :

$$SE_M = \frac{8,3}{\sqrt{30}} = 1,52$$

Luego usamos SE_M para determinar t . Si $M = 35$, entonces

$$t = \frac{M - \mu}{SE_M} = \frac{35 - 31}{1,52} = 2,63$$

Observe que la respuesta para t es ligeramente más pequeña que la puntuación Z (2,92). Esto demostró que el poder de detectar diferencias entre las muestras y la población es mayor cuando se conoce a la población que cuando se estima. Para encontrar el nivel de confianza, se compara la razón t (2.63) con las razones críticas de una prueba de dos colas de la distribución t en la tabla 9.1 para $gl = 30 - 1 = 29$. La relación crítica en $gl = 29$ en la tabla 9.1 para $p = 0.05$ es 2.045 y para $p = 0.01$ es 2.756. Nuestro valor t cae entre estos valores, por lo que aceptamos el valor menos significativo, $p = <0.05$. Tenga en cuenta que se tiene un nivel de confianza más bajo (95%) para rechazar H_0 cuando se estima la población que cuando se conocen los parámetros de la población (99%). Mientras que la tabla 9.1 está limitada a tres niveles (0,10; 0,05 y 0,01), las computadoras generarán valores p exactos. Es preferible usar una computadora para analizar datos reales.

Prueba t para un grupo

Paso 1. La hipótesis nula del investigador era que el porcentaje medio de grasa corporal para la población era 18, por lo que la media poblacional hipotética (μ) es 18 ($H_0: \mu = 18$). Como alternativas a esta hipótesis, el investigador pensó que es posible que la media sea menor que 18 ($H_1: \mu < 18$) o mayor que 18 ($H_2: \mu > 18$).

Paso 2. Se seleccionó un nivel alfa de 0.05.

Paso 3. Sabiendo que una prueba t sería la prueba estadística y el tamaño de la muestra sería 41, el investigador fue a las tablas t y descubrió que para rechazar la hipótesis nula al nivel alfa 0.05, el valor de la prueba t debería ser mayor o igual a 2.021, o menor o igual a -2.021. Este paso no es necesario si se utiliza un programa de computadora, ya que calcula los valores p automáticamente.

Paso 4. El investigador seleccionó al azar a 41 hombres de la población como muestra, midió su porcentaje de grasa corporal, calculó la media y la desviación estándar para las mediciones de grasa corporal, y calculó la prueba t .

$$M = 16,50 \quad s = 2,50 \quad n = 41$$

$$t = \frac{M - \mu}{s/\sqrt{n}} = \frac{16,50 - 18}{2,50/\sqrt{41}} = \frac{-1,50}{2,50/6,40} = \frac{-1,50}{0,39} = -3,85$$

Paso 5. Dado que el valor calculado de t (-3,85) en el paso 4 es menor que el valor de la tabla de t (-2.021) en el paso 3, la investigación concluye que la diferencia entre la hipótesis nula valorada (18) en el paso 1 y el valor de muestra (16,50) en el paso 4 es real y no se debe a un error de muestra. Por lo tanto, el investigador rechaza la hipótesis nula y acepta la hipótesis alternativa más probable, que en este caso es H_1 , "la media de la población es inferior a 18", ya que la media de la muestra fue inferior a 18.

Ahora algunos puntos de información sobre lo que se presentó en estos cinco pasos. En el paso 3, cuando el investigador fue a las tablas t para identificar el valor t necesario para rechazar la hipótesis nula, tuvo que alinear un valor llamado grados de libertad y el nivel alfa para encontrar el valor t . Los grados de libertad para esta prueba t ($n-1$) donde n es el tamaño de la muestra. Como había dos hipótesis alternativas, el investigador utilizó el nivel alfa en una prueba de dos colas. Para una prueba de dos colas, este valor siempre se interpreta como más y menos. Si hubiera habido solo una hipótesis alternativa, se habría utilizado el nivel alfa bajo una prueba de una cola y el valor de la tabla se habría interpretado como más y menos, dependiendo de si la hipótesis alternativa era mayor que (valor de la tabla más) o menor que (valor de la tabla menos).

El valor de la tabla a veces se denomina valor crítico o región crítica. En el paso 5, si el valor de la prueba estadística en el paso 4 es igual o mayor que el valor de la tabla en el paso 3, la diferencia y la prueba estadística se denominan significativas (la diferencia se considera real) y la hipótesis nula se rechaza. Si el valor de la prueba estadística en el paso 4 no es igual o mayor que el valor de la tabla en el paso 3, la diferencia y la prueba estadística se denominan no significativas (se considera que la diferencia se debe a un error de muestreo) y se acepta la hipótesis nula.

Esta prueba t es útil en cualquier situación de medición en la que se compara el rendimiento medio de un grupo con una media esperada o un estándar o norma expresada como una media. ¿La media del grupo realmente difiere de la media esperada?

Comparar dos grupos independientes y dos grupos dependientes. En este apartado se plantea una generalidad sobre la comparación de muestras independientes, que puede profundizarse posteriormente de acuerdo al interés en un estudio de este tipo y el requerimiento de la investigación en sí. Los mismos conceptos utilizados para comparar una media muestral con la media poblacional pueden aplicarse para comparar dos muestras extraídas de la misma población. Si la relación excede la relación crítica de la tabla 9.1, la hipótesis nula H_0 se rechaza. H_0 afirma que ambas medias representan (o fueron extraídas al azar) de la misma población. Si la relación t no excede la relación crítica, se acepta H_0 .

En este diseño, dos muestras independientes se extraen aleatoriamente de una población. Por muestras independientes, queremos decir que los sujetos en una muestra son

personas diferentes y no están relacionados, o correlacionados, de ninguna manera con los sujetos en la otra muestra. Una muestra, el grupo experimental, se trata, y la otra, el grupo de control, no. Se realiza una prueba t para comparar la diferencia real entre las medias de las dos muestras con la diferencia esperada si solo estuviera funcionando el azar. Esta relación se utiliza para determinar si los dos grupos representan a la población de la que se extrajeron. También pueden ser dos muestras extraídas de la misma población y a cada muestra se le aplica un tratamiento diferente, o muestras que se extraen de dos poblaciones diferentes.

En la prueba t para dos grupos dependientes, un ejemplo es un grupo de personas que son medidas en dos ocasiones diferentes, por lo general al comienzo y al final de un período de preparación, entrenamiento o clases. En este caso hay algún grado de correlación o dependencia entre las dos columnas de puntajes. Esta prueba t es útil en cualquier situación de medición cuando se requiera determinar si la media del desempeño de un grupo cambia a través del tiempo, o como resultado de un programa instruccional o de entrenamiento. Si la razón t de cálculo es mayor que la razón crítica de la distribución t en la tabla 9.1, se rechaza la posibilidad y se supone una causa probable. Cuando se rechaza H_0 para un experimento realizado bajo condiciones cuidadosamente controladas, el tratamiento se identifica como la causa probable.

A diferencia de la prueba t para un grupo la cual es fácil de calcular a mano, en las pruebas t para dos grupos independientes o dos grupos dependientes no es tan fácil hacerlo a mano por lo que, afortunadamente, hay paquetes estadísticos para computadora que lo calculan rápidamente como el SPSS.

Análisis de varianza

Un equipo de investigadores (científicos del ejercicio y nutricionistas), interesados en los efectos relativos de la dieta y el ejercicio sobre la composición corporal, realizó un estudio con cuatro grupos seleccionados al azar: 1) Un grupo de control que no recibió tratamiento, 2) un grupo con dieta sin un programa de ejercicio, 3) un grupo con un programa de ejercicio, pero sin dieta, y 4) un grupo combinado que siguió tanto una dieta como un programa de ejercicio. El porcentaje de grasa corporal de los sujetos en los cuatro grupos se midió al final del programa de 6 semanas.

El análisis de varianza (ANOVA) es una técnica estadística paramétrica utilizada para determinar si existen diferencias significativas entre las medias para tres o más conjuntos de datos de muestra. En una prueba t , las diferencias entre las medias de dos grupos se comparan con la diferencia esperada solo por azar. El análisis de varianza compara la variabilidad entre tres o más medias grupales (la variabilidad entre grupos) con la variabilidad de los puntajes dentro de los grupos (la variabilidad dentro del grupo). Esto produce un valor de relación llamado F (F = varianza promedio entre grupos dividido por la varianza promedio dentro de los grupos).

Si la variabilidad entre grupos excede la variabilidad dentro del grupo en más de lo esperado por azar, se puede concluir que al menos una de las medias del grupo difiere

significativamente de la media de otro grupo. La hipótesis nula (H_0) para una prueba F se designa como $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$.

Tabla de distribución t de Student.

gl	Prueba de dos colas			gl	Prueba de una cola		
	0,10	0,05	0,01		0,10	0,05	0,01
1	6.314	12.706	63.657	1	3.078	6.314	31.821
2	2.920	4.303	9.925	2	1.886	2.920	6.965
3	2.353	3.182	5.841	3	1.638	2.353	4.541
4	2.312	2.776	4.604	4	1.533	2.312	3.747
5	2.015	2.571	4.032	5	1.476	2.015	3.365
6	1.943	2.447	3.707	6	1.440	1.943	3.143
7	1.895	2.365	3.499	7	1.415	1.895	2.998
8	1.860	2.306	3.355	8	1.397	1.860	2.896
9	1.833	2.262	3.250	9	1.383	1.833	2.821
10	1.812	2.228	3.169	10	1.372	1.812	2.764
11	1.796	2.201	3.106	11	1.363	1.796	2.718
12	1.782	2.179	3.055	12	1.356	1.782	2.681
13	1.771	2.160	3.012	13	1.350	1.771	2.650
14	1.761	2.145	2.977	14	1.345	1.761	2.624
15	1.753	2.131	2.947	15	1.341	1.753	2.602
16	1.746	2.120	2.921	16	1.337	1.746	2.583
17	1.740	2.110	2.898	17	1.333	1.740	2.567
18	1.734	2.101	2.878	18	1.330	1.734	2.552
19	1.729	2.093	2.861	19	1.328	1.729	2.539
20	1.725	2.086	2.845	20	1.325	1.725	2.528
21	1.721	2.080	2.831	21	1.323	1.721	2.518
22	1.717	2.074	2.819	22	1.321	1.717	2.508
23	1.714	2.069	2.807	23	1.319	1.714	2.500
24	1.711	2.064	2.797	24	1.318	1.711	2.492
25	1.708	2.060	2.787	25	1.316	1.708	2.485
26	1.706	2.056	2.779	26	1.315	1.706	2.479
27	1.703	2.052	2.771	27	1.314	1.703	2.473
28	1.701	2.048	2.763	28	1.313	1.701	2.467
29	1.699	2.045	2.756	29	1.311	1.699	2.462
30	1.697	2.042	2.750	30	1.310	1.697	2.457
40	1.684	2.021	2.704	40	1.303	1.684	2.423
60	1.671	2.000	2.660	60	1.296	1.671	2.390
120	1.658	1.980	2.617	120	1.289	1.658	2.358
∞	1.645	1.960	2.576	∞	1.282	1.645	2.326

Tabla 9.1 Valores de la distribución t de Student

La hipótesis nula supone que las medias de dos o más poblaciones son iguales o que una sola población es el origen de varias muestras no procesadas y extraídas de ella.

Por lo tanto, las medias no tratadas de muestras extraídas al azar de la misma población no deberían diferir más que por azar. Cuando al menos una media muestral es significativamente diferente de cualquier otra, F es significativa y se rechaza la hipótesis nula.

Al igual que la prueba t , los conceptos teóricos de análisis de varianza se basan en muestras aleatorias extraídas de una población y las características de la curva normal. Si se extrae al azar un gran número de muestras de una población, la varianza entre las puntuaciones de todos los sujetos en todos los grupos es la mejor estimación de la varianza de la población. Cuando se conoce la varianza entre todos los puntajes en un conjunto de datos, se puede usar para determinar la probabilidad de que un puntaje desviado no se extraiga al azar de la misma población. Este argumento se puede ampliar para inferir que, si las puntuaciones extraídas al azar se dividen aleatoriamente en subgrupos, se puede esperar que la varianza entre las medias de los subgrupos sea de la misma magnitud relativa que la varianza entre todas las puntuaciones individuales que comprenden los diversos grupos. Con datos no procesados, cuando solo funcionan factores aleatorios entre las medias del grupo y dentro de los puntajes de los grupos, las variaciones entre grupos y dentro del grupo deben ser aproximadamente iguales. Se espera que la proporción F , la proporción de la varianza promedio entre grupos dividida por la varianza promedio dentro del grupo, sea de aproximadamente 1.00. Cuando el valor de la relación F excede 1.00 en más de lo que se esperaría por azar, la variación entre la media se considera significativa (es decir, la diferencia fue causada por un factor diferente al azar), y se rechaza H_0 .

La relación F es análoga a la relación t en que ambas comparan las diferencias medias reales u observadas con las diferencias esperadas por azar. Cuando esta proporción excede los límites del azar a un nivel de confianza dado, se rechaza el azar como causa de las diferencias. En la prueba F , las diferencias reales son las variaciones entre las medias del grupo, y las diferencias esperadas son las variaciones dentro de los puntajes individuales que componen los grupos. La prueba t es en realidad un caso especial de análisis de varianza con dos grupos. Debido a que t usa la desviación estándar y ANOVA usa la varianza (V) para estimar las diferencias de medias, un $SD^2 = V$, cuando solo hay dos grupos en ANOVA, $t^2 = F$.

El análisis de varianza es una de las técnicas estadísticas más utilizadas en investigación. Pero a menudo surgen preguntas: ¿Por qué se necesita esta nueva técnica cuando se puede usar una prueba t entre cada uno de los grupos? Por ejemplo, en un estudio de cuatro grupos, ¿por qué no realizar seis pruebas entre los grupos A y B, A y C, A y D, B y C, B y D, y C y D? Hay tres razones por las que las pruebas t múltiples no son apropiadas. - Hay una mayor probabilidad de cometer un error tipo I (rechazando la hipótesis nula cuando es realmente cierta) al realizar múltiples pruebas t en muestras tomadas de la misma población. Cuando se realiza una sola prueba, los resultados se comparan con la probabilidad del azar. Cuando se realizan múltiples pruebas t de muestras extraídas al azar de la misma población, aumentan las probabilidades de encontrar la única conclusión desviada que se espera por azar.

- La prueba t no utiliza toda la información disponible sobre la población de la que se extrajeron las muestras. La prueba t se basa en el supuesto de que solo dos grupos han sido seleccionados al azar de la población. En la prueba t , la estimación del error estándar de

la diferencia entre medias se basa solo en datos de dos grupos. Cuando se han seleccionado tres o más grupos, la información sobre la población de tres o más muestras está disponible y debe usarse en el análisis, pero t considera solo dos muestras a la vez. El análisis de varianza emplea toda la información disponible de todos los grupos simultáneamente.

- Las pruebas t múltiples requieren más tiempo y esfuerzo que un ANOVA simple. Es más fácil, especialmente con una computadora, realizar una prueba F que realizar múltiples pruebas t .

Debido a estas razones, el análisis de varianza se emplea en lugar de múltiples pruebas t cuando están involucrados tres o más grupos de datos. El análisis de varianza puede determinar si existe una diferencia significativa entre cualquiera de los grupos representados en el experimento, pero ANOVA no identifica el grupo o grupos que difieren. Un valor F significativo solo indica que existe al menos una diferencia significativa en algún lugar entre las muchas combinaciones posibles de grupos. Cuando se encuentra una F significativa, se deben realizar pruebas adicionales post hoc (después del hecho) para identificar el grupo o grupos que difieren. Si F no es significativo, no se necesitan más análisis porque sabemos que no hay diferencias significativas entre ninguno de los grupos.

Supuestos de ANOVA. La prueba F se basa en los siguientes supuestos:

- La (s) población (es) de las cuales se extraen las muestras está(n) normalmente distribuidas. La violación de este supuesto tiene poco efecto sobre el valor F entre las muestras. La prueba F produce resultados válidos incluso cuando la población no está normalmente distribuida. Por esta razón, F se considera robusto.

- La variabilidad de las muestras en el experimento es igual o casi igual (homogeneidad de la varianza). Al igual que con el supuesto de normalidad, la violación de este supuesto no cambia radicalmente el valor de F . Sin embargo, como regla general, la varianza del grupo más grande no debe ser más de dos veces la varianza del grupo más pequeño.

- Los puntajes en todos los grupos son independientes; es decir, los puntajes en cada grupo no dependen, no están correlacionados o no se toman de los mismos sujetos que los puntajes en cualquier otro grupo. Las muestras se seleccionaron aleatoriamente de la población y se asignaron aleatoriamente a las condiciones. Si existe una relación conocida entre las puntuaciones de los sujetos en los diversos grupos, se utilizan medidas repetidas de análisis de varianza.

- Los datos se basan en una escala paramétrica, ya sea intervalo o relación.

La prueba F , como la prueba t , se considera robusta. Proporciona respuestas confiables incluso cuando hay violaciones de los supuestos. Las violaciones son más críticas cuando los tamaños de muestra son pequeños o N no son iguales. Si se cometen violaciones que no pueden controlarse y que el investigador cree que pueden aumentar la posibilidad de un error tipo I, se debe usar un valor p más conservador para compensar las violaciones (es decir, use $p = 0.01$ en lugar de $p = 0.05$).

Fuentes de varianza. El cálculo de F es simple en teoría y no involucra matemáticas difíciles. Cuando se comparan varios grupos de datos, cada grupo tiene una media (la media del grupo) y todo el conjunto de datos, todos los grupos combinados, tiene una media (la gran media). La gran media se calcula sumando las puntuaciones de todos los

grupos y dividiendo por el número total de sujetos en todos los grupos combinados (N). La media de cada uno de los grupos puede diferir de la gran media. La varianza, o desviación, de las medias grupales de la gran media se llama varianza entre grupos.

Además de la varianza entre las medias, cada puntaje individual se desvía de la media de su grupo en una cierta cantidad. Esta fuente de variación se llama variación dentro del grupo. Estas dos fuentes de varianza, entre grupos y dentro del grupo, son los componentes básicos utilizados para calcular el análisis de varianza, F .

Una tercera fuente de variación, la variación total, puede calcularse determinando la desviación de cada puntaje en cada grupo de la gran media. La varianza total es igual a la suma de la varianza entre grupos y dentro de la varianza:

$$\text{Varianza}_{total} = \text{Varianza}_{entre} + \text{Varianza}_{dentro}$$

La varianza total y la varianza entre grupos son relativamente fáciles de calcular. Pero calcular la varianza dentro del grupo, aunque no es matemáticamente difícil, es tedioso. Como la varianza total es la suma de la varianza entre y dentro del grupo, la varianza dentro del grupo puede determinarse restando la varianza entre grupos de la varianza total. La varianza total no es esencial para determinar F , pero es útil para calcular la varianza dentro del grupo.

Suma de cuadrados y cuadrado medio. La suma de cuadrados (SS) es la suma de los cuadrados de las desviaciones de cada puntaje de una media (M representa la media):

$$SS = \Sigma(X - M)^2$$

Se calcula restando cada puntaje de la media, cuadrando los puntajes de desviación y sumándolos. La suma de los cuadrados dentro de cualquier grupo se puede calcular a partir de los puntajes individuales y la media del grupo:

$$SS_w = \Sigma(X - M_{grupo})^2$$

La suma de los cuadrados entre los grupos se puede calcular restando la gran media de cada media del grupo:

$$SS_B = \Sigma(M_{grupo} - M_{grande})^2$$

La suma de cuadrados para el total se puede calcular restando cada puntaje individual y la gran media:

$$SS_T = \Sigma(X - M_{grande})^2$$

La suma total de cuadrados siempre es igual a la suma de cuadrados entre grupos más la suma de cuadrados dentro del grupo:

$$SS_T = SS_B + SS_w$$

En ANOVA, el tamaño de la suma de cuadrados entre grupos se compara con el tamaño de la suma de cuadrados dentro de los grupos. El tamaño de la suma de cuadrados depende de (a) el número de puntajes sumados y (b) el tamaño de las desviaciones de cuadrados de la media o la varianza. Para tener en cuenta las diferencias en el número de puntajes que conforman la suma de cuadrados entre grupos (cantidad de grupos) y dentro del grupo (cantidad de puntajes individuales), el cuadrado medio (MS) se calcula dividiendo cada SS por los grados de libertad apropiados ($MS = SS / df$). Este proceso hace que la variabilidad media o promedio para SS_B y SS_W sea comparable.

Los grados de libertad dentro del grupo se determinan restando el número de grupos (k) del número total de sujetos en todos los grupos (N):

$$gl_W = N - k$$

Los grados de libertad entre están determinados por el número de grupos (k) menos 1:

$$gl_B = k - 1$$

Para determinar el cuadrado medio dentro, divide SS_W por grados de libertad dentro del grupo:

$$MS_W = SS_W / gl_W$$

Para determinar el cuadrado medio entre, divide SS_B por los grados de libertad entre:

$$MS_B = SS_B / gl_B$$

Estos valores cuadrados medios ahora son directamente comparables y pueden usarse para calcular F :

$$F = MS_B / MS_W$$

El cuadrado medio dentro de grupos (MS_W) es el denominador de la prueba F . Al igual que el SE_D en la prueba t , representa la cantidad de varianza que puede esperarse debido solo a ocurrencias fortuitas. La relación F compara MS_W con las diferencias entre las medias representadas por el cuadrado medio entre grupos (MS_B).

En ANOVA, MS_W a menudo se denomina error cuadrático medio y puede denominarse MS_E . El término error no significa una equivocación; significa la variación que puede ocurrir solo por azar, o la variación que no se explica por los efectos del tratamiento. A partir de este momento, utilizaremos la representación MS_E , pero es sinónimo de MS_W .

Cuando el MS_E es grande, tiende a enmascarar los pequeños efectos del tratamiento. Cuando el MS_E es pequeño, es más fácil identificar los efectos del tratamiento. La intención de todos los diseños de investigación es mantener MS_E lo más pequeño posible. El valor de MS_E depende de la variabilidad en los datos y del número de puntajes (N) que componen los datos. De estos dos factores, solo el investigador puede controlar N . Por lo tanto, los estudios con grandes N son más potentes para detectar diferencias de medias.

Cálculo de F . Se puede calcular F usando las fórmulas presentadas hasta ahora. Los conceptos para este cálculo de F son fáciles de entender porque se basan en la definición de términos ya introducidos. Pero, al igual que las fórmulas de definición para la desviación estándar y la correlación de Pearson, el método de definición para ANOVA es tedioso cuando las puntuaciones sin procesar contienen valores decimales y N es grande. Un segundo método computacional, el método de puntajes originales (brutos), es más fácil de usar. El método de definición casi nunca se usa en aplicaciones prácticas, y el método de puntajes originales debe usarse para resolver problemas reales; ambos métodos, cuando se aplican correctamente, producen las mismas respuestas.

Método de puntajes originales. El método de definición para calcular F es bastante tedioso de llevar a cabo para proporcionar datos que contienen valores decimales y grandes N . Las fórmulas de cálculo alternativas para SS_B y SS_W que usan puntajes originales en lugar de puntajes de desviación hacen que el cálculo sea mucho más fácil. Recuérdese la ecuación donde $SS_T = SS_B + SS_W$. Debido a que SS_T y SS_B son más fáciles de calcular que SS_W , SS_W puede determinarse mediante la ecuación $SS_W = SS_T - SS_B$. Una vez que se encuentran SS_B y SS_W , el resto del cálculo de F es el mismo que se describe en el método de definición.

Las fórmulas computacionales para calcular SS_B , SS_T y SS_W son las siguientes:

$$SS_B = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} \dots + \frac{(\Sigma X_k)^2}{n_k} - \frac{(\Sigma X_T)^2}{N}$$

$$SS_T = \Sigma X^2 - \frac{(\Sigma X_T)^2}{N}$$

$$SS_W = SS_T - SS_B$$

El método con puntajes originales requiere que cada valor de datos sin procesar (X) sea cuadrado y se calculen las sumas de cuadrados para cada grupo. Los datos originales de una prueba de fuerza de brazos luego de entrenar en este ejemplo, se cuadran y se suman.

Puntajes originales						Suma de cuadrados de puntajes originales				
X_1	X_2	X_3	X_4	X_5		$\frac{(X_1)^2}{2}$	$\frac{(X_2)^2}{2}$	$\frac{(X_3)^2}{2}$	$\frac{(X_4)^2}{2}$	$\frac{(X_5)^2}{2}$
4	5	5	8	5		16	25	25	64	25
5	7	4	4	4		25	49	16	16	16
6	9	6	6	3		36	81	36	36	9
7	8	5	8	4		49	64	25	64	16
4	9	5	5	6		16	81	25	25	36
6	7	6	6	4		36	49	36	36	16
7	1	4	7	5		25	100	16	49	25
	0									
Σ	3	5	3	4	3	$\Sigma=2$	$\Sigma=4$	$\Sigma=1$	$\Sigma=2$	$\Sigma=1$
X	7	5	5	4	1	03	49	79	90	43
n	7	Σ	2			$\Sigma X^2 = 203 + 449 + 179 + 290 + 143 = 1264$				
		X_T	0							
		=	2							
M	5,	7,	5,	6,	4,					
	2	8	0	2	4					
	9	6	0	9	3					

Tabla 9.2 Puntajes originales y suma de los cuadrados de una prueba de fuerza de brazos

A continuación, se procede a calcular SS_B , SS_T y SS_W , según las ecuaciones arriba anotadas:

$$SS_B = \frac{37^2}{7} + \frac{55^2}{7} + \frac{35^2}{7} + \frac{44^2}{7} + \frac{31^2}{7} - \frac{202^2}{35} = 1216,57 - 1165,83 = 50,74$$

$$SS_T = 1264 - \frac{202^2}{35} = 98,17$$

$$SS_W = 98,17 - 50,74 = 47,43$$

Los cálculos de los grados de libertad, los valores cuadrados medios y F son los mismos en ambos métodos:

$$gl_B = 5 - 1 = 4$$

$$gl_E = 35 - 5 = 30$$

$$MS_B = \frac{SS_B}{gl_B} = \frac{50,74}{4} = 12,69$$

$$MS_E = \frac{SS_W}{gl_W} = \frac{47,43}{30} = 1,58$$

$$F = \frac{MS_B}{MS_E} = \frac{12,69}{1,58} = 8,03$$

Los resultados del análisis de varianza se reportan como en la tabla 9.3:

Fuente de la varianza	Suma de los cuadrados	gl	Media de los cuadrados	F	p
Entre grupos	50,74	4	12,69	8,03	<0.01
Dentro de grupos (error)	47,43	30	1,58		
Totales	98,17	34			
Tabla 9.3 Informe del análisis de varianza					

Determinación del significado de F . El significado de F se determina consultando las tablas correspondientes en el apéndice de cualquier texto de estadística, que muestra los valores de F para los grados de libertad apropiados entre y dentro de los grupos. En estas tablas, gl_B se lee en la parte superior y horizontal de la tabla y gl_E se lee en el lado izquierdo y vertical.

La tabla 9.4 muestra los valores de F para $p = 0.10$; cuando $gl_B = 4$ y $gl_E = 30$, el valor crítico F es 2.14; debido a que el F (8.03) obtenido excede este valor, se pasa a la siguiente tabla.

En la tabla 9.5 se enumeran los valores F para $p = 0.05$, donde el F crítico = 2.69. La F obtenida (8.03) sigue siendo mayor que la F crítica, por lo que se pasa a la siguiente tabla.

Y en la tabla 9.6, que tiene los valores F para $p = 0.01$ y en el cual F crítico = 4.02, la F obtenida (8.03) también excede este valor F crítico, por lo que la F obtenida se declara significativa en $p < 0.01$.

Esto significa que las probabilidades son menores de 1 en 100 de que una F mayor a 4.02 se obtenga por azar. Por tanto, el nivel de confianza alcanzado por el valor F obtenido (8.03) es superior al 99%. Se concluye que el tratamiento tuvo efecto en al menos uno de los grupos y se rechaza H_0 .

gl_E	df_B									
	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60

Tabla 9.4 Valores de la distribución F ($p = 0,10$)

gl_E	df_B									
	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Tabla 9.5 Valores de la distribución F ($p = 0,05$)

gl_E	df_B									
	1	2	3	4	5	6	7	8	9	10
1	4052	5000	5403	5625	5764	5859	5628	5982	6022	6056
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.46	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.56	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Tabla 9.6 Valores de la distribución F ($p = 0,01$)

Pruebas *post hoc*. Acá solo se hace una referencia a este tipo de procedimiento, si bien es necesario considerarlo al calcular F cuando se analiza tres o más grupos mediante ANOVA.

Una F significativa por sí sola no especifica cuáles grupos difieren entre sí. Solo indica que hay diferencias en algún lugar entre los grupos. Para identificar los grupos que difieren significativamente del otro, se debe realizar una prueba *post hoc*.

Una prueba *post hoc* es similar a una prueba t , excepto que las pruebas *post hoc* tienen incorporada una corrección para los errores alfa. En este tipo de prueba, algunas

son más conservadoras que otras. Conservadora significa que las pruebas son menos poderosas porque requieren mayores diferencias medias antes de poder declarar la significancia. Las pruebas conservadoras ofrecen una mayor protección contra errores de tipo I, pero son más susceptibles a errores de tipo II.

Se pueden aplicar varias pruebas post hoc para determinar la ubicación de las diferencias de grupo después de que se haya encontrado una F significativa. Dos de las pruebas más utilizadas son el intervalo de confianza de Scheffé (I) y la diferencia honestamente significativa (HSD) de Tukey. Scheffé permite todas las comparaciones posibles, mientras que Tukey sólo permite comparaciones por pares. Tukey es más fácil de aplicar y es apropiado en la mayoría de los diseños de investigación.

Análisis de varianza con mediciones repetidas

Uno de los diseños de investigación más comunes en el campo de la actividad física implica medir a los sujetos antes del tratamiento (prueba previa) y luego de este proceso (prueba posterior). Se utiliza una prueba t dependiente con muestras emparejadas o correlacionadas para analizar tales datos, porque los mismos sujetos se miden dos veces (medidas repetidas). En ANOVA, este tipo de diseño se denomina diseño intra-sujetos. Cuando se administran tres o más pruebas (por ejemplo, si se administran una prueba previa, intermedia y posterior, con el tratamiento antes y después de la prueba intermedia) se necesita ANOVA con medidas repetidas para analizar adecuadamente las diferencias entre las tres pruebas.

El ANOVA simple (descrito anteriormente) asume que los valores medios se toman de grupos independientes que no tienen relación entre sí. En este diseño de grupo independiente, la variabilidad total es la suma de

- variabilidad entre personas en los diferentes grupos (variabilidad interindividual)
- variabilidad dentro de las puntuaciones de una persona (variabilidad interindividual)
- variabilidad entre grupos debido a los efectos del tratamiento, y
- variabilidad debida al error (variabilidad inexplicable).

Cuando solo hay un grupo de sujetos, pero se miden más de una vez, los conjuntos de datos son dependientes. Se espera que la variabilidad total para un solo grupo de sujetos medidos más de una vez sea menor que si los puntajes provinieran de diferentes grupos de personas (es decir, si los puntajes fueran independientes), porque la variabilidad interindividual se ha eliminado al usar un grupo único. Esto tiende a reducir el término MS_E en el denominador de F de una manera similar a la corrección realizada al error estándar de la diferencia en la prueba t .

La ventaja del diseño de medidas repetidas es que los sujetos sirven como su propio control. Si se han controlado todos los demás factores relevantes, cualquier diferencia observada entre las medias debe deberse a: (a) el tratamiento, (b) las variaciones dentro de los sujetos (variabilidad intraindividual) o (c) el error (variabilidad inexplicable). La va-

riabilidad entre sujetos (variabilidad interindividual) ya no es un factor. Las fórmulas presentadas anteriormente para ANOVA simple se modifican para tener en cuenta las medidas repetidas.

Supuestos de ANOVA con mediciones repetidas. Excepto por la independencia de las muestras, las suposiciones para ANOVA simple también son válidas para ANOVA de mediciones repetidas (diseños intra-sujetos). Pero con diseños de medidas repetidas, también debemos considerar las relaciones entre las medidas repetidas. ANOVA de medidas repetidas debe cumplir con el supuesto de esfericidad, a veces denominado simetría compuesta. La esfericidad requiere que lo repetido demuestre homogeneidad de varianza y homogeneidad de covarianza. La homogeneidad de la covarianza significa que la relación o correlaciones de la variable dependiente entre las tres o más medidas repetidas son iguales. Cuando solo se emplean dos medidas repetidas (como las medidas pre-post para una prueba t), esta suposición no es aplicable, porque solo hay un coeficiente de correlación que se puede calcular, la correlación entre las puntuaciones anteriores y posteriores a la prueba.

Cálculo de ANOVA con mediciones repetidas. Para demostrar cómo calcular ANOVA con medidas repetidas, se analiza un estudio hipotético. Un estudiante de posgrado que estudia el comportamiento motor estaba interesado en la disminución de la capacidad de equilibrio que experimentan los ciclistas a medida que aumenta su fatiga en la carrera. Para medir esto, el investigador colocó una bicicleta de carreras en un ergómetro de ruedas. Las ruedas delanteras y traseras de la bicicleta se colocaron en la parte superior de rodillos cilíndricos de modo que a medida que las ruedas y los rodillos giraban, los ciclistas rodaban en el lugar.

Se pintó una franja de 4 pulgadas de ancho en el medio del rodillo delantero y en la misma dirección de la bicicleta, y se le pidió al ciclista que mantuviera la rueda delantera sin salirse de la franja. Los rodillos traseros se conectaron a un sistema de frenado para brindar resistencia a la rueda trasera de la bicicleta. El equilibrio se indicó mediante el bamboleo en la rueda delantera y se midió contando el número de veces por minuto que la rueda delantera de la bicicleta se desviaba de la franja de 4 pulgadas.

A medida que aumentaba la resistencia en las ruedas traseras, aumentaba la fatiga fisiológica y se hacía cada vez más difícil mantener la rueda delantera en la raya. Los sujetos trabajaron en bicicleta durante 15 minutos, divididos en cinco períodos de 3 minutos con el fin de recopilar datos. Se recopilaron datos sobre el número de errores de equilibrio durante el último minuto de cada período de 3 minutos y la resistencia se incrementó al final de ese período. En este diseño, la variable dependiente son los errores de equilibrio y la variable independiente es el aumento de la resistencia (fatiga).

La Tabla 9.7 presenta los datos brutos en columnas y filas. Los datos (en errores por minuto) de los sujetos ($N = 10$) están en las filas y los datos de los ensayos ($k = 5$ medidas repetidas) están en las columnas. La suma de las filas (ΣR) es la puntuación total de cada sujeto en los cinco ensayos, y la suma al cuadrado $(\Sigma R)^2$ se presenta a la derecha de la tabla. La suma de cada columna (ΣC) es el total de todos los sujetos en una prueba

determinada; ΣX_T se presenta en la parte inferior de la tabla. La tabla 9.8 presenta el cuadrado de los datos brutos.

Su- jeto	Mi- nuto 3	Mi- nuto 6	Mi- nuto 9	Minuto 12	Minuto 15	ΣR	$(\Sigma R)^2$
1	7	7	23	36	70	143	20499
2	12	22	26	26	20	106	11236
3	11	6	9	31	30	87	7569
4	10	18	16	40	25	109	11881
5	6	12	9	28	37	92	8464
6	13	21	30	55	65	184	33856
7	5	0	2	10	11	28	784
8	15	18	22	37	42	134	17956
9	0	2	0	16	11	29	841
10	6	8	27	32	54	127	16129
ΣC	85	114	164	311	365	1039	129165
Mean	8,5	11,4	16,4	31,1	36,5		
$\Sigma X_T =$ 85 + 114 + 164 + 311 + 365 = 1039							
Tabla 9.7 Datos brutos: errores de equilibrio por minuto							

Sujeto	Minuto 3	Minuto 6	Minuto 9	Minuto 12	Minuto 15
1	49	49	529	1296	4900
2	144	484	676	676	400
3	121	36	81	961	900
4	100	324	256	1600	625
5	36	144	81	784	1369
6	169	441	900	3025	4225
7	25	0	4	100	121
8	225	324	484	1369	1764
9	0	4	0	256	121
10	36	64	729	1024	2916
Total	905	1870	3740	11091	17341
ΣX^2	905 + 1870 + 3740 + 11091 + 17341 = 34947				
Tabla 9.8 Cuadrado de datos brutos					

Pasos para el cálculo de F

1. Organice los datos brutos (X) de forma que se observen los datos de los sujetos en las filas (R) y las medidas repetidas en las columnas (C) como se muestra en la tabla 9.4.

2. Calcule los totales de las filas ($\Sigma R_1, \Sigma R_2, \Sigma R_3, \dots$), los totales de las columnas ($\Sigma C_1, \Sigma C_2, \Sigma C_3, \dots$) y el gran total general: $\Sigma X_T = 1039$.
3. Cuadre cada fila total $(\Sigma R)^2$ y sume estos valores: $\Sigma (\Sigma R)^2 = 129165$.
4. Calcule los valores medios de las columnas.
5. Eleve al cuadrado cada puntuación bruta (X^2) y calcule los totales de la columna al cuadrado y el total general: $\Sigma X^2 = 34947$, como se muestra en la tabla 9.5.
6. Calcule la suma de cuadrados entre columnas (SS_C). Ésta es la variabilidad debida a los efectos del tratamiento.

$$SS_C = \frac{(\Sigma C_1)^2 + (\Sigma C_2)^2 + (\Sigma C_3)^2 + \dots + (\Sigma C_K)^2}{N} - \frac{(\Sigma X_T)^2}{(N)(k)}$$

Por tanto, en el problema de la bicicleta:

$$SS_C = \frac{(85)^2 + (114)^2 + (164)^2 + (311)^2 + (365)^2}{10} - \frac{(1039)^2}{(10)(5)} =$$

$$SS_C = 27706,3 - 21590,42 = 6115,88$$

7. Calcule la suma de cuadrados entre filas (SS_R). Esta es la variabilidad debida a las diferencias entre los sujetos.

$$SS_R = \frac{\Sigma (\Sigma R)^2}{k} - \frac{(\Sigma X_T)^2}{(N)(k)}$$

Por tanto, en el problema de la bicicleta:

$$SS_R = \frac{129165}{5} - \frac{1039^2}{(10)(5)} = 25833,0 - 21590,42 = 4242,58$$

8. Calcule la suma total de cuadrados (SST). Esta es la variabilidad debida a los sujetos (filas), el tratamiento (columnas) y la variabilidad residual inexplicable (error).

$$SS_T = \Sigma X^2 - \frac{(\Sigma X_T)^2}{(N)(k)}$$

Por tanto, en el problema de la bicicleta:

$$SS_T = 34947 - \frac{(1039)^2}{(10)(5)} = 34947 - 21590,42 = 13356,58$$

9. Calcule la suma de cuadrados debida al error (SSE). Esta es la varianza inexplicable que se utilizará en el denominador de F . Es la cantidad de varianza que se puede atribuir al azar. Debido a que la variabilidad de las columnas (tratamientos),

las filas (sujetos) y el error deben ser iguales a la variabilidad total, podemos encontrar SS_E por sustracción:

$$SS_E = SS_T - SS_C - SS_R$$

Por lo tanto:

$$SS_E = 13356,58 - 6115,88 - 4242,58 = 2998,12$$

10. Calcule los grados de libertad para cada fuente de varianza.

Columnas (tratamiento):

$$gl_C = k - 1$$

$$gl_C = 5 - 1 = 4$$

Filas (sujetos):

$$gl_R = N - 1$$

$$gl_R = 10 - 1 = 9$$

Error:

$$gl_E = (k - 1)(N - 1)$$

$$gl_E = (5 - 1)(10 - 1) = 36$$

Total:

$$gl_T = (N)(k) - 1$$

$$gl_T = (10)(5) - 1 = 49$$

11. Calcule la media cuadrada para cada fuente de varianza

Media cuadrada de columnas (intentos):

$$MS_C = \frac{SS_C}{gl_C}$$

$$MS_C = \frac{6115,88}{4} = 1528,97$$

Media cuadrada de filas (sujetos)

$$MS_R = \frac{SS_R}{gl_R}$$

$$MS_R = \frac{4242,58}{9} = 471,40$$

Media cuadrada del error:

$$MS_E = \frac{SS_E}{gl_E}$$

$$MS_E = \frac{2998,12}{36} = 83,28$$

12. Calcule F para las columnas (tratamiento)

$$F_C = \frac{MS_C}{MS_E}$$

$$F_C = \frac{1528,97}{83,28} = 18,36$$

Es posible calcular un valor F para sujetos o filas ($F_R = MS_R / MS_E$). Pero este valor no es de interés en este momento porque simplemente representa una prueba de la variabilidad entre los sujetos. Este valor de F_R no es importante para la pregunta de investigación que se está considerando: ¿Un aumento en la fatiga da como resultado un aumento en los errores de equilibrio medio entre los intentos (columnas)? Se usará F_R más adelante cuando se discuta la fiabilidad intraclase.

Determinación del significado de F . Para determinar la importancia de F para las columnas, miramos en las tablas correspondientes. Los grados de libertad para las columnas gl_C son los mismos que df_B en el análisis de varianza simple porque las columnas representan la variabilidad entre intentos. Los grados de libertad para el error df_E se utilizan de la misma tabla para medir la variabilidad dentro del grupo. La tabla 9.6 muestra que para gl (df_B 4, df_E 36) y F de 4.02 se necesita para alcanzar $p = 0.01$. Debido a que nuestra F (18.36) obtenida excede fácilmente 4.02, se concluye que existen diferencias en algún lugar entre los valores medios para los cinco intentos en $p < 0.01$.

Corrección de las violaciones de los supuestos de esfericidad. Recuerde que uno de los supuestos del ANOVA de medidas repetidas es que la varianza de varias medidas repetidas, o intentos, es igual (homogeneidad de varianza) y las correlaciones entre todas las combinaciones posibles de intentos (homogeneidad de covarianza) son iguales. A esto se le llama el supuesto de esfericidad. Si el supuesto de esfericidad no es cierto, aumenta la probabilidad de cometer un error de tipo I (rechazar H_0 cuando es cierto). Greenhouse-Geisser y Huynh-Feldt han sugerido métodos para corregir una infracción del supuesto. Ambos procedimientos de corrección modifican los valores gl para columnas (tratamiento) y error.

Fiabilidad intraclase

El coeficiente de correlación de Pearson a veces se usa incorrectamente para calcular la fiabilidad test-retest. La correlación compara las desviaciones (fluctuaciones en las puntuaciones de los sujetos) de la media en dos mediciones, pero no es sensible a los cambios en las medias de las puntuaciones. Si las fluctuaciones en las puntuaciones de la primera a la segunda prueba ocurren todas de manera sistemática (es decir, todas aumentan o disminuyen en la misma cantidad), el orden y las desviaciones de las puntuaciones en las dos pruebas serán los mismos (es decir, el coeficiente de correlación será alto), pero las dos

medias pueden diferir significativamente. En tales casos, la r de Pearson puede llevarnos a concluir erróneamente que los puntajes son altamente fiables, cuando en realidad están cambiando considerablemente. La r de Pearson es incapaz de detectar el cambio en las medias. Otra limitación de la correlación es que solo se puede aplicar a dos conjuntos de datos a la vez, pero es posible que el interés sea la fiabilidad de varias mediciones.

La fiabilidad intraclase resuelve estas dos objeciones: se puede usar en dos o más medidas simultáneamente, y es sensible a los cambios tanto en el orden como en la magnitud (diferencias medias) de los valores repetidos. Se llama intraclase porque está diseñado para analizar datos de medidas repetidas sobre la misma variable. Interclase (es decir, Pearson) es apropiado para el análisis de dos variables diferentes. La correlación de Pearson indica el mantenimiento o el orden relativo entre dos conjuntos de datos, que pueden usarse en diferentes escalas de medición y, por lo tanto, es apropiado para el análisis interclase. De hecho, a menudo se usa de esta manera para evaluar la validez. En estas condiciones, no nos preocupan las diferencias en los valores medios porque las unidades de medida pueden diferir (por ejemplo, tiempo frente a distancia).

Sin embargo, al determinar la fiabilidad, los investigadores suelen estar interesados en cambios tanto en el orden de los valores como en su magnitud. Se puede usar ANOVA con medidas repetidas para evaluar la fiabilidad con sensibilidad al orden y magnitud combinados. Para determinar la fiabilidad de la medición de los errores de equilibrio en la cinta de correr bicicleta descrita anteriormente, el investigador recopiló los datos de los errores de equilibrio en 5 sujetos piloto (no los mismos que en el estudio) durante 15 minutos (midiendo cada 3 minutos), sin resistencia en los rodillos traseros. Los datos de los 5 sujetos piloto se muestran en la tabla 9.9.

Sujetos	Minuto 3	Minuto 6	Minuto 9	Minuto 12	Minuto 15
1	10	12	11	9	10
2	14	15	13	15	16
3	6	5	6	7	8
4	12	10	11	13	12
5	8	9	7	8	9
Media=	10,0	10,2	9,6	10,4	11,0
Tabla 9.9 Evaluación de la fiabilidad en los sujetos del estudio piloto					

El análisis de varianza de medidas repetidas con software produjo los resultados en la tabla 9.10. La variabilidad entre los sujetos (filas) se incluye en esta tabla. El valor alto de F (43.63, $p < 0.0000$) para las filas indica que los sujetos fueron significativamente diferentes en el número de errores que cometieron en todos los intentos. Esto se espera porque no todos los sujetos tienen la misma capacidad de conducir bicicletas. La F no significativa para los tratamientos indica que no hay diferencias significativas entre las medias de los cinco intentos. Esto no es sorprendente porque las condiciones de tratamiento para todos los intentos siguieron siendo las mismas (es decir, sin aumento de la resistencia).

Fuente	Suma de cuadrados	gl	Media cuadrada	F	p
Tratamiento (columnas)	5,36	4	1,34	1,18	N.S.
Sujetos (filas)	198,96	4	49,74	43,63	0.0000
T X S (error)	18,24	16	1,14		
Total	222,56	24			

Tabla 9.10 ANOVA: Errores en la bicicleta en un estudio piloto

Para calcular la fiabilidad intraclase, debemos calcular un valor medio cuadrado que represente la suma de los cambios en la media (columna o efectos del tratamiento) y el error (MS_{C+E}):

$$MS_{C+E} = \frac{SS_C + SS_E}{gl_C + gl_E}$$

Para calcular MS_{C+E} , se utilizan los datos de la tabla 9.7:

$$MS_{C+E} = \frac{5,36 + 18,24}{4 + 16} = 1,18$$

Luego se determina la fiabilidad intraclase (R_I):

$$R_I = \frac{MS_R - MS_{C+E}}{MS_R}$$

A continuación, se utilizan los datos de la tabla 9.7:

$$R_I = \frac{49,74 - 1,18}{49,74} = 0,976$$

La fiabilidad o consistencia de los sujetos en las pruebas (R_I) de los datos de la bicicleta, incluidos los efectos de los cambios en los valores medios, es de 0,976. Este es un valor de fiabilidad muy alto. Como regla general, los valores de R_I por encima de 0,90 se consideran altos, de 0,80 a 0,89 moderados y por debajo de 0,80 cuestionables para datos fisiológicos. En las ciencias del comportamiento, los valores entre 0,70 y 0,80 pueden considerarse aceptables, dependiendo del tipo de instrumento de medición aplicado.

Debido a que los valores de los hombres no cambian significativamente (F para columnas, o tratamiento, no es significativo), y el R_I (0,976) es positivo alto, concluimos que los sujetos pudieron pedalear con poca variación en su rendimiento minuto a minuto. Esto confirma que el instrumento de cinta de correr para bicicleta diseñado para medir errores de equilibrio es confiable. De vez en cuando, se desea medir la fiabilidad sin considerar los cambios en los valores medios durante los ensayos. Si ignoramos la varianza debida a cambios en los valores medios (columnas), podemos calcular la varianza debida

al orden (filas) solo entre los sujetos durante las diversas pruebas. Este valor se designa como R_2 y, a veces, se denomina coeficiente alfa. Para calcular R_2 , el valor de MS_E se sustituye por MS_{C+E} :

$$R_2 = \frac{MS_R - MS_E}{MS_R}$$

El valor de R_2 se calcula tomando en cuenta los datos de la prueba piloto en la tabla 9.7:

$$R_2 = \frac{49,74 - 1,14}{49,74} = 0,977$$

Cuando los valores medios no cambian, R_2 no se desviará mucho de R_1 . El valor de R_2 (0,977) difiere poco de R_1 (0,976) porque las medias son muy similares en los cinco ensayos. Es apropiado utilizar R_1 para evaluar la fiabilidad intraclase, incluido el orden y la magnitud, mientras que R_2 es apropiado cuando los cambios en las medias deben ignorarse.

Interpretación de la fiabilidad intraclase. Si tanto el orden de la puntuación de los sujetos individuales como los valores medios de las mediciones posteriores no cambian mucho, tanto R_1 como R_2 serán altos. Si tanto el orden como los valores medios cambian, entonces R_1 y R_2 disminuirán, pero R_1 será el que más se reducirá.

Para ilustrar este concepto, analícense los datos de la bicicleta cuando la resistencia se incrementó durante las pruebas. Si aplicamos el análisis de fiabilidad intraclase a los datos de error de la bicicleta de la tabla 9.4, $R_1 = 0,517$ y $R_2 = 0,823$. Esta diferencia refleja el hecho de que R_1 es sensible a los cambios de valor medio durante los ensayos, pero R_2 no lo es; se confirma que los valores medios sí cambian drásticamente. Cuando se factorizan las diferencias de medias entre los intentos mediante el uso de R_2 , la fiabilidad aumenta a 0,823.

El método adecuado para determinar la fiabilidad es calcularlo con datos piloto y luego confirmarlo con los datos experimentales. Si los datos piloto no demuestran una fiabilidad aceptable, las técnicas utilizadas para medir la variable de interés en el estudio deben revisarse hasta que se obtenga una fiabilidad aceptable. Sin evidencia de que las mediciones son fiables en un estudio piloto, no podemos estar seguros de que los datos obtenidos en el estudio real sean correctos.

Coeficiente alfa. El coeficiente alfa se usa a menudo para determinar la fiabilidad de datos dicotómicos. Cuando los datos son ordinales, se puede usar el coeficiente alfa para determinar la fiabilidad. De hecho, con datos ordinales o de intervalo el coeficiente alfa es el mismo que R_2 . El coeficiente alfa es una estimación de la fiabilidad de una puntuación de criterio que es la suma de las puntuaciones del ensayo en un día. No hay ninguna razón por la que el coeficiente no pueda aplicarse a datos de varios días o de varios jueces para estimar estabilidad, fiabilidad u objetividad. Determinamos el coeficiente alfa utilizando la siguiente fórmula:

$$r_{\alpha} = \left(\frac{k}{k-1} \right) \left(\frac{S_x^2 - S_j^2}{S_x^2} \right)$$

donde r_{α} es el coeficiente alfa, k es el número de intentos, s_x^2 es la varianza de la puntuación del criterio y s_j^2 es la suma de las varianzas de los ensayos. En el cálculo del coeficiente alfa, el procedimiento también implica múltiples pasos:

Paso 1. La forma más sencilla de hacer esto es establecer una tabla como la siguiente:

	Intento 1	Intento 2	Intento 3	TOTAL X
Persona	X ₁	X ₂	X ₃	X
A	5	6	7	18
B	3	3	4	10
C	4	4	5	13
D	7	6	6	19
E	6	7	5	18
	25	26	27	78

Tabla 9.11 Datos para el cálculo del coeficiente alfa

Paso 2. Con la computadora se calcula la varianza para X_1 , X_2 , X_3 y X . Las varianzas son las siguientes:

$$s_1^2 = 2,50; s_2^2 = 2,70; s_3^2 = 1,30; s_x^2 = 15,30$$

Paso 3. A continuación se calcula r_{α} , donde k es 3, s_x^2 es 15,30 y s_j^2 es 6,50 (la suma de s_1^2 , s_2^2 y $s_3^2 = 2,50 + 2,70 + 1,30$) teniendo como resultado de coeficiente alfa r 0,86:

$$r_{\alpha} = \left(\frac{3}{3-1} \right) \left(\frac{15,30 - 6,50}{15,30} \right) = \left(\frac{3}{2} \right) \left(\frac{8,80}{15,30} \right) = \frac{26,4}{30,60} = 0,86$$

Además de proporcionar las variaciones necesarias, la computadora también puede proporcionar las correlaciones entre las puntuaciones de los intentos y la suma de las puntuaciones de los intentos, lo que podría ser útil para decidir qué pruebas sumar como puntuaciones de criterio. Sólo deben retenerse los intentos que se correlacionen positivamente en un valor razonablemente alto con la suma de las puntuaciones de los intentos. Normalmente, con las pruebas de rendimiento motor, cada prueba se correlaciona con la suma de todas las pruebas.

Chi-cuadrado (clasificación única)

Como se discutió anteriormente, los datos se pueden clasificar en cuatro categorías: nominal, ordinal, intervalo y razón. Cuando se recopilan datos de razón o intervalo, el aná-

lisis mediante técnicas estadísticas paramétricas es apropiado. El coeficiente de correlación de Pearson, la prueba t y el ANOVA en todas sus variedades son técnicas estadísticas paramétricas. Pero cuando los datos son de tipo nominal u ordinal no se cumplen los supuestos de normalidad y se deben utilizar técnicas estadísticas no paramétricas.

Hay varias técnicas estadísticas no paramétricas más utilizadas entre dos conjuntos. El Chi-cuadrado compara dos o más conjuntos de datos nominales que se han organizado en categorías por conteos de frecuencia. El coeficiente de correlación del orden de rango de Spearman determina la relación entre dos conjuntos de datos ordinales. La prueba U de Mann-Whitney determina la significancia de la diferencia entre las clasificaciones ordinales de dos grupos de sujetos clasificados en la misma variable. El ANOVA de Kruskal-Wallis por rangos se utiliza para comparar el ranking de tres o más grupos independientes, y es similar a un ANOVA simple. El ANOVA de dos vías de Friedman por rangos es similar al ANOVA de medidas repetidas, determina la significancia de la diferencia entre rangos sobre los mismos sujetos. A continuación, se presenta lo referido a chi-cuadrado.

El chi-cuadrado se utiliza para comparar dos o más conjuntos de datos nominales que se han organizado en categorías por recuentos de frecuencia. El género es un ejemplo de datos nominales. Una persona se clasifica como hombre o mujer, y simplemente contamos el número en cada categoría. No hay variabilidad dentro de la categoría; todos los sujetos tienen igual valor. La clasificación debe ser mutuamente excluyente; un sujeto puede clasificarse en una sola de las categorías.

Las calificaciones son otro ejemplo de este tipo de clasificación. Después de considerar todos los puntajes de las pruebas de un estudiante, el maestro le asigna al estudiante una calificación para el curso. La calificación es solo un nombre (datos nominales) para el desempeño que tiene un cierto significado. Las calificaciones representan valores ordenados (por ejemplo, A, B, C, D y F), pero aquí los trataremos como nominales contando el número de estudiantes en cada categoría.

Un cierto número de alumnos (la frecuencia) reciben A, otros B, C, D, y F. Cuando esto se completa, todas las A son igualmente valiosas, todas las B son igualmente valiosas, y así sucesivamente para las C, D, E y F. La prueba de chi-cuadrado debe usarse para comparar dos conjuntos de datos, clasificados por categoría, y determinar si las frecuencias de las categorías difieren en cantidades mayores de lo que cabría esperar por azar. Revela la importancia de las diferencias en los recuentos de frecuencia. El siguiente ejemplo muestra cómo se puede aplicar chi-cuadrado.

Un ejemplo de aplicación de chi-cuadrado. Un departamento de educación física estableció una política según la cual los docentes deben otorgar ciertos porcentajes de las diversas calificaciones en sus clases. Después de mucha discusión, los docentes acordaron mantener la distribución de calificaciones en aproximadamente 10% en A, 30% B, 35% C, 20% D y 5% F. Los docentes reconocieron que, para una clase pequeña, estos porcentajes pueden ser irreales, pero cuando se combinan todas las clases de cualquier docente, los porcentajes finales deben aproximarse mucho a la política establecida.

Todos aceptaron la política y durante un año o dos la facultad la siguió de cerca. Pero luego, un docente comenzó a otorgar calificaciones más altas de lo que cabría esperar según dicha política. En una lista de 141 estudiantes, este docente otorgó 30 A, 57 B, 32 C, 15 D y 7 F. El director quiso saber si dicha distribución difería significativamente de la política ya aceptada. Con base en 141 estudiantes, la política departamental requeriría 14 A (10%), 42 B (30%), 49 C (35%), 28 D (20%) y 8 F s (5%).

La hipótesis nula (H_0) supone que las diferencias ocurrieron por azar. Si se demuestra que H_0 es falsa con un nivel de confianza aceptable, se acepta H_1 y se debe concluir que una influencia diferente al azar causó las diferencias en el recuento de frecuencias. Para determinar las probabilidades de que la distribución del docente pudiera ocurrir solo por azar (y, por lo tanto, no fue una violación de la política), el director del departamento aplicó la técnica de chi-cuadrado. La fórmula para chi-cuadrado es la siguiente:

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

donde O = la frecuencia observada y E = la frecuencia esperada.

Para aplicar esta fórmula, se debe encontrar la diferencia entre las frecuencias observadas (calificaciones del docente) y esperadas (política del departamento) para cada categoría y elevar al cuadrado cada una de estas diferencias. Luego, cada diferencia al cuadrado se divide por la frecuencia esperada para su categoría. Finalmente obtenemos la suma de estas diferencias cuadradas divididas por la frecuencia esperada. La Tabla 9.12 muestra el proceso.

Calificaciones	A	B	C	D	F	Total
Observadas (calificaciones del docente)	30	57	32	15	7	141*
Esperadas (política del Departamento)	14	42	49	28	8	141*
$O - E$	16	15	-17	-13	-1	
$(O - E)^2$	256	225	289	169	1	
$(O - E)^2 / E$	18,29	5,36	5,90	6,04	0,13	35,72
* La suma de las frecuencias esperadas y observadas debe ser igual						
Tabla 9.12 Cálculo de chi-cuadrado						

A la tabla 9.12 se le llama uno x cinco (1 x 5) porque las frecuencias observadas ocupan solo una fila y tiene cinco columnas.

En este problema, chi-cuadrado = 35,72. El grado de libertad en chi-cuadrado es el número de categorías, k (en este caso k = número de categorías de grado), menos uno ($gl = k-1$, o $5 - 1 = 4$). La tabla 9.13 indica que para $gl = 4$, chi-cuadrado debe ser 13,28 para alcanzar el nivel de confianza del 99% ($p = 0,01$). Debido a que 35,72 excede fácilmente 13,28, las probabilidades de que las dos distribuciones difieran solo por azar son

menos de 1 en 100. El director del departamento rechazó H_0 y concluyó que el docente estaba otorgando calificaciones significativamente más altas que las permitidas por la política departamental.

Chi-cuadrado para dos o más clasificaciones

Chi-cuadrado también se puede utilizar para resolver problemas más complicados. El segundo ejemplo usa dos clasificaciones en una tabla de contingencia, que es una distribución de frecuencias en filas y columnas. Una categoría de frecuencia ocupa las filas y otra categoría ocupa las columnas. El siguiente problema de comportamiento motor involucra una tabla de contingencia de 2 x 3 (dos filas y tres columnas).

gl	0.10	0.05	0.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	14.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.73
12	18.55	21.03	26.22
13	19.81	22.36	27.09
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
25	34.38	37.65	44.31
30	40.26	43.77	50.89
40	51.81	55.76	63.69
50	63.17	67.50	76.15
60	74.40	79.08	88.80
70	85.53	90.53	100.43
80	96.58	101.88	112.33
90	107.57	113.15	124.12
∞	118.50	124.34	135.81

Tabla 9.13 Valores de la distribución chi-cuadrado (χ^2)

Un ejemplo de comportamiento motor. Un docente quería saber cuál de los dos métodos de enseñar el estilo de crol frontal era más efectivo para producir nadadores con una buena técnica de brazada. El primer método fue el método tradicional de las partes, en el que se enseñaba primero la patada, luego la brazada, luego la respiración y finalmente todas las partes se juntaban como una brazada completa. El segundo método fue el método global, en el que después de aprender la flotación y el deslizamiento en decúbito prono, los estudiantes fueron introducidos a todas las partes del estilo simultáneamente y lo practicaron completo durante el resto del semestre.

Después de enseñar a 55 estudiantes con el método parcial y a 52 estudiantes con el método global durante 15 semanas, dos veces por semana durante 1 hora en cada clase, el docente le pidió a otro experto en natación que clasificara la brazada de cada estudiante en una de tres categorías: bueno, promedio o deficiente. Los resultados se muestran en la tabla 9.14; los valores en estas celdas representan el número de estudiantes en cada clase que fueron clasificados de esa manera.

Tipo de instrucción	Categorías de destreza		
	Bueno	Promedio	Deficiente
Partes	15	27	13
Global	21	19	12

Tabla 9.14 Tabla de contingencia chi-cuadrado: clasificación 2 x 3

Ahora surge la pregunta, debido a que hay más estudiantes con el método global en la categoría bueno, y más estudiantes con el método de las partes en las categorías promedio y deficiente, ¿es el método global mejor o estas frecuencias podrían haber ocurrido por casualidad? Entonces, se usa la técnica de chi-cuadrado para obtener la respuesta.

Primero, debemos asumir que el grupo total de 107 estudiantes representa la mejor estimación que se puede obtener para las frecuencias esperadas en cada categoría. Debido a que no existe un grupo de control con el que se pueda comparar cada método, calculamos cuántos estudiantes se esperaría que cayeran en cada categoría si todos hubieran sido enseñados por igual por cualquiera de los métodos (es decir, sin la influencia de los métodos de instrucción global o parcial). El total para cada grupo y para cada categoría debe calcularse como se muestra en la tabla 9.15.

El porcentaje del número total de estudiantes que se espera en cada categoría se calcula dividiendo el número total en esa categoría por el total general, en este caso 107.

Tipo de instrucción	Categorías de destreza			
	Bueno	Promedio	Deficiente	Total
Partes	15	27	13	55
Global	21	19	12	52
Total	36	46	25	107
% del gran total*	33,64	42,99	23,26	----
* Bueno: (36 / 107 = 33,64%) Promedio: (46 / 107 = 42,99%) Deficiente: (25 / 107 = 23,26%)				
Tabla 9.15 Determinación del porcentaje del total				

Con esta información, se puede calcular cuántos estudiantes caerían en cada categoría de cada grupo si a todos se les enseñara por igual. Ésta es la frecuencia esperada para el cálculo de chi-cuadrado. Se determina multiplicando el porcentaje de estudiantes que se espera en cada categoría por el número total del grupo.

Grupo de instrucción por partes			Grupo de instrucción global		
Bueno	0,3364	x 55=	Bueno	0,3364	x 52=
		18,50			17,49
Promedio	0,4200	x 55=	Promedio	0,4299	x 52=
		23,64			22,35
Deficiente	0,2336	x 55=	Deficiente	0,2336	x 52=
		12,85			12,15
Tabla 9.16 Determinación de la frecuencia esperada si se les enseñara por igual					

Como método alternativo y más corto para determinar la frecuencia esperada por celda, es aplicar la siguiente fórmula:

$$\text{Frecuencia esperada} = \frac{(\Sigma \text{fila})(\Sigma \text{columna})}{N}$$

Por ejemplo, para los estudiantes de la categoría bueno del grupo de la instrucción por partes se calcula así:

$$\text{Frecuencia esperada} = \frac{(55)(36)}{107} = 18,50$$

Las frecuencias esperadas se ingresan luego en una tabla junto con las frecuencias observadas (Tabla 9.17):

Instrucción	Bueno		Promedio		Deficiente	
	Observado	Esperado	Observado	Esperado	Observado	Esperado
Partes	15	18,50	27	23,64	13	12,85
Global	21	17,49	19	22,35	12	12,15

Tabla 9.17 Frecuencias observadas y esperadas

Las diferencias entre las frecuencias observadas y esperadas ahora se calculan, se elevan al cuadrado, se dividen por el valor esperado y se suman para calcular el estadístico chi-cuadrado. La tabla 9.18 demuestra el proceso; para este ejemplo, chi-cuadrado (χ^2) es 2,35.

Observado	Esperado	$O - E$	$(O - E)^2$	$(O - E)^2 / E$
15	18,50	-3,50	12,25	0,662
21	17,49	3,51	12,32	0,702
27	23,64	3,36	11,29	0,478
19	22,35	-3,35	11,22	0,502
13	12,85	0,15	0,02	0,002
12	12,15	-0,15	0,02	0,002
			Chi-cuadrado =	2,35

Tabla 9.18 Tabla de contingencia del cálculo de chi-cuadrado

Los grados de libertad para una tabla de contingencia de chi-cuadrado son el número de filas menos uno por el número de columnas menos uno:

$$gl = (F - 1)(C - 1)$$

En este ejemplo tenemos:

$$gl = (2 - 1)(3 - 1) = (1)(2) = 2$$

Para $gl = 2$, la tabla 9.13 indica que se necesita un valor de chi-cuadrado de 4.61 para el nivel de confianza del 90%. Debido a que el valor calculado para χ^2 (2.35) es menor que eso, debemos aceptar H_0 . Concluimos que las frecuencias de los grupos no son significativamente diferentes y que los estudiantes aprenden igualmente bien con cualquiera de los métodos. Las pequeñas diferencias observadas son el resultado de sucesos por azar.

Limitaciones de chi-cuadrado. Chi-cuadrado no se aplica a muestras pequeñas, especialmente aquellas en una tabla de 2 x 2. El número total de frecuencias (N) debe ser al menos 20, y el valor de cada celda en la fila de frecuencias esperadas no debe ser menor que 1.

Además, en la tabla 2 x 2, algunos estadísticos sugieren restar 0,5 de cada de los valores ($O-E$) antes de elevarlos al cuadrado para proteger contra errores de tipo I.

La técnica de chi-cuadrado puede resultar muy útil para analizar los recuentos de frecuencias por categorías. Con este método se puede analizar cualquier número de grupos o categorías. Chi-cuadrado es relativamente fácil de aplicar y es aplicable a muchos problemas en kinesiología.

Capítulo 10. Diseño y selección de pruebas de condición física

Muchas de las pruebas utilizadas para medir la condición física no son ni efectivas ni eficientes, mientras que otras producen resultados que no permiten interpretaciones erróneas porque sus responsables han sido muy cuidadosos tanto en la selección de las pruebas como en la planificación de los procedimientos. Para la planificación de una prueba de condición física se debe, en primer lugar, seleccionar o diseñar la prueba. En segundo lugar, se administra la prueba y, en tercer lugar, se evalúan los resultados.

Existen diversas maneras de aplicar una o varias pruebas. En el caso de los docentes de educación física, pueden aplicar una prueba a todo el grupo en forma simultánea en la prueba de capacidad aeróbica, o aplicar una prueba a la mitad del grupo mientras el resto ayuda en el caso de una prueba de abdominales; también se puede medir a un miembro del grupo mientras los demás observan como en el caso de las mediciones de los pliegues de grasa. En ambientes diferentes al educativo se mide una persona a la vez, lo que conforma un modelo clínico propio del entrenamiento deportivo, la terapia física o la rehabilitación cardíaca.

El examen y evaluación de los participantes debe basarse en los puntajes que obtenga en una o varias pruebas, comparándolos para tomar decisiones que dependen si se está en el ámbito educativo, médico, del entrenamiento deportivo o de la condición física relacionada con la salud.

Requerimientos de las pruebas

Conocer las características de una prueba le permite al personal especializado diseñar instrumentos de medición eficientes y efectivos, y reconocer las características esenciales de las pruebas construidas por otros especialistas. Las características de una prueba se relacionan con el contenido, el participante y los procedimientos de administración, así como los principios de confiabilidad, objetividad y validez.

Las características más importantes de una prueba son la confiabilidad y objetividad de los puntajes de las pruebas y la validez de la interpretación de esos puntajes. Si una prueba no cumple con estos requerimientos no se debe aplicar, aun cuando no haya estándares rígidos del nivel de aceptación de esas características. Su aceptación está determinada tanto por la situación que se examina en sí y por otros valores que se hayan obtenido en la medición de la cualidad física en cuestión.

Aspectos relativos al contenido de las pruebas. Estas características se relacionan con la naturaleza del contenido de la prueba: qué mide y cómo lo hace.

Selección de los aspectos más importantes. Se ha recomendado que no se emplee más del 10% del tiempo de clase o de un programa en actividades de evaluación. Para alcanzar esta meta sólo se deben medir las destrezas y habilidades más importantes. Dichas destrezas y habilidades son las señaladas en los objetivos didácticos de la unidad respectiva o las que conforman los componentes claves de cualquier tipo de programa, de condición física relacionada con las destrezas o de condición física relacionada con la salud.

Discriminación. Una prueba debe discriminar los diferentes niveles de habilidad dentro del rango total de habilidad. Por lo general se espera que haya muchos puntajes diferentes en una prueba y que su distribución sea normal, o al menos que no estén marcadamente sesgados. También es importante seleccionar una prueba lo suficientemente difícil para que nadie reciba un puntaje perfecto pero lo suficientemente fácil para que nadie reciba un cero. Consideremos el problema de dos sujetos que obtienen un puntaje mínimo o máximo. Aunque dos individuos que obtienen un cero en una prueba de flexiones de codo son débiles, lo más probable es que ellos no sean idénticos en fuerza por kilogramo de peso corporal. En condiciones clínicas o de investigación donde se requiera que los sujetos eleven un peso tantas veces como sea posible, se debe seleccionar un peso que cualquiera pueda elevar al menos una vez pero que nadie pueda elevar una gran cantidad de veces, quizá no más de 30 repeticiones, por ejemplo. Recuérdese, sin embargo, el hecho de que nadie obtenga un puntaje perfecto o un cero no es garantía que una prueba discrimina satisfactoriamente e inversamente, el hecho de que alguien reciba un puntaje perfecto o un cero no indica que la prueba sea mala.

Semejanza con la actividad. Al participar en una prueba, especialmente en una prueba de destreza deportiva, se les exige a los participantes que demuestren una técnica deportiva apropiada siguiendo las reglas de la disciplina deportiva y ejecutando los movimientos de acuerdo con sus demandas. Por ejemplo, si en una prueba de baloncesto se permite correr con el balón en la mano y no se exige la utilización del drible, no se siguen las reglas del deporte ni el desempeño se adecúa a sus características. La validez de los datos de una prueba y su interpretación es cuestionable si la prueba no es semejante al tipo de actividad que se pretende evaluar.

Especificidad. Cuando una prueba mide una sola cualidad física, es posible determinar fácilmente el por qué una persona la está ejecutando bien o mal; cuando una prueba mide una cualidad que tiene varios componentes es más difícil determinar por qué un sujeto lo está ejecutando en un determinado nivel.

Por ejemplo, veamos una prueba de lanzamiento estacionario en baloncesto en 10 oportunidades. Si un estudiante lanza desde un metro de distancia del aro y fracasa en los 10 intentos, es fácil determinar que no lanza bien. Pero si lanza desde 12 metros de distancia y falla en los 10 intentos, es difícil determinar si su fracaso depende de una mala técnica de lanzamiento o de la falta de fuerza. Por ello, las pruebas que pretenden medir cualidades como la flexibilidad y la fuerza deben ser tan específicas como sea posible.

En muchas oportunidades no es viable medir una sola cualidad y en otras ocasiones el objetivo de la prueba es medir cuán bien una persona combina varias cualidades. En

todo caso, una prueba debe ser tan específica como sea posible de acuerdo con lo que se desea medir.

Pruebas con baja correlación. Cuando una cualidad física o una destreza deportiva tengan varios componentes, se implementa una batería conformada por varias pruebas. Las pruebas que se incluyan en una batería no pueden estar relacionadas, es decir, la correlación entre las pruebas debe ser baja, lo que ahorra tiempo en la realización de las pruebas favoreciendo a los sujetos que están participando. Por supuesto, todas las pruebas de una batería deben presentar una correlación alta con respecto al criterio utilizado para determinar la validez de las mismas.

Cuando dos pruebas están altamente relacionadas, es muy probable que midan la misma habilidad. Si se aplican ambas se incrementa el tiempo de las pruebas, además de dársele un mayor peso dentro de la batería de pruebas lo cual desfavorece a los sujetos con debilidad en esa habilidad. Si dos pruebas dentro de una batería presentan una correlación alta, se mantiene la que se considere más apropiada y se desecha la otra.

Aspectos relativos a los participantes. Se refiere a las características de los sujetos que influyen en el proceso:

Adecuación a los participantes. El desempeño es influenciado por la madurez, género y experiencia de las personas. Por ejemplo, los estudiantes más avanzados generalmente obtienen un mejor puntaje en las pruebas de fuerza que los más jóvenes. Los puntajes de una prueba de fuerza por lo general son confiables y su interpretación es válida según la edad y el género, lo que orienta la conformación de los estándares de desempeño. Las pruebas de destrezas deportivas, por otra parte, no son universalmente aplicables: deben implementarse según la edad, el género, fuerza, nivel de destreza, y otras capacidades. La confiabilidad y validez asociadas a las destrezas deportivas que son apropiadas para un nivel escolar, no son aceptables para otro nivel, y las apropiadas para las hembras no son aceptables para los hombres.

El desempeño también es influenciado por la edad y las discapacidades de la persona. Las pruebas de condición física para estudiantes por lo general no son apropiadas para personas mayores de 30 años y mucho menos para los de 60 o más, o para niños en edad preescolar. Los individuos con discapacidades no llegan a puntajes tan buenos como los individuos sin discapacidades en la mayoría de las pruebas de condición física por lo que las pruebas para estos grupos deben ser diferentes.

Puntajes individuales. Los puntajes de una persona en una prueba no deben ser afectados por el desempeño de otras personas. Esto es, en la organización de una prueba no se debe permitir que los individuos interactúen y que el resultado obtenido dependa de esa interacción. Por ejemplo, si en una prueba de lanzamiento en bandeja un estudiante corre hacia el tablero y otro estudiante le pasa el balón para que lance, el mejor jugador fallará si el pase es deficiente.

Aprecio. Cuando los individuos aceptan participar en una prueba y entienden el por qué son evaluados, se sienten motivados a hacer lo mejor posible y sus puntajes por lo general

demuestran su máxima capacidad. Para que la prueba sea verdaderamente apreciada, debe ser interesante y retadora, dentro de lo razonable. Lo más probable es que la gente aprecie una prueba si tiene la oportunidad de lograr un puntaje aceptable. El confort dentro de la prueba también es apreciado. Dado que algunas pruebas de capacidad aeróbica y otras pruebas de esfuerzo máximo no son muy agradables, deben evitarse pruebas fuertes que sólo unas pocas personas puedan hacer bien.

Seguridad. Por supuesto, no deben utilizarse pruebas que constituyan un peligro para los participantes. Hay que revisar cada prueba para determinar si los sujetos podrían sobreexponerse o cometer faltas que les pueda causar lesiones. Debe asegurarse el área con los equipos apropiados y utilizar implementos que no tengan probabilidad de romperse representando un potencial riesgo para los participantes.

Confidencialidad y privacidad. Muchos participantes en pruebas de condición física o en programas de rehabilitación prefieren que los demás no conozcan cuán bien o mal fueron sus puntajes en las pruebas. Los estudiantes podrían sentirse desconcertados si obtienen un puntaje considerablemente más alto o más bajo que sus compañeros, y los participantes en programas de condición física o de rehabilitación pueden tener sentimientos similares o simplemente pensar que sus puntajes no tienen que ser conocidos por los demás. Se requiere que todos aquellos que dirijan programas de medición acepten este principio. Evaluar a una persona a la vez antes que en grupo es la mejor manera de cumplir con esa exigencia.

Motivación para alcanzar el máximo potencial. Los estudiantes hacen un gran esfuerzo en las pruebas cuando su clasificación puede ser afectada. Por lo general, la gente que está en programas de certificación o cuyos trabajos, salarios o premios dependan de sus puntajes en las pruebas, hacen el máximo esfuerzo para salir bien. Por ello, este tipo de evaluación puede motivar a la gente a practicar para lograr el más alto puntaje posible.

Pero, ¿qué motiva a la gente a hacer el mejor esfuerzo si no se halla en las situaciones anteriores? Un esfuerzo muy leve por parte de las personas que están siendo evaluadas arroja puntajes que no son confiables y la interpretación que se base en esos puntajes no es válida. El examinador debe motivar a los participantes a demostrar lo mejor y constantemente observa la posible falta de esfuerzo por parte de los participantes. Los especialistas en ejercicio que efectúan pruebas en programas de terapia necesitan ser particularmente conscientes del esfuerzo de la gente.

Aspectos administrativos

Pruebas para grupos numerosos. Cuando hay que evaluar un gran número de personas en un corto período de tiempo, las pruebas deben ser adecuadas para que todos los participantes rindan apropiadamente. Mientras más tiempo dure una prueba, menor en el número de pruebas que pueden aplicarse. Cuando se miden grupos numerosos es importante que la gente sea medida rápidamente, en forma sucesiva o simultáneamente. Una prueba es apropiada para grandes grupos cuando un sujeto tarda ejecutándola entre 10 y 15 segundos. Por ejemplo, una prueba de abdominales es apropiada para grandes grupos cuando la

mitad del grupo es evaluado y la otra mitad sirve de ayuda. Asimismo, las pruebas cortas mantienen a los participantes activos y ayudan a prevenir los problemas disciplinarios que puedan surgir debido a la inactividad; también reduce la posible insatisfacción si hay mucha gente en el caso de que sea un programa de investigación o de mantenimiento de la condición física.

Tanto el docente como el investigador y el entrenador deben preocuparse por la confiabilidad de los datos, así como por la validez de la interpretación de los datos recolectados. Con una minuciosa planificación y reflexión no se producirán problemas en estos temas.

Oportunidades de práctica. La gente debe estar familiarizada con las pruebas y permitírseles que practiquen antes de su ejecución. La familiarización inmediatamente antes de la prueba, o durante el desarrollo del programa o en clase, o en sesiones prácticas previas al día de la evaluación, disminuye el tiempo de práctica y explicación el día de la prueba. Aun cuando una prueba sea desconocida, si es fácil de comprender se requiere de poco tiempo para explicarla. Evite las pruebas que requieran instrucciones muy complejas o un considerable tiempo de práctica.

Cuando la evaluación se produzca en programas de investigación, rehabilitación o en pruebas de fuerza, donde los participantes son evaluados con cintas rodantes, bicicletas u otros equipos sin que tengan experiencia previa en ellos y con una mínima explicación y con pocos intentos de práctica en la prueba, ¿son confiables los puntajes obtenidos? Un equipo puede tener el potencial de suministrar datos muy confiables y la interpretación de ellos es muy válida, pero si no se utilizan las técnicas apropiadas de administración de una prueba al evaluar con ese equipo, la confiabilidad y validez de la interpretación de los datos no es aceptable.

Personal y equipamiento. Para muchos docentes, las pruebas que requieran de poco equipamiento y personal de aplicación no son satisfactorias. Pero la realidad dice que los equipos generalmente son costosos tanto para adquirirlos como para darles mantenimiento, y también puede llevar tiempo ensamblarlos. De la misma manera, cuando se necesita personal para administrar una prueba debe emplearse tiempo en encontrarla y entrenarla. Aun si se planifica que los miembros de un curso o de un programa sean los aplicadores de una prueba, se debe emplear tiempo en entrenarlos.

La falta de entrenamiento de los aplicadores de las pruebas se traduce en falta de objetividad de los puntajes. En los laboratorios y programas de rehabilitación en los que los participantes son examinados en intervalos regulares por un período prolongado para determinar si están mejorando, la misma persona o aplicador debe evaluar al participante cada vez para maximizar las oportunidades de que el mejoramiento del participante sea realmente un mejoramiento y no un producto de la falta de objetividad.

Preparación de los materiales. Seleccionar pruebas con materiales que sean fáciles de colocar e iniciar es preferible a pruebas donde se lleve demasiado tiempo cumplir con esas tareas, lo cual hace que el trabajo de evaluación se inicia con buen pie. Las pruebas que utilizan equipos complejos o varias partes del equipo colocado en lugares diferentes, o

que requieran demarcaciones de límites y dimensiones en el piso o en paredes, generalmente se inician con mucho retraso si no se tiene el personal adiestrado para hacerlo rápidamente.

Instrucciones adecuadas. Un importante requerimiento en el diseño de una prueba son las instrucciones. Al utilizar las pruebas diseñadas por otros se deben poseer las instrucciones completas. Las instrucciones se especifican tal cómo se estableció en la prueba, así como la preparación de los sujetos y los procedimientos de administración y anotación.

Las instrucciones en muchas pruebas no orientan adecuadamente en el caso de que una persona que está siendo medida se desvíe de la forma establecida y aun así se le conceda el puntaje en la prueba. Por ejemplo, las instrucciones en la prueba de flexiones de codo en la barra especifican que el sujeto debe elevar el cuerpo hasta que el pecho toque la barra y luego descender hasta que los codos estén extendidos, y así tantas veces como sea posible. Solo se anotan las ejecuciones correctas. Se supone que no se cuenta si la barra casi toca el pecho o si los codos están casi extendidos. Sin embargo, un aplicador contará una repetición ligeramente incorrecta porque la forma o técnica es básicamente correcta, mientras que otro aplicador no contará la repetición porque no es perfectamente correcta. O, un día el aplicador contará una repetición ligeramente incorrecta, y otro día no la hará. La falta de claras y suficientes instrucciones de cómo anotar una ejecución ligeramente incorrecta contribuye a la disminución de la objetividad y confiabilidad de la prueba.

Normas. Si una prueba y las normas correspondientes son recientes y apropiadas, se ahorra el tiempo que se necesita para desarrollar las normas locales, o al menos ofrecen estándares temporales hasta que las normas locales se desarrollen. Desafortunadamente, las normas que se originan con base a una prueba pueden ser tan antiguas que ya no son apropiadas, o pueden estar basadas en edades, género o experiencia diferente, por lo que no son las más adecuadas para la evaluación de los sujetos.

Puntajes útiles. Una prueba debe producir puntajes útiles. Estos son puntajes que se utilizan para procesar fórmulas con poco esfuerzo. Muchas mediciones físicas pueden utilizarse inmediatamente después de la medición y si se requiere que los puntajes sean introducidos a una fórmula antes de que sean divulgados y empleados por los interesados, entonces la fórmula debe ser lo suficientemente simple para que los cálculos se hagan rápidamente.

Por ejemplo, con la fórmula $Y = 2X + 5$ donde Y es el puntaje calculado y X es el puntaje recolectado, se observa que es una simple fórmula que se puede procesar sin complicaciones. Ahora, si la prueba requiere que se calcule $Y = 0,4521X_1 + 0,3334X_2 + 1,2$ llevaría mucho tiempo realizar los cálculos a mano para entregárselos a cada participante. Apoyándose en una computadora se puede incluir con antelación la fórmula para hallar Y introduciéndose sólo los puntajes X en el programa de procesamiento de datos que se utilice.

Cuando se analizan estadísticamente un conjunto de puntajes, con una calculadora o con una computadora, los puntajes deben expresarse en la más simple unidad de medida.

Al medir la estatura de las personas los datos se recolectan en metros y centímetros, convirtiendo luego los metros a centímetros para procesar los datos en esta última unidad de medida. Los datos se recolectan en metros y centímetros porque así lo expresa el equipo y la gente está habituada a ese tipo de expresión. Este concepto es muy útil en pruebas de salto y de lanzamiento, donde por lo general se recolecta la información en metros y centímetros, por lo que la conversión se debe hacer cuando se vayan a introducir los datos.

Administración

La clave de una buena evaluación es la planificación que se realiza antes de la aplicación de la prueba, y luego seguir esa planificación durante y después de la prueba.

Procedimientos anteriores a la aplicación de la prueba (Pretest). Se planifica antes de la aplicación de la prueba para estar seguro de que la preparación es la apropiada y que la aplicación fluirá sin contratiempo. La planificación previa a la prueba es la preparación de todo lo que ocurra antes de su aplicación. Esto implica cierto número de tareas: conocer la prueba, desarrollar los procedimientos, desarrollar las instrucciones, preparar los sujetos que van a ser evaluados, prever el calentamiento y los intentos de práctica, preparación de los materiales de la prueba incluyendo los equipos de seguridad, elaborar las hojas de anotación, estimar el tiempo que se necesita y, finalmente, aplicar la prueba.

Conocimiento de la prueba. Cuando se planifica la administración de una prueba por primera vez, lea el manual de instrucciones varias veces pensando en la prueba en la medida que vaya leyendo. Esta es la única manera de evitar que se le pasen pequeños detalles sobre procedimientos, dimensiones, materiales, etc.

Desarrollo de los procedimientos. Una vez que se familiarice con la prueba, se inicia el desarrollo de los procedimientos para su administración. Se deben seleccionar los procedimientos más eficientes decidiendo si se evalúa a todos los sujetos a la vez o en grupos y determinando si se hace en forma individual o en pareja cuyos miembros se miden uno al otro. Si se planifica la aplicación de varias pruebas el mismo día, deben ordenarse de manera tal que la fatiga sea minimizada, sin organizar pruebas consecutivas donde se empleen los mismos grupos musculares. Asimismo, si se incluyen eventos muy fatigantes, tal como las carreras de distancia o de máximo esfuerzo, se colocan de últimas.

El próximo paso es identificar la unidad de medición y la forma de anotación. Por ejemplo, en una prueba de abdominales que requiera que el sujeto comience en posición supina con las escápulas en el piso, las rodillas flexionadas y los brazos cruzados con las manos colocadas en el hombro contrario, para que luego eleve el tronco hasta que los codos toquen los muslos y retorne a la posición inicial, se anota como un abdominal, es decir, una repetición. Cuando sea necesario debe seleccionarse la unidad de medición con la que se anotará el resultado; en el caso de las pruebas de distancia determinar si se anotará en metros y centímetros o en centímetros, y en las pruebas de tiempo en segundos y centésimas de segundo o en centésimas de segundo. Para obtener un puntaje que pueda

ser analizado matemáticamente de inmediato, debe utilizarse sólo una unidad de medición que por lo general es la más pequeña.

También debe decidirse qué hacer si un sujeto comete un error durante la prueba. Anticipándose a todas las situaciones y posibles sentencias, se tratará con mayor equidad a los sujetos. Por ejemplo, ¿Qué hacer si un individuo no llega al piso en un abdominal? Si se hace caso omiso del error y se le advierte al sujeto y se cuenta un abdominal, se debe tomar la misma medida con todos los participantes.

Finalmente, es muy importante prever medidas de seguridad. Planifique siempre la colocación de observadores en las pruebas donde la gente pueda lesionarse o cuando se evalúe a personas mayores. Considere la colocación de dispositivos para la demarcación también. En una carrera de obstáculos los aparatos deben ser blandos, irrompibles y lo suficientemente altos como para que los participantes no puedan pasar por encima de ellos. Se deben utilizar conos en lugar de sillas o botellas de refresco para delimitar el área de evaluación o el curso de la carrera de obstáculos. Prevéase también el área de evaluación. Organícense las pruebas de carrera en un área lo suficientemente amplia como para los participantes no corran con dificultad, así como a los demás participantes que esperan para pasar a la prueba de manera tal que no se corra entre ellos.

Si nunca se ha administrado una prueba, tratar de practicarla una o dos veces antes de aplicarla. Esta es una buena manera no sólo de determinar posibles cambios en los procedimientos sino también de entrenar a los aplicadores.

Desarrollo de las instrucciones. Luego de desarrollar los procedimientos, se desarrollan las instrucciones exactas. Es perfectamente aceptable leer las instrucciones al grupo antes de aplicarles la prueba. Las instrucciones deben ser fáciles de comprender con especificaciones sobre:

- Procedimientos de aplicación
- Instrucciones sobre la ejecución
- Procedimientos de anotación y medidas sobre una ejecución incorrecta
- Sugerencias sobre la técnica para mejorar el resultado.

Preparación de los participantes. Debe anunciarse la prueba con suficiente antelación para que la gente pueda practicar si ellos piensan obtener los mejores puntajes. Cuando un grupo no está familiarizado con la prueba se emplea cierto tiempo antes del día de la prueba para explicarla, proporcionar la forma de mejorar el puntaje y supervisar las prácticas previas.

Aun cuando la gente tenga experiencia en una prueba, pueden necesitar de cierto tiempo para recordar la forma de ejecución. Ellos deben conocer con anticipación que van a ser evaluados, cual es la prueba, y lo que ello implica. Esto le permite a cada uno estar psicológica y fisiológicamente preparados para participar y tratar de obtener un puntaje elevado acorde con su potencial. Ello es vital cuando se hallan en juego asuntos importantes como la clasificación y ubicación en un curso, la admisión a una institución, la

aprobación de un programa o la determinación de la condición física para la salud o las destrezas.

Planificación del calentamiento y los intentos de práctica. La confiabilidad de una prueba mejora con un calentamiento previo. La cantidad y naturaleza del calentamiento debe planificarse. Idealmente, el calentamiento debe ser específico a la destreza que se va a evaluar. Se ha demostrado que un calentamiento supervisado en el cual un aplicador le dice a los sujetos qué hacer y asegura que cada uno reciba el mismo tipo de práctica, es mejor que un calentamiento sin supervisión. En algunas ocasiones es aceptable que el aplicador permita una gran cantidad de intentos y considerar luego los primeros intentos como el calentamiento, y dándole un puntaje a cada individuo para su preparación y conocimiento de acuerdo con la media de sus últimos intentos.

Preparación de los materiales de la prueba y equipos de seguridad. Se debe tener a mano todo el equipamiento antes del día de la prueba. Los equipos deben ser chequeados para asegurarse que funcionan y se conozca cómo utilizarlos. Igualmente, antes de la prueba, todos los límites y demás marcaciones deben ser ubicados. Teniendo todo el equipo disponible, las líneas de demarcación y otras demarcaciones colocadas correctamente cuando la gente que va a ser evaluada llegue, ahorra tiempo y evita los problemas que surgen cuando la gente tiene que esperar mientras todo esté listo.

Elaboración de las hojas de anotación. Mucho antes de la prueba se debe localizar o preparar la hoja de anotación general para introducir los puntajes, o tarjetas individuales. Los nombres de los sujetos se anotan antes de la prueba.

La implementación de la ficha individual es más ventajosa que la hoja general. La ficha individual permite que una persona se desplace entre las estaciones de evaluación y registre rápidamente el puntaje cuando está siendo medida por un compañero. Cuando un aplicador está midiendo y registrando los puntajes de un grupo puede ahorrar tiempo al reunir las fichas individuales en el orden en que están alineados para ser medidos, mucho más que organizando a los individuos en una línea tal como aparecen en la hoja general.

La alternativa entre usar la hoja general de anotación o la ficha individual está, cuando se mide la gente, en introducir los datos en un programa informático para su inmediato análisis. Al igual que en la hoja general, se ingresan los datos personales en el programa antes de la medición.

Estimación del tiempo de medición. Cuando una prueba no se aplica a una sección completa o a toda la gente en un programa, o cuando la mitad del grupo se mide un día y la otra mitad al siguiente, se deben planificar actividades para llenar el tiempo libre y mantener ocupado al grupo restante. La estimación del tiempo de medición para la aplicación de una prueba o batería de pruebas en todos los casos minimiza las situaciones confusas que surjan y maximiza el uso del tiempo disponible.

Aplicación de la prueba. Si se ha planificado cuidadosamente, la medición debe realizarse en forma fluida. Aunque la principal preocupación el día de la medición es la admi-

nistración de la prueba o pruebas en sí, también debe haber preocupación por la preparación, motivación y seguridad de los participantes. Si después de la aplicación de las pruebas los aplicadores pueden decir, “los participantes estaban preparados y la prueba fue administrada de la manera que nos gustaría que nos evaluaran”, las mediciones indudablemente fueron un éxito.

Preparación. Los participantes ya deben conocer cuál es la prueba y porqué van a ser medidos; luego se procede con el calentamiento y la práctica. A continuación, se dan las instrucciones y se explican los procedimientos de la prueba, y de ser necesario se hace la demostración. Se responden las preguntas antes y después de la demostración. Cuando la destreza que se va a medir o los procedimientos son muy complicados, se permite que los participantes hagan nuevas prácticas de la prueba.

Motivación. Todos los participantes deben recibir la misma cantidad de estímulos y motivación. Aunque existe la tendencia a estimular a los ejecutantes deficientes y elogiar a los mejores, lo justo es que el comentario se haga para todos o para ninguno. Siempre que sea posible, al participante se le indica el puntaje inmediatamente después del intento. Esto puede motivarlos a superarse en el segundo o tercer intento, teniendo cuidado de no avergonzarlos cuando el intento sea verdaderamente deficiente.

Seguridad. Durante una sesión de medición, se observa si surgen problemas de seguridad. Los sujetos ejecutan la prueba bajo condiciones de inseguridad si no siguen las indicaciones respectivas. Hay que anticiparse a estas y otras situaciones de inseguridad.

Procedimientos posteriores a la aplicación de las pruebas (Posttest). La razón fundamental de la medición es recolectar información sobre los participantes. Sólo después de aplicarse las pruebas, la información puede ser utilizada. De modo sorprendente, muchos docentes y entrenadores no hacen el suficiente esfuerzo e incluso, cualquier esfuerzo, para realizar un análisis posttest. Mientras más rápido se les den los resultados de las pruebas a los participantes, más significativo será el proceso de evaluación. Los procedimientos posteriores a la aplicación de las pruebas incluyen el registro, análisis y reporte de los resultados.

Análisis de los puntajes de las pruebas. Inmediatamente después de las pruebas, los puntajes se analizan de acuerdo con las técnicas estadísticas apropiadas. Ello requiere la introducción de los datos a la computadora para el análisis, registro y recuperación de los datos. El análisis se utiliza para revelar los elementos que podrían influir en el proceso de enseñanza aprendizaje o en la conducción de un programa de otra índole, así como suministrarle información al grupo y preparar los datos para cumplir con los propósitos de la evaluación. Las personas por lo general se interesan en sus puntajes, su ubicación con respecto al grupo, y el grado de mejoramiento o declinación desde la última prueba realizada. Informarles los resultados a los participantes es un efectivo mecanismo motivacional.

Registro de los resultados de las pruebas. El registro de los resultados de las pruebas usualmente no es más que la introducción de los puntajes en una hoja de cálculo en el

archivo correspondiente y el respectivo análisis. Esa información hace posible las comparaciones entre clases o grupos y dentro y entre años, la evaluación del programa a través de los años, y el desarrollo de las normas. Nótese que estas normas son estándares del grupo o de la clase y no son estándares individuales, normas basadas únicamente en los puntajes y su análisis. Puede ser innecesario hacer públicos los puntajes, especialmente si se lleva una ficha de registro permanente para cada persona. Es a partir de esta ficha que se le hace un seguimiento de grado a grado o de año a año, para que se pueda indicar el mejoramiento o rendimiento individual a través del tiempo.

Evaluación

Luego de administrarse las pruebas y obtenerse los puntajes, es decir, después de la medición, se sigue con el proceso de evaluación que toma la forma de un juicio sobre la calidad de un desempeño. Supóngase, por ejemplo, que cada participante corrió una milla y que sus puntajes fueron registrados. No importa si el evaluador clasifica las mediciones como excelente, buena, promedio, o como A, B, C; simplemente se está haciendo una evaluación.

La evaluación puede ser subjetiva: un juez no utiliza los estándares establecidos para cada clasificación pudiendo evaluar durante el desempeño sin necesidad de registrar las mediciones. La objetividad de la evaluación se incrementa cuando está basada en estándares definidos.

Tipos de evaluación. La evaluación formativa ocurre en el transcurso del proceso instruccional o del programa de condición física, para informarle tanto al participante como al docente o entrenador del estatus del sujeto. Esta información permite juzgar la efectividad de la unidad en desarrollo y planificar para el futuro. Los participantes están interesados por conocer en qué medida están confluyendo hacia los objetivos establecidos. Así pues, la evaluación formativa continua durante toda la unidad o programa de enseñanza y se halla totalmente vinculada con los objetivos previstos.

La evaluación sumativa comprende el final de la unidad o programa de enseñanza, así como de la medición última del desempeño del participante, lo que también lleva a la clasificación definitiva. Este tipo de evaluación implica la comparación entre los estudiantes, más que la comparación del nivel de logro. En un programa de entrenamiento se compara el puntaje de los participantes con respecto a la meta de desempeño establecida al inicio o a un estándar ideal, pero la evaluación puede llevar a que se le permita al participante retirarse del programa o a que se le invite a continuar en él.

Estándares de evaluación

Estándares referidos a criterios. Los estándares referidos a criterios representan el nivel de logro que casi todos los participantes deben ser capaces de alcanzar en concordancia con una apropiada instrucción y con suficiente práctica. Los estándares son valiosos para los participantes porque especifican el nivel esperado de desempeño, y valiosos para el

docente o entrenador porque definen claramente el estatus del participante de acuerdo con el estándar.

Los estándares referidos a criterios deben relacionarse con objetivos explícitos, específicos, objetivos que normalmente pueden ser cumplidos, antes que objetivos muy amplios, generales, cuyo logro exige un mayor plazo. De esta manera los estándares referidos a criterios pueden utilizarse en la evaluación formativa para diagnosticar las debilidades y determinar cuándo los participantes están listos para progresar.

Los estándares referidos a criterios se utilizan en cursos cuyo objetivo sea la certificación. Por ejemplo, estos estándares se usan en el examen para obtener la licencia de conducir. Existe un puntaje mínimo que una persona debe alcanzar tanto en la prueba de conocimiento como en la prueba de conducción para obtener la licencia. Alcanzar al menos en puntaje mínimo representa la aprobación del examen, pero cualquier puntaje menor a ese puntaje mínimo indica el fracaso en el examen.

Los estándares referidos a criterios en una unidad de condición física para alumnos de noveno grado podían ser una carrera de milla y media en 12 minutos, 35 abdominales en dos minutos y 3 flexiones en suspensión. Así, estos estándares referidos a criterios tienen un cierto nivel de dificultad. En este caso, el desempeño en las tres pruebas sumando sus puntajes estándar, puede utilizarse como evaluación sumativa del objetivo general de condición física.

Estándares referidos a normas. Los estándares referidos a normas, aquellos que permiten la comparación del desempeño dentro del grupo, son útiles para determinar el grado en el cual los participantes logran el objetivo general. Para el desarrollo de estándares referidos a normas, los niveles de desempeño que distinguen la habilidad entre los grupos debe ser especificada, esto es, los estándares se establecen para que algunos participantes sean clasificados, por ejemplo, de “alta habilidad” y otros de “baja habilidad”. Donde se utiliza el sistema de clasificación A, B, C, D, E, se han basado en estándares referidos a normas.

En general, todas las personas no responden de la misma manera a las mediciones. La misma prueba, el protocolo de las pruebas o las expectativas sobre el puntaje del desempeño no son aplicables para todas las personas. Comprender las características de la población que se va a medir y tener experiencia en la medición de esa gente, es vital para obtener puntajes en los cuales fundamentar la validez de las interpretaciones.

Pruebas de campo de condición física

CARRERA DE 1 MILLA (1609 metros)

Objetivo	Medir la capacidad funcional y resistencia del sistema cardiovascular.
Equipo	Pista de atletismo o terreno perfectamente medido y acondicionado con banderolas de señalamiento, cronómetros, tablas de anotación, planillas de registro de resultados, lápices.
Procedimiento	El juez de salida instruye a los participantes para que corran la distancia en el menor tiempo posible. Deben comenzar a la señal de “listos, ya” empleando la salida alta. Se permite caminar, pero el objetivo es hacerlo rápido. El cronometrista registra el tiempo que realiza cada participante y luego se cheque el orden de llegada para anotar el tiempo correspondiente. Tan pronto como se cruce la línea final, se sale de la pista.
Anotación	Se registra el tiempo empleado en culminar el recorrido en minutos y segundos. Para procesar los datos y asignar el percentil que le corresponde a cada uno, se hace una conversión llevando el tiempo a segundos.

ABDOMINALES

Objetivo	Medir la resistencia de la musculatura abdominal, ejecutando el máximo número de repeticiones en 1 minuto.
Equipo	Cronómetro, colchoneta, tablas de anotación, planillas de registro de resultados, lápices.
Procedimiento	El sujeto asume una posición decúbito dorsal, con las rodillas flexionadas y los talones a 45 cm de los glúteos; las manos se colocan a los lados de la cabeza con los dedos tocando alrededor de las orejas, sin separarlos durante la prueba. Un ayudante sujeta los pies, evitando el levantamiento de los talones. Desde esta posición y al darse la orden de inicio por parte del cronometrista, se eleva el tronco para tocar con los codos las rodillas o los muslos lo cual constituye 1 abdominal. Se debe volver a la posición inicial estableciendo el contacto de las escápulas con la colchoneta para repetir la acción constantemente tan rápido como pueda durante 1 minuto.
Anotación	Un juez cuenta el número de repeticiones efectuadas durante el tiempo señalado y se lo indica al anotador para que registre la marca lograda.

SALTO VERTICAL

Objetivo	Medir la potencia de los miembros inferiores al saltar verticalmente hacia arriba.
Equipo	Listón de madera con una cinta métrica adosada, tiza, pared.
Procedimiento	El sujeto se coloca de pie y en forma lateral a la pared, adosado para elevar el brazo con el codo extendido marcando con tiza el punto más alto que alcanza con la yema de los dedos. Luego se separa 20 cm de la pared permaneciendo alineado lateralmente con la marca realizada, se impulsa verticalmente para marcar con los dedos el punto más alto que alcance con el salto. Efectúa dos intentos con una separación mínima de 2 minutos entre cada intento.
Anotación	Se registra en centímetros la distancia lograda en cada intento con lo cual se puede determinar las diferencias entre los participantes. Asimismo, se puede establecer la potencia del salto según la siguiente fórmula: $P = \sqrt{4.9 * p * h}$ donde p = peso en kg y h = distancia alcanzada en metros.

AGILIDAD

Objetivo	Medir la capacidad para desplazarse cambiando de dirección.
Equipo	Cinta métrica, tirro o pintura, cronómetro, dos tacos de madera de 5 * 5 * 10 cm.
Procedimiento	Se trazan dos líneas paralelas de 1.50 metros a una distancia de 9,15 metros una de otra. Debe haber un área de seguridad de, al menos, dos metros alrededor de la zona de la prueba. La salida se realiza detrás de una de las líneas. Inmediatamente después de la línea opuesta se colocan los dos tacos en forma longitudinal y separados 50 cm entre sí. El juez – cronometrista da la salida y el sujeto corre a toda velocidad a tomar un taco, regresa a colocarlo después de la línea de salida retornando inmediatamente a tomar el segundo taco para trasladarlo y colocarlo de la misma manera. Se cronometra hasta que coloque el segundo taco en el piso sin que toque la línea ni lo suelte. Se permiten dos intentos con una diferencia mínima de 5 minutos de descanso entre cada intento.
Anotación	Se mide el tiempo empleado en segundos y centésimas de segundo, tomándose el mejor de los dos intentos para los cálculos correspondientes.

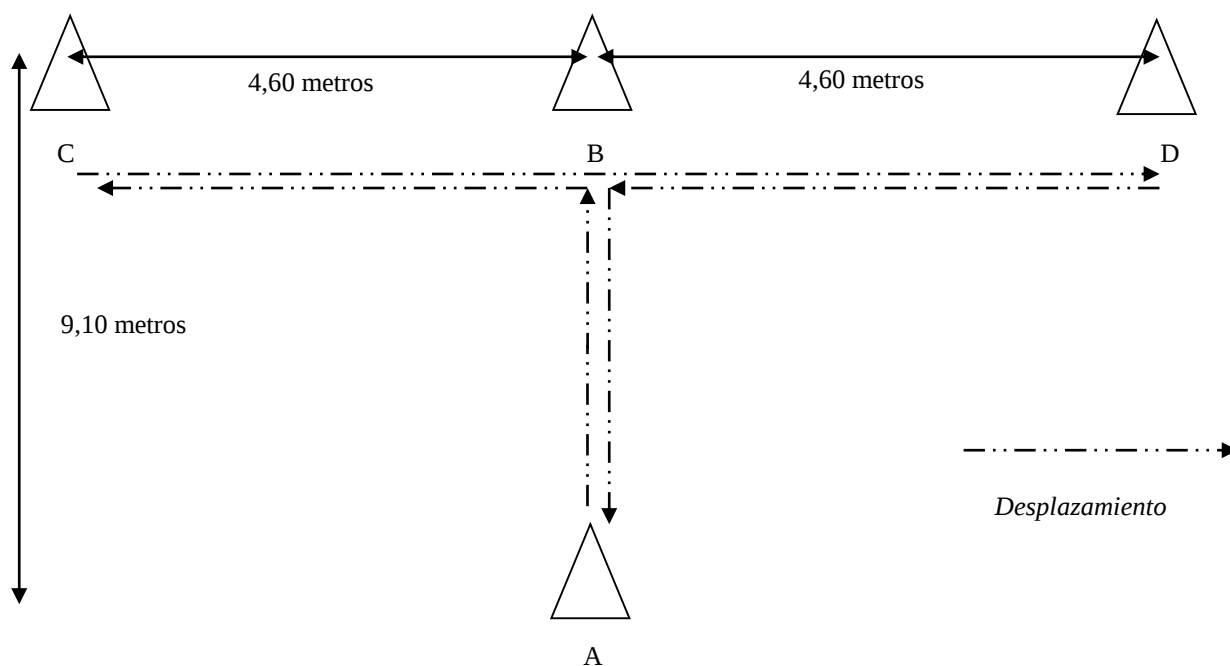
IMPULSO DE BALÓN

Objetivo	Medir la potencia de los miembros superiores al lanzar un objeto.
Equipo	Cinta métrica, balón de baloncesto.
Procedimiento	El sujeto se sienta con la espalda a la pared, con el tronco erguido y las piernas extendidas y separadas. Se lanza el balón impulsándolo desde el pecho con ambas manos, tratando de alcanzar la mayor distancia posible sin despegarse de la pared. Se efectúan dos intentos.
Anotación	Se mide la distancia alcanzada en metros y centímetros, tomándose el mejor de los dos intentos para los cálculos correspondientes.

TEST T

Objetivo	Medir la velocidad, agilidad y el control corporal al desplazarse cambiando de dirección.
Equipo	4 conos, cronómetro, cinta métrica.
Procedimiento	Se coloca un cono (A) como punto de partida, otro cono (B) al frente del cono de salida a 9,10 metros; a partir del segundo cono se colocan dos conos (C – D), uno hacia cada lado, a 4,60 metros. El sujeto se desplaza corriendo hacia adelante y toca con la mano derecha la base del cono b. De allí desliza lateralmente hacia la izquierda y toca con la mano izquierda la base del cono c, desliza luego lateralmente hacia la derecha hasta el cono d y toca con la mano derecha la base del cono. A continuación, desliza de la misma manera hasta el cono b tocando con la mano izquierda su base. Sigue con carrera hacia atrás hacia el primer cono donde se detiene el cronómetro. Se efectúan dos intentos, con un intervalo entre uno y dos minutos. Se descalifica al sujeto si no toca las bases de los conos, cruza los pies durante el deslizamiento, o corre de manera diferente al desplazamiento indicado.
Anotación	Se registra el tiempo en segundos y centésimas de segundo, y se toma el mejor de los dos intentos para el procesamiento de los datos.

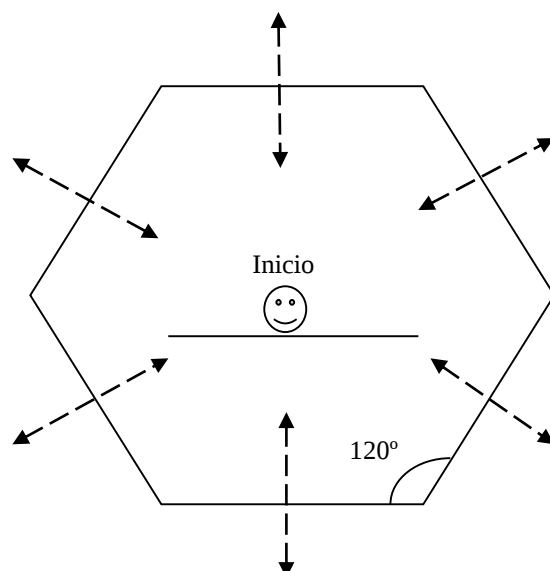
Test T:



TEST HEXAGONAL (SALTO RASANTE BIPODAL)

Objetivo	Valorar la capacidad para producir potencia, velocidad, equilibrio, agilidad y control corporal por medio de saltos rasantes bipodales.
Equipo	Superficie plana no deslizante, tirro, cronómetro.
Procedimiento	Se traza un hexágono con líneas de tirro de 60 cm de largo y una pulgada de ancho, que formen ángulos de 120° en sus uniones. El sujeto se coloca en el centro del hexágono y salta en forma bipodal adentro – afuera – adentro sin tocar las líneas ni dar pasos adicionales. Se salta hacia cada lado del hexágono y se regresa siempre de frente al punto de partida y en el sentido de las agujas del reloj, completando el trazado en tres oportunidades, es decir, saltando un total de 36 veces. No es válido el intento si toca las líneas, se da un paso adicional o se pierde el equilibrio. Entre cada intento se concede un descanso de uno o dos minutos.
Anotación	Se toma el tiempo en cada intento, se registran y luego se promedian para obtener una valoración más precisa de la capacidad del sujeto.

Test hexagonal:



PRUEBA DE LANZAMIENTO HACIA ATRÁS

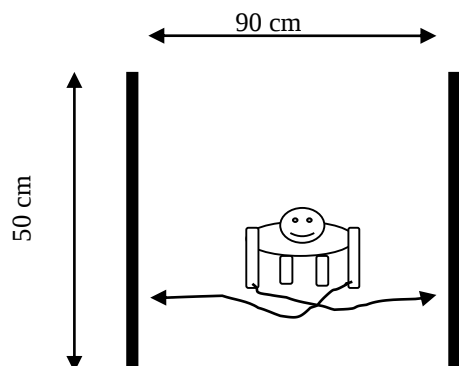
Objetivo	Determinar la fuerza y potencia de los miembros superiores.
Equipo	Balón medicinal de 1.5 kg (f) y 2 kg (m), tirro, cinta métrica.
Procedimiento	En un extremo del área de lanzamiento se traza una línea de 50 cm como punto de lanzamiento. El sujeto se coloca de espalda hacia el área de lanzamiento con los pies separados a la misma anchura de los hombros, sin que los talones toquen la línea. Se sostiene el balón adelante sin flexionar los codos, a la altura de los hombros. Se permite que se haga un contramovimiento a través de un movimiento de flexión de tronco y rodillas para llevar el balón abajo hasta la altura de las rodillas, se eleva el balón al frente y se repite el contramovimiento para tomar impulso, lanzando el balón hacia atrás por arriba de la cabeza con los codos extendidos tratando de alcanzar la mayor distancia posible. Al lanzar el balón no se debe separar los pies del piso ni dar un paso para guardar el equilibrio.
Anotación	Se mide la distancia alcanzada en metros y centímetros. Se permiten dos intentos y se toma en cuenta el mejor lanzamiento para el procesamiento de los datos.

Lanzamiento hacia atrás



TEST DE EXTREMIDADES SUPERIORES

Objetivo	Medir la fuerza, resistencia y la estabilidad cinética continua de las extremidades superiores
Equipo	Tirro, cronómetro.
Procedimiento	Se colocan dos líneas de 50 cm de largo paralelas y separadas entre sí a 90 cm. El sujeto se coloca en apoyo facial con las manos entre los dos trozos de tirro; los varones con apoyo de la punta de los pies y las hembras apoyadas en las rodillas. Se toca con la palma de la mano derecha la línea de la izquierda retornando al apoyo, luego toca con la mano izquierda la línea de la derecha y regresa al apoyo, en un movimiento continuo de lado a lado durante 15 segundos.
Anotación	Se registra el número de toques realizados durante el tiempo de ejecución. Se puede calcular la potencia multiplicando el 68% del peso corporal en libras por el número de toques realizados y dividirlos entre 15 segundos: Potencia = (68% del peso en libras * # de toques) / 15



Capítulo 11. Desarrollo de pruebas de destreza deportiva

Los docentes, entrenadores e investigadores del área de la actividad física deben interesarse en la implementación de baterías de pruebas para alguna disciplina deportiva o determinadas cualidades de la condición física. En el análisis del rendimiento deportivo, las pruebas de destreza deportiva y de desempeño motor se utilizan en conjunción con la medición de otras cualidades físicas a fin de darle información adicional acerca de la situación actual y futuro de los atletas. A continuación, se presenta un flujograma con 10 pasos para el diseño de pruebas de esta naturaleza que también orienta el diseño de pruebas de habilidad motora.

Paso 1. Estudiar los criterios exigidos a una buena prueba. Básicamente, este criterio se refiere a los elementos estadísticos de fiabilidad, validez y objetividad. El evaluador debe estar familiarizado con los equipos, el personal, el espacio y el tiempo disponible para las mediciones. Es imperativo que la prueba se adecúe a la edad y género de los estudiantes y estar íntimamente relacionada con la destreza que se pretende medir. Aquí se deben tomar en cuenta los factores de seguridad implícitos en la prueba.

Paso 2. Analizar la disciplina deportiva para determinar las destrezas o habilidades que se medirán. Si se piensa evaluar la situación actual del sujeto o del grupo las pruebas de destrezas deportivas son más apropiadas, pero si se trata de determinar el potencial deportivo, las pruebas de habilidad motora son las requeridas. Se podrían combinar estas baterías si se considera que es útil de acuerdo con los objetivos trazados.

Paso 3. Revisar la literatura. Una vez que se haya analizado el aspecto sobresaliente de la disciplina deportiva, se hace una revisión de la literatura asociada con las baterías de pruebas para destrezas o áreas de desempeño motor específicas sin olvidarse de consultar los expertos.

Paso 4. Seleccionar o diseñar las pruebas. Asegurarse que las pruebas son: (a) Representativas del desempeño que se pretende analizar; (b) Aplicadas con relativa facilidad; (c) Estrechamente asociadas con el desempeño real de los sujetos; (d) De importancia verdadera. También es importante que cada prueba esencialmente mida cualidades diferentes. No habrá un uso eficiente del tiempo si se aplican pruebas que midan la misma destreza o habilidad motora.

Paso 5. Determinar los procedimientos de medición. Esto incluye la determinación del número de intentos mínimos para la medición, la forma de establecer el puntaje que se procesará, y el orden de las pruebas.

Paso 6. Escrutinio de los expertos. Hacer que los expertos como otros docentes o entrenadores que se hallen familiarizados con la actividad, examinen la batería de pruebas.

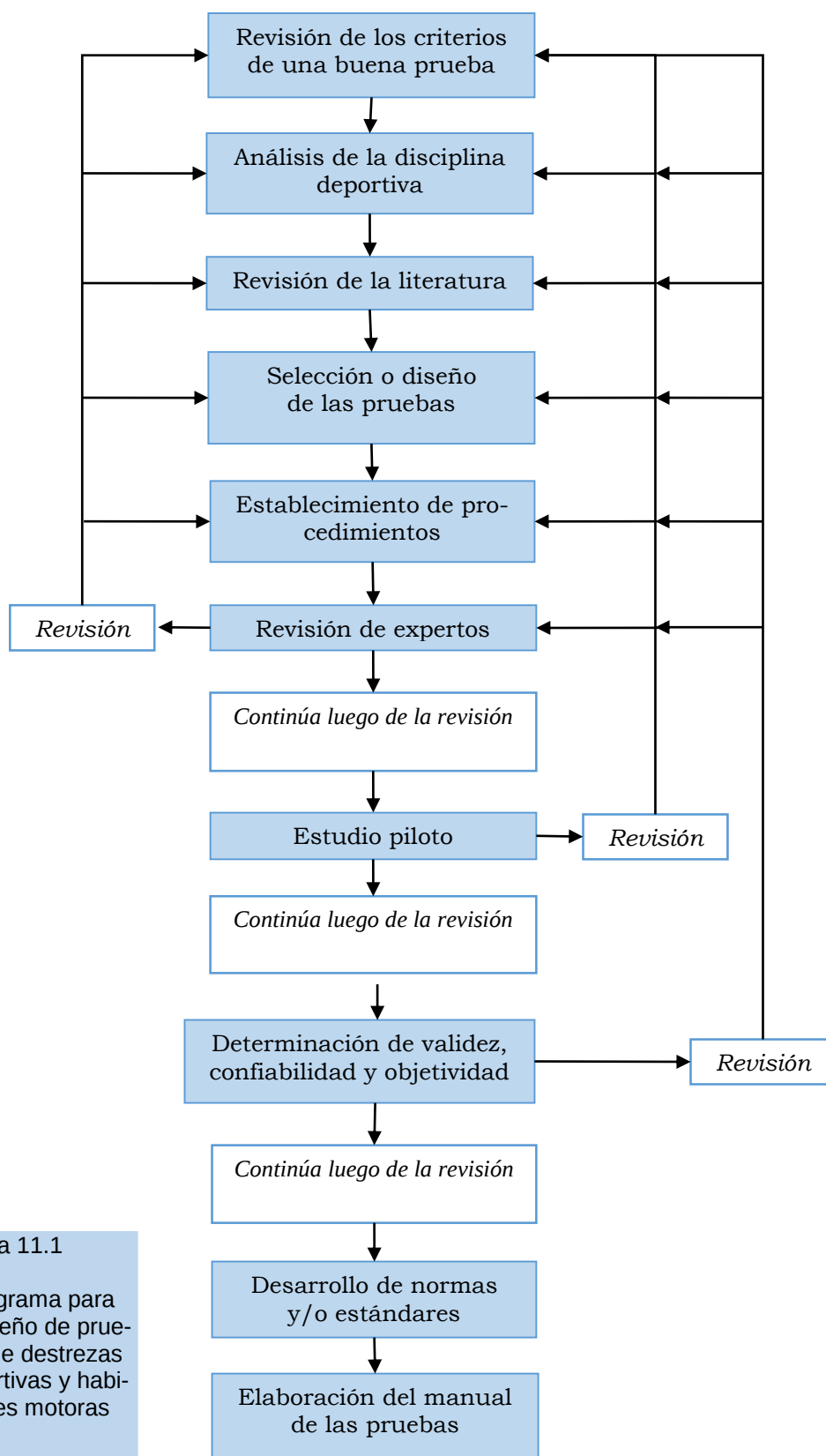


Figura 11.1
Flujograma para el diseño de pruebas de destrezas deportivas y habilidades motoras

Paso 7. Estudio piloto. Si las pruebas seleccionadas se basan en la opinión de los expertos y en la revisión de la literatura, el análisis del estudio piloto ayudará a determinar las propiedades de la prueba. La prueba piloto es un paso importante que se da antes de culminar con el diseño de la prueba, ya que ayuda a determinar el tiempo total de su administración y la claridad de las instrucciones al mismo tiempo que puede señalar las posibles fallas en la prueba.

Paso 8. Determinar las características estadísticas de la prueba. Se trata de la fiabilidad, validez y objetividad. Los coeficientes de fiabilidad se estiman y son precisos sólo para los grupos medidos. Deben ser grupos específicos especialmente si se trata de adolescentes; es importante que ambos géneros y todas las edades y niveles de destreza incluidos en la muestra normativa sean medidos. También es importante validar el contenido antes de la evaluación estadística de la validez. Se puede determinar la validez concurrente o de constructo al momento de acometer los procedimientos respectivos. La estabilidad de las pruebas de desempeño motor generalmente se establece con el coeficiente de fiabilidad que implican las mediciones repetidas.

Paso 9. Establecer las normas para las pruebas referidas a normas o determinar los estándares para las pruebas referidas a criterios.

Paso 10. Elaborar un manual que describa completamente la prueba, los procedimientos de anotación y las características estadísticas (fiabilidad y validez).

Paso 11. Evaluar el instrumento cada cierto tiempo. Cuando el tiempo pasa, los estudiantes pueden tener un nivel de preparación diferente por lo que las normas y estándares que fueron apropiados en una época pudieran no ser apropiados para otra época.

Elementos de la medición de las destrezas

Los elementos básicos de la medición más importantes para la evaluación de las destrezas son la confiabilidad y la validez. También hay otros elementos que destacar, uno de los cuales es la factibilidad. Las pruebas de destrezas toman su tiempo para aplicarlas apropiadamente por lo que el evaluador debe hacerse una pregunta: ¿Es más importante enseñar las destrezas y que los estudiantes la practiquen, o emplear el tiempo midiendo y evaluando las destrezas? Un segundo elemento es la determinación de la mejor manera de evaluar la destreza. ¿Se aplicarán pruebas objetivas que no tienen similitud con el juego, o es mejor implementar pruebas subjetivas similares a alguna situación de juego, aunque tengan muy poca fiabilidad?

Cuando se evalúan destrezas deportivas el docente o entrenador tiene que definirse entre una prueba que sea extremadamente objetiva y confiable, pero sin aplicabilidad en el juego, o una prueba que sea altamente válida pero menos objetiva y que consume más tiempo. Por ejemplo, un docente o entrenador podría medir varias destrezas asociadas con técnicas de voleibol, junto con la realización de alguna prueba de desempeño motor. Una de las destrezas es el voleo de pelotas bajas, que podría evaluarse mediante tres pruebas:

el autopase, el voleo contra la pared y el pase en la cancha. El autopase y el pase a la pared son pruebas individuales que pueden administrarse a grupos grandes en un espacio relativamente pequeño. Ambas pruebas producen consistentemente altos grados de confiabilidad. El problema con estas pruebas es que no son aplicables en situaciones de juego. En el autopase el ejecutante se pasa el balón repetidamente a sí mismo, con el criterio de que el balón sea elevado por lo menos a 3 metros en cada repetición. En el pase a la pared, el ejecutante dirige el balón por arriba de una línea trazada en la pared a 3 metros del piso mientras permanece detrás de una línea de restricción separada a 1,80 metros de la pared. En ambos casos, los participantes pueden practicar y medirse a sí mismo. Este tipo de pruebas se diseñan para su administración en grupo y la anotación del compañero; son factibles de aplicar y requieren de pocas marcas en la cancha, así como un mínimo de materiales y tiempo de clase. Sin embargo, los expertos en voleibol argumentarán que estas pruebas no son transferibles a situaciones de juego como la habilidad para recibir un servicio o un remate del equipo contrario.

Por otra parte, en la prueba de pase a la cancha se coloca a los participantes en la zona de zagueros, en una posición donde sea más probable que reciba el servicio contrario. Ante un balón que es enviado por un aplicador, se les pide que hagan un pase hacia un área predeterminada. Se utiliza un sistema de puntos para determinar la precisión en la destreza al realizar 10 intentos. Esta prueba posee una buena confiabilidad y es válida en términos de la validez de contenido. Sin embargo, requiere de un largo tiempo para su aplicación que en las pruebas anteriores. Asimismo, es indispensable que los aplicadores sean bastante consistentes al realizar el servicio que se envía al receptor.

Clasificación de las pruebas de destreza deportiva

Cuando se diseñen pruebas de destreza deportiva se debe precisar si se utilizarán procedimientos de carácter objetivo o subjetivo, asegurando también que se va a medir sólo un aspecto en cada prueba; es apropiado incluir pruebas subjetivas como formas alternas de medición al diseñar una batería de pruebas para evaluar técnicas complejas en algunas disciplinas deportivas. Una gran cantidad de factores como el tiempo, los materiales, el número de aplicadores y el número y tipo de pruebas que se aplicarán, influyen en la decisión de medir las destrezas bajo el enfoque específico del carácter objetivo o subjetivo de la prueba.

Pruebas objetivas. Hay cuatro formas de clasificar las pruebas de destrezas deportivas objetivas:

- Pruebas de precisión
- Pruebas de repetición
- Pruebas de movimiento total del cuerpo
- Pruebas de potencia y distancia

Pruebas de precisión. Las pruebas de precisión por lo general implican las destrezas de servicio de objetos como el balón de voleibol o la pelota en tenis y bádminton. También

se incluyen acá los lanzamientos en fútbol americano y béisbol, el tiro libre y otros lanzamientos en baloncesto, o golpear el balón hacia el arco en fútbol de campo. El principio fundamental de la medición relacionada con las pruebas de precisión es el desarrollo de un sistema de puntaje que provea de resultados válidos y también confiables.

Se deben prever situaciones que disminuyan la fiabilidad y validez de una prueba. Veamos el caso de una prueba de tiro a gol en fútbol de campo. La prueba consiste en golpear el balón hacia el arco desde una distancia de 11 metros. Se divide el arco colocando cuerdas a 90 centímetros de cada poste en forma vertical siendo el blanco el área entre el poste y la cuerda. Se otorgan 2 puntos si el balón entra en el blanco, 1 punto si entra por el centro, y ningún punto si no entra. Al ejecutante se le permiten 6 intentos, tres hacia el lado derecho y tres hacia el lado izquierdo. El problema con esta prueba es que los mejores tiradores tratan en todo momento de que el balón entre por el blanco de los extremos, pero pueden terminar desviados y el puntaje es cero, mientras los ejecutantes menos competentes golpean hacia el centro para asegurar al menos un punto. Así, los mejores golpeadores podrían obtener menos puntos que los menos diestros, situación que reduce la validez y fiabilidad de la prueba.

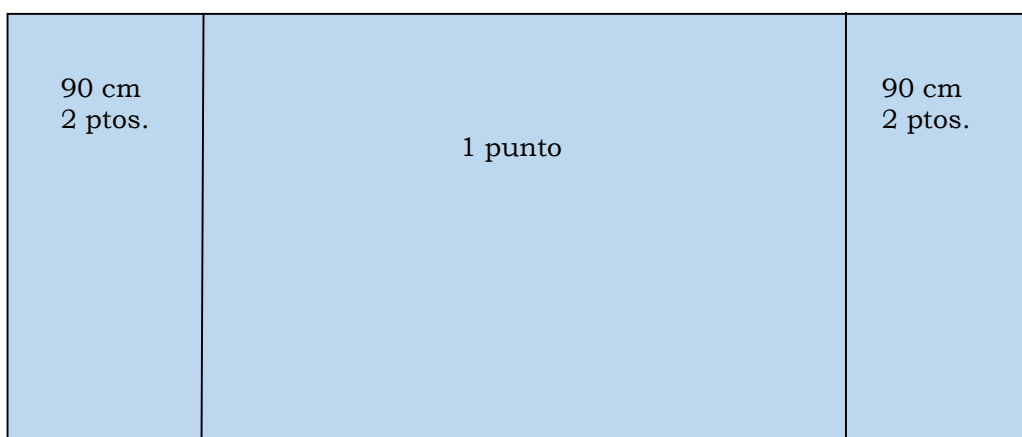


Figura 11.2 Diseño de una prueba de precisión en fútbol

En las pruebas de voleibol, la cancha es marcada para delinear el área hacia donde es más difícil dirigir el balón en un servicio. En problema señalado en la prueba de tiro en fútbol de campo también se presenta en este caso porque al servicio que caiga fuera de la cancha se le da un puntaje de cero. Para rectificar este problema, el valor del puntaje más bajo del blanco puede asignarse al servicio que esté ligeramente por fuera de ese blanco. Sin embargo, esto puede acarrear un problema de factibilidad como lo es el marcaje del terreno previa determinación de la distancia máxima permitida para otorgar el puntaje más bajo al servicio que esté fuera del blanco.

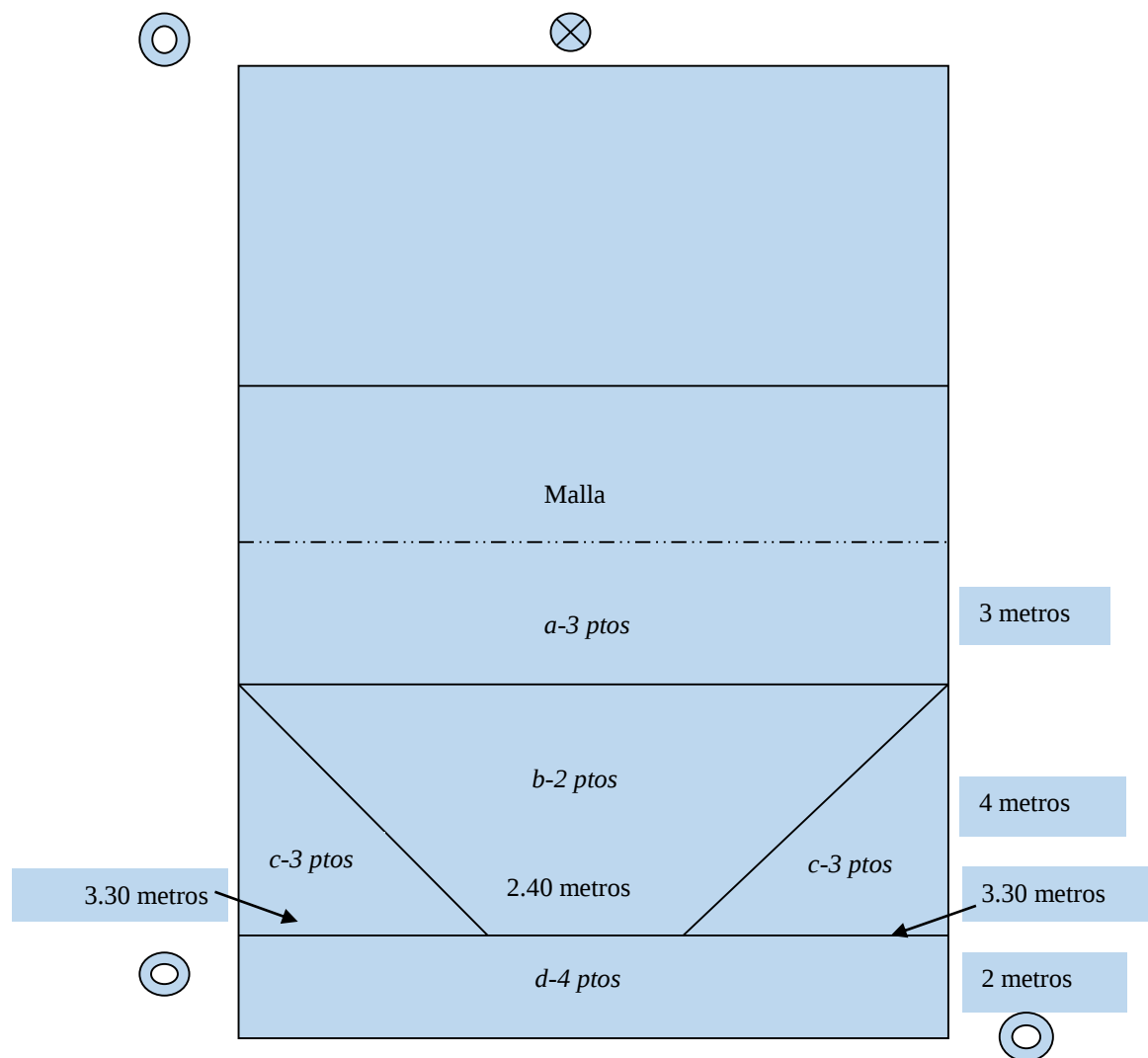


Figura 11.3 Diseño de una prueba de precisión en voleibol

Otra consideración con respecto a las pruebas de precisión es el número de repeticiones necesarias para producir puntajes confiables. Cuando se diseñaron las primeras pruebas de destrezas en tenis de campo, el componente de servicio no obtuvo el valor de fiabilidad necesario como para ser incluido en la batería y en el estudio estadístico correspondiente se recomendó elevar el número de intentos al doble para mejorar la fiabilidad de la prueba significativamente, lo que también impactó la factibilidad de la prueba. Luego se modificó el sistema de puntuación para elevar la fiabilidad a un nivel aceptable. Por otro lado, si se utilizan múltiples intentos en el servicio, es necesario maximizar la fiabilidad mientras se minimiza el número de intentos.

Pruebas de repetición. Son pruebas que implican un desempeño continuo en una actividad y un período específico de tiempo. Unas pruebas utilizadas en voleibol son el voleo contra la pared y el autopase. En otras disciplinas se pueden medir los golpes en los deportes de

raqueta como los golpes de derecha y revés. Las pruebas de repetición por lo general poseen un alto grado de fiabilidad, pero al menos que sean cuidadosamente construidas, podrían estar lejanas al tipo de movimiento que se utiliza en el juego por lo que se reduce sensiblemente su validez. Además, si la prueba no es similar a la forma de efectuarse el juego, no se puede transferir al desempeño real en competencia como puede suceder en otras pruebas. Por ello, es extremadamente importante cuando se utilicen pruebas de repetición que los evaluadores estén seguros de que los sujetos empleen la técnica apropiada.

Pruebas de movimiento total del cuerpo. Las pruebas de movimiento total del cuerpo son llamadas también pruebas de velocidad ya que ellas evalúan la rapidez con la cual un sujeto completa una tarea que implica el movimiento total del cuerpo en un área restringida. En una prueba de drible en baloncesto o fútbol de campo se mide esta destreza, al igual que en las pruebas de correr las bases en béisbol y softbol. Estas pruebas por lo general poseen un alto grado de confiabilidad porque tienen una gran cantidad de variabilidad interindividual asociada al desempeño medido.

Las pruebas pueden ser administradas rápidamente, aunque existen dos problemas inherentes a las propias destrezas. Primero, una prueba debe estar próxima al desempeño en situaciones reales de juego, aunque en muchas ocasiones la velocidad no sea requerida durante la competencia, incluso en situaciones como en el rompimiento rápido en baloncesto donde debe haber algún grado de velocidad controlada tal que permita el control del balón. Por ello, para medir el drible en baloncesto se selecciona una prueba que implique driblar siguiendo el curso de los conos como se muestra en el siguiente gráfico. Cuando este tipo de prueba se utiliza, el curso debe ser creado para aproximarlos válidamente a la destreza aplicable en un juego. Segundo, cuando estas pruebas se utilizan para evaluar el tiempo ejecutado es el criterio válido, pero si el interés está en la eficiencia del desempeño las pruebas deben relacionarse en un alto grado con la velocidad del ejecutante. Es obvio que un ejecutante rápido puede completar la prueba más rápido que uno más lento, aunque el ejecutante más lento posea más destreza en el manejo del balón.

Una manera de eliminar el problema de la velocidad inherente a las pruebas de movimiento total del cuerpo es crear una proporción del desempeño, dividiendo el tiempo de desempeño (tiempo empleado en driblar) entre el tiempo de movimiento (tiempo en efectuar el mismo recorrido corriendo sin balón) del mismo sujeto. Esta proporción del desempeño para esta destreza se denomina eficiencia en el drible. La proporción de desempeño puede ser una poderosa herramienta motivacional tanto para los más diestros como los menos diestros porque se están enfrentando consigo mismo, tratando de reducir la proporción tan cercana a uno (1) como les sea posible. Teóricamente, una proporción de 1 debería ser el valor mínimo para este tipo de prueba. Inversamente, ya que el tiempo se anota de acuerdo con este tipo de proporción, los sujetos más rápidos podrían ser excesivamente penalizados. Asimismo, si los sujetos son conscientes de la manera como se crean estas proporciones, podrían dejar de hacer el mejor esfuerzo al correr sin balón lo que les permitiría lograr una mejor proporción de lo que su desempeño puede demostrar. Estas proporciones no son apropiadas en disciplinas deportivas por cuanto el desempeño ocupa un lugar preponderante como en los equipos que compiten en campeonatos, pero es un enfoque efectivo en la medición de la eficiencia de la destreza en el ambiente escolar.

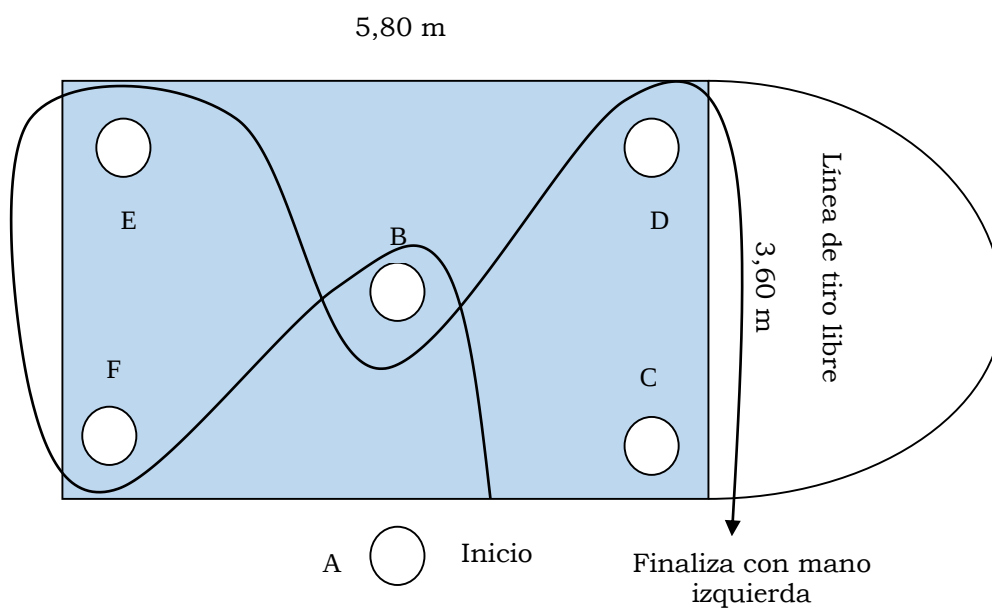
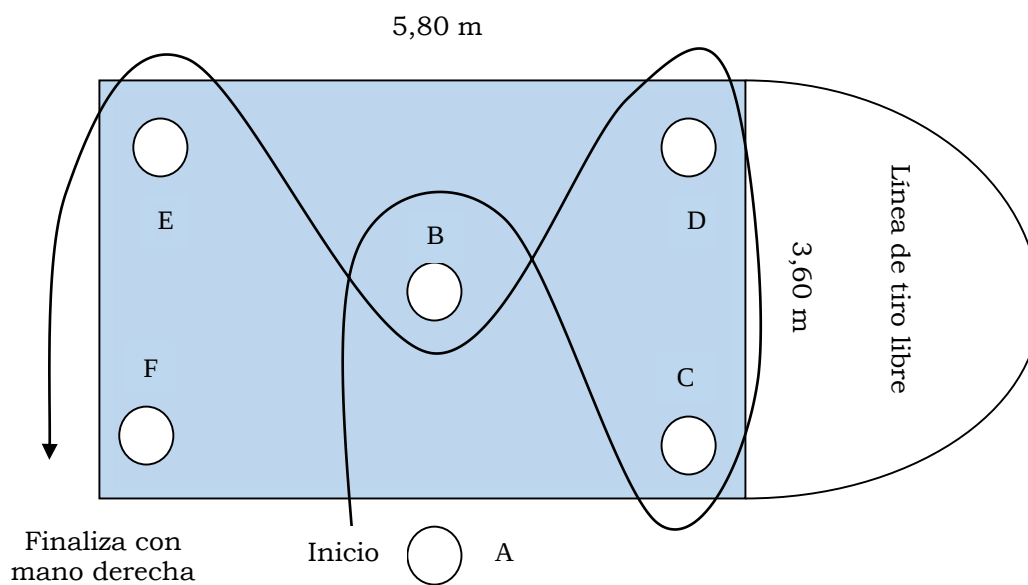


Figura 11.4 Diseño de una prueba de movimiento total del cuerpo en baloncesto

Pruebas de potencia y distancia. Las pruebas de potencia y distancia valoran la habilidad para proyectar un objeto con máxima fuerza o mayor distancia posible. La prueba de impulso en bádminton, potencia del servicio en racquetbol, distancia de lanzamiento en béisbol y softbol, así como las competencias de golpear, pasar y batear, son ejemplos de este tipo de prueba. Un problema con estas pruebas es que debe asegurarse que sean efectuadas

de forma similar a las situaciones reales de juego. Otra cuestión que se debe resolver es si se toma en cuenta la precisión. Por ejemplo, en una prueba de golpear, pasar y batear la distancia fuera de la línea o área de proyección se resta de la distancia de proyección. Debido a esto, un sujeto podría abstenerse de usar su máxima fuerza por miedo a perder precisión. Por el contrario, en las pruebas de campo en atletismo como el disco y la bala, esas correcciones no se hacen porque que el objeto debe caer dentro del área establecida. En consecuencia, antes de aplicar una prueba es importante determinar si la prueba requiere o cuenta con correcciones sobre la precisión. En la prueba de potencia del brazo en voleibol donde se lanza el balón de baloncesto, la precisión no se considera un componente importante de la potencia del brazo asociada con la habilidad para golpear en voleibol; la distancia absoluta alcanzada es el criterio de la prueba. Sin embargo, en las pruebas de otras destrezas deportivas como lanzar en fútbol americano y béisbol, la precisión es un componente muy importante que requiere de algún tipo de corrección. La corrección más simple que se utiliza es la sustracción de la distancia fuera de la línea o área de proyección, de la total alcanzada.

Prueba de lanzamiento sobre el brazo en softbol

Objetivo: Medir la capacidad para lanzar una pelota de softbol con la mayor precisión y distancia posible.

Equipamiento: Terreno de softbol, 2 cintas métricas, pelotas de softbol, 2 conos o estacas de marcación.

Procedimiento: Se marca una línea de lanzamiento de 3 metros de longitud con otra línea paralela a ella a 3 metros de distancia, ambas de unos 5 centímetros de ancho y en el centro del campo (centerfield) a 2 metros de la zona de seguridad, conformando así el área de lanzamiento para proyectar la pelota hacia el home. Desde el centro de la línea de lanzamiento se coloca una cinta métrica o se marca el terreno cada metro desde el centro del campo en línea recta. Cada sujeto debe realizar lanzamientos de calentamiento por 3 o 4 minutos.

El sujeto se coloca en el área de lanzamiento pudiendo dar los pasos de impulso que desee en ese espacio. Se le dan 2 intentos para lanzar tan lejos y directo como le sea posible en forma rasante, de forma tal que pique lo más cerca posible de la línea demarcada o la cinta métrica. No se puede sobrepasar o pisar la línea de lanzamiento al realizar el intento, ni se permite lanzar en globo. Si se sobrepasa o pisa la línea debe repetir el intento, con sólo una oportunidad de repetir cada vez que comete el error. La repetición debe ser inmediatamente después del error, pero el primer y el segundo intento no deben ser consecutivos. Si la prueba se hace para un grupo, cada uno pasa la primera vez y luego se procede con el segundo intento.

Registro: Se anota la distancia alcanzada en centímetros, procediéndose de la siguiente manera. Si la pelota toca el terreno hacia la derecha o izquierda de la cinta métrica o de la línea demarcada, el juez marca el punto donde pica la pelota y proyecta ese punto hacia la cinta métrica en forma paralela a la línea de lanzamiento. Se mide la distancia entre el punto de caída y la cinta métrica o la línea en centímetros y se le resta a la distancia del lanzamiento, anotándose el resultado en la hoja correspondiente. El puntaje del sujeto es el mejor de los dos intentos.

Capítulo 12. La evaluación cualitativa en educación física

Antecedentes

En la praxis educativa, la evaluación se ha desnaturalizado debido a su exclusivo uso como un medio para la obtención de una calificación parcial o final, e incluso se tergiversa el sentido de la investigación, el informe, la intervención, la discusión y otras importantes técnicas, en razón de la presión administrativa para entregar las calificaciones. Además, los instrumentos de evaluación que se aplican están colmados de falta de validez y confiabilidad.

En múltiples ocasiones se ha afirmado que la evaluación formativa y sumativa practicada en la realidad no cumple con los preceptos teóricos y legales. Sin duda, el tipo de evaluación sumativa provoca más temor, ansiedad y estrés que la evaluación formativa. En la evaluación sumativa no se revisa, por ejemplo, si el aplazamiento de un grupo de estudiantes obedece a que no entendieron ni practicaron, o que el método del docente no era pertinente, o que la información fue incompleta. En muchos casos se implementan métodos basados en procesos de aprendizaje, pero se evalúa en forma tradicional. En el peor caso, el docente otorga una calificación sin cumplir con su compromiso pedagógico y metodológico, ni conocer las verdaderas capacidades de los estudiantes.

La evaluación formativa, que debe ser eminentemente cualitativa, sólo se utiliza para señalar un porcentaje de la calificación final en el boletín. Se ha transformado en una evaluación sumativa disfrazada, sin acercarse a la definición de fondo: hay ausencia del diálogo como el principal medio de la evaluación formativa.

Ante la contundente realidad de la evaluación escolar, se ha propuesto un profundo cambio de paradigma apoyado en lo sustancial: cómo aprenden los estudiantes. La evaluación en este caso tiene carácter cualitativo, basada en la descripción de las condiciones del proceso de enseñanza – aprendizaje y su relación con los resultados obtenidos. Es una evaluación concebida como investigación, análisis, valoración y reflexión acerca de la forma de aprender y los factores que propician u obstaculizan el aprendizaje. Asimismo, la evaluación es un proceso que construyen los estudiantes, el docente y los demás actores que participan y contribuyen a elaborar el conjunto de datos. Este enfoque se apoya en la visión del aprendizaje centrado en la construcción de significados, tratando de asegurar las condiciones para que los estudiantes desplieguen sus potencialidades y capacidades, aprendiendo mientras se les evalúa.

En este contexto, el acto pedagógico no se presenta como una mirada “sobre” (el estudiante), sino como una mirada “conjunta” (docente – alumno, alumno – alumno), donde cada uno acepta compartir un trozo de camino con el otro, a pesar de sus diferencias y creencias. Pero, ¿cómo se podrá abrirles el mundo a los estudiantes enseñándoles a ser,

a hacer, a conocer, y a convivir a pesar de las diferencias, si la evaluación, núcleo de toda acción educativa, frena e incluso impide la adquisición de tales aprendizajes? Además, si en el contexto del aula de clase, por ejemplo, el diálogo se convierte en la fuente de la evaluación formativa, ¿cuáles procedimientos son los apropiados en educación física?

Bases de la evaluación cualitativa

En los últimos años se ha extendido el uso del término valoración en oposición a la evaluación y medición tradicional, utilizándose ahora técnicas de valoración auténtica o alternativa para darle forma y significado a la evaluación cualitativa. En algunos casos, al hablar de valoración se ha entendido como sinónimo de evaluación cualitativa.

La valoración es el proceso de recolección de información, mientras la evaluación resulta de un juicio o interpretación sobre la valoración efectuada, un juicio sobre los productos y procesos de la instrucción. Es decir, la valoración utiliza los datos recolectados al medirse el aprendizaje de los estudiantes, que luego sirven de soporte a la evaluación. Entre las bases teóricas de la valoración, que coadyuvan al mejoramiento de la evaluación del estudiante, encontramos las siguientes:

La valoración orienta la instrucción. No se puede ver la valoración del estudiante como un producto final. Se espera una valoración frecuente (evaluación formativa) de las destrezas, conocimientos y conducta, de forma tal que los docentes utilicen la información obtenida de esas evaluaciones para diseñar una clase centrada en los intereses del estudiante.

Los docentes deben evaluar todos los dominios: psicomotor (destrezas), cognitivo (conocimientos) y afectivo (conductas, actitudes). Hoy estamos viviendo una era de sobrecarga informativa. En el tiempo asignado a la escuela, es imposible enseñar y evaluar cada unidad de información. La selección del contenido requiere del docente unas decisiones difíciles, y aún más difícil es la tarea de diseñar una valoración que mida las competencias del estudiante en relación a ese contenido.

Los docentes deben evaluar más allá del nivel cognitivo. El mundo real requiere que los estudiantes apliquen la información y lo hagan en forma competente. Los estudiantes necesitan participar en estrategias de valoración diseñadas para demostrar habilidades en las áreas de comprensión, interpretación, aplicación, análisis, síntesis y evaluación.

Los docentes deben equilibrar la selección de las estrategias de valoración. Los educadores necesitan ir más allá de las pruebas estandarizadas para la medición del aprendizaje de los estudiantes. Se debe utilizar una amplia variedad de herramientas de valoración que reflejen los principios de la valoración auténtica, alternativa e incluso tradicional.

Se deben considerar los propósitos de la valoración en educación física.

- Suministrar a los estudiantes información sobre su progreso;
- Motivar a los estudiantes para que mejoren su desempeño;
- Formarse un juicio acerca de la efectividad de la enseñanza;

- Suministrarle al docente información sobre el nivel de dominio de los objetivos por parte de los estudiantes, para que pueda ajustar el proceso de enseñanza – aprendizaje.
- Evaluar el currículo o programa.
- Ubicar a los estudiantes en un grupo de trabajo apropiado.
- Presentar una información objetiva al docente sobre el avance de los estudiantes, para efectos de la calificación.

Tipos de valoración

La valoración o evaluación cualitativa es muy diferente a la evaluación tradicional, la cual utiliza la evaluación unidireccional efectuada en un ambiente artificial y simplificado. Las principales preocupaciones de las estrategias de evaluación tradicional se limitan a un examen del contenido, bajo la forma de pruebas de verdadero – falso o de selección múltiple. Ejemplos de ello en el campo de la educación física son las pruebas de destrezas motoras, las pruebas de condición física y las pruebas escritas sobre la historia del deporte o sus reglas. A continuación, se distinguen algunos tipos de valoración o evaluación cualitativa.

Valoración auténtica. Se diseña para aplicarse en situaciones de la vida real, al contrario del ambiente artificial típico de la evaluación tradicional. Por ejemplo, la evaluación auténtica implica la observación de la estrategia que selecciona un jugador y la ejecución de las destrezas durante una competencia y no como se hace en un test de destreza deportiva, que mide la destreza fuera del contexto de juego. En educación física, se necesita una valoración auténtica que permita al estudiante aplicar sus destrezas y conocimientos de manera que le sean significativas dentro y fuera de los confines del programa y la clase. Esto puede incluir la anotación del puntaje que demuestre las competencias alcanzadas en el curso de judo, monitorear y registrar el tiempo empleado dentro de la zona de ritmo cardíaco al caminar o trotar utilizando un monitor del ritmo cardíaco, o anotar el entrenamiento diario en su registro de actividad física.

Valoración alternativa. Se considera diferente a la forma convencional de pruebas estandarizadas (selección múltiple, verdadero – falso, tests de destrezas deportivas o condición física). La valoración alternativa se apoya en una amplia variedad de técnicas no tradicionales para la valoración del logro, tal como el portafolio, el desempeño de roles, entrevistas, organización de eventos, proyectos individuales y en grupo, diarios o periódicos estudiantiles, exhibiciones, la observación. Esta valoración alternativa se basa en la evaluación con base en criterios y la puntuación otorgada se basa en juicios subjetivos. La valoración alternativa requiere que los estudiantes utilicen destrezas de pensamientos superiores y de toma de decisiones. Idealmente, la valoración alternativa implica la solución activa de problemas reales, por ejemplo, a través de la aplicación de la nueva información.

Valoración del desempeño. Incluye métodos o pruebas para crear una respuesta que demuestre un conocimiento o destreza. Los estudiantes deben dirigir una actividad o crear un producto, antes que seleccionar o responder la lista de alternativas suministrada por el docente. Los exámenes tradicionalmente se han concentrado en la respuesta correcta, sin preocuparse si los estudiantes comprenden el problema o si pueden aplicar aquello que

comprenden. En la valoración del desempeño se distingue entre ser capaz de describir cómo una destreza debe ser ejecutada (conocimiento) y ser capaz realmente de ejecutarla: el desempeño, una importante característica en educación física y deporte. Un estudiante puede diseñar los pasos a seguir para practicar los pases en fútbol, mientras otro presenta un portafolio sobre los puntajes obtenidos en las sucesivas y complejas rutinas gimnásticas. En cada caso, las tareas pueden ser individuales, en pareja o en grupo, la valoración puede reflejar el tipo de aprendizaje cognitivo, afectivo o psicomotor, mientras la información del desempeño puede ser inherente a la tarea, autorregulada o suministrada externamente.

Valoración analítica. Al evaluarse el desempeño del estudiante en una destreza en particular, como el tiro libre en baloncesto o el remate en voleibol, haciendo énfasis en la forma de realizar el movimiento técnico, lo que puede ser útil cuando se efectúa con el fin de diagnosticar específicamente sólo la destreza o tarea en cuestión. Sin embargo, ello indica poco sobre la capacidad del estudiante para jugar baloncesto o voleibol.

Valoración holística. Se ha sugerido que más allá de la observación aislada de las destrezas, los docentes se concentren en valorar holísticamente la actuación del estudiante, juzgando la calidad total de su actuación deportiva.

Actualmente y de conformidad con la noción de que la valoración sea progresiva, acumulativa y orientada al aprendizaje, ella debe reflejarse en la práctica formativa. Si se selecciona una valoración auténtica o del desempeño, dado que se conoce que el aprendizaje y la valoración se encuentran relacionados, la práctica de la valoración debe reflejar el proceso de aprendizaje. Si los estudiantes sólo aprenden lo que será evaluado, entonces la construcción de la valoración y el mismo proceso de aprendizaje deben incrementar el nivel de dominio alcanzado.

La valoración debe ser acompañada de estándares que orienten la puntuación, es decir, el criterio de desempeño apropiado. Sin algún tipo de estándar para otorgar una puntuación no habrá una valoración sino una actividad instruccional. Muchos docentes han seguido esta corriente en el marco de la evaluación continua, pero la mayoría de los instrumentos de valoración que se han creado en nombre de la evaluación continua son inadecuados por el simple hecho de que no contienen el componente de comparación.

La valoración en educación física

La valoración siempre ha sido parte del modelo teórico del proceso instruccional plan – enseñanza – evaluación, pero ello no ha recibido mayor atención por parte de los docentes de educación física. Probablemente la principal razón haya sido que los docentes no suministran información sobre el desempeño del estudiante a nadie, sino que asignan una calificación. En muchas ocasiones se carece de datos para efectuar la valoración. La falta de valoración del proceso de enseñanza – aprendizaje y de la evaluación de los planes también puede atribuirse al tipo de material de valoración que se les ha dicho que utilicen,

la operatividad de tales materiales y el tiempo que se programa para la valoración del proceso. Los docentes exponen que muchos tests tienen poca relación con lo que están enseñando, o que los tests requieren de mucha preparación y tiempo como para aplicarlos a gran cantidad de estudiantes.

Al hablar de evaluación puede pensarse en calificación, y esto es una realidad en el medio educativo. Muchos docentes, estudiantes, padres y representantes no están satisfechos por la manera cómo se califica. Realmente, en educación física la calificación no se basa en las mediciones efectuadas en ciertas ocasiones. La tendencia, más bien, es reflejar una “nota” efectuada cerca del final del lapso. Los docentes tienden a calificar a los estudiantes sobre la base de variables como participación, actitudes, conductas y asistencia, pero no por el conocimiento o la ejecución de destrezas. En ocasiones los docentes de educación física se limitan a calificar al estudiante con las únicas alternativas de aprobar/aplazar. En otras palabras, los docentes de educación física tienden a calificar con base a criterios diferentes al aprendizaje y al desempeño. No sólo no se califica el aprendizaje: tampoco se evalúa a los estudiantes de manera sustantiva. Varias razones se han expuesto para explicar esta falta de atención al logro de las metas en educación física, incluyendo secciones superpobladas, equipos inadecuados de evaluación, falta de tiempo para una evaluación real, la creencia de los docentes que las pruebas son innecesarias o, incluso, que ellos no han sido formados para conducir unas pruebas elaboradas en forma apropiada y significativa. También se ha sugerido que directivos y administradores afirman que mantener a los estudiantes activos es más importante que su progreso.

El nuevo enfoque postula cambiar el qué y el cómo de la evaluación, de forma tal que se integren al proceso de enseñanza – aprendizaje y así obtener información relevante sobre los logros de los estudiantes. Mientras más fuerte sea el vínculo entre la evaluación y los resultados del proceso de enseñanza – aprendizaje, más profundos serán los logros del estudiante. Esto confirma que el diseño de los objetivos, la implementación de experiencias de aprendizaje que oriente el progreso de los estudiantes hacia esas metas y hacer que ellos se responsabilicen de lograrlas, hace que la integración de la evaluación al proceso de enseñanza - aprendizaje sea más efectiva.

La observación en el proceso de valoración. Se ha dicho que los docentes de educación física deben tener “ojos en la nuca” para saber lo que cada estudiante hace aun si se hallan a su espalda, lo cual sugiere que se debe conocer todo lo que sucede en el aula, cuáles estudiantes tienen dificultades, quién es exitoso, y cuándo dejan de practicar la tarea. La observación se utiliza en la clase como un medio para suministrar información sobre el desempeño, organizar el grupo, informarse sobre el proceso de enseñanza – aprendizaje o revisar la valoración que hacen los estudiantes de sus compañeros. Cuando el docente se habitúa a la observación continua, puede convertirla en criterio de evaluación que puede formalizar y registrar en un instrumento de valoración.

La realización de un movimiento ofrece oportunidades para que un estudiante demuestre comportamientos que pueden ser observados por otros, lo cual es una ventaja para quienes laboran en el campo del dominio psicomotor. La naturaleza de la ejecución de una destreza motora hace que la valoración, a través de la observación y el juicio subjetivo,

una herramienta útil para muchos docentes de educación física. Las investigaciones realizadas sobre la medición y evaluación de los docentes de educación física han mostrado confianza en los métodos de observación y valoración subjetivos.

Sin embargo, la valoración que depende del juicio de los observadores ha sido criticada por sustentarse en una validez y confiabilidad muy cuestionadas, además que ha sido susceptible de factores como el sesgo personal, el error de generosidad o de severidad, la falta de anotación objetiva, su falta de conducción sistemática o que no se compara con criterios definidos. A pesar de estas críticas, la observación y el juicio apropiado representan la mejor esperanza para la valoración auténtica en educación física, que provea de información significativa y relevante.

La observación tiene una profunda importancia para la educación física, tanto desde el punto de vista de la valoración o evaluación cualitativa como del proceso de enseñanza – aprendizaje. En ambos casos se define observación como el proceso de obtener, organizar y darle significado a la información proveniente del desempeño del sujeto observado. La observación abarca diferentes modos de percepción, pero debe limitarse la variabilidad de la información que se recolecta, ya que la tarea de observación se concentra en esa recolección y no en su evaluación.

Instrumentos de observación

Los instrumentos para juzgar el desempeño y progreso de los estudiantes han sido llamados de muchas maneras: lista de chequeo, escala de estimación, lista de cotejo, escala descriptiva, etc. Estos instrumentos representan la columna vertebral de un efectivo sistema de valoración del desempeño. Una escala descriptiva es comparable a las escalas de estimación que se han utilizado en educación física por mucho tiempo, pero difieren por el hecho de que la escala descriptiva facilita la instrucción y puede ser entregada a estudiantes, directivos, padres y representantes, al contrario del sistema secreto mediante el cual se pretende calificar al estudiante. A continuación, se presentan varios de estos instrumentos.

Lista de chequeo o cotejo. Están diseñadas para juzgar si unas conductas o características específicas están presentes o no en el desempeño de los estudiantes o en el resultado de su trabajo. Este sistema de puntuación utiliza sólo dos niveles para la estimación del desempeño. No se intenta juzgar cuán bien se demuestra cada característica del movimiento sino verificar si la característica se halla presente o no. En la mayoría de las ocasiones la lista de chequeo se construye con base en las características críticas del movimiento que se valora, y el docente determina si el estudiante exhibe o no esas características. Puede utilizarse para medir el desempeño durante el proceso (la secuencia) o como producto (el resultado) y el docente, un compañero u otro observador determina si se demostró el criterio. Las listas de chequeo, en educación física, están asociadas con la observación del desempeño en vivo, pero también pueden utilizarse para la valoración escrita o de un video grabado.

Escala de estimación. Mientras en una lista de chequeo se anota la demostración de un criterio específico, en una escala de estimación se registra el grado de dominio o calidad de ese criterio. Ese grado puede anotarse en forma numérica (1 al 5, siendo 5 el puntaje más alto), descriptiva (siempre, a menudo, algunas veces, rara vez, nunca), o por nivel (principiante, intermedio, avanzado). Dado que con la escala de estimación requiere de dos decisiones, una sobre la presencia y otra sobre la calidad del movimiento, se necesita de más tiempo y atención lo cual sugiere que en este tipo de escala deben incluirse pocos criterios de observación. La escala de estimación suministra datos más completos y detallados que una lista de chequeo, pudiendo ser utilizada por los estudiantes para informar sobre su propio desempeño o por los docentes para mejorar determinadas necesidades de los aprendices. Adicionalmente, este tipo de escala estimula a los estudiantes a concentrarse más en la calidad del desempeño que en su cantidad. Cuando se estimula a los estudiantes a considerar el grado de dominio sobre algo y a hacerse más analíticos sobre su desempeño, se concentran más en la calidad del movimiento. En forma similar, los docentes que usan una escala de estimación para valorar el desempeño de los estudiantes se hacen más aptos para recolectar datos útiles que realmente describan el progreso hacia las metas.

Escala descriptiva. Este instrumento suministra un esquema de las pautas para el registro del desempeño del estudiante. Una escala descriptiva bien diseñada identifica los criterios esenciales del desempeño y tiene la propiedad de poder ser utilizada por el docente o por un observador externo. En una escala descriptiva holística se registra el desempeño como una totalidad, en un ambiente real. La escala descriptiva analítica se concentra en los elementos esenciales de una técnica u otro tipo de movimiento. Dado que las conductas complejas pueden valorarse en varias dimensiones, con la escala descriptiva se facilita la observación de varias conductas en forma simultánea. Al construirla, los criterios de valoración más importantes se ubican en el nivel superior, y luego se anotan los demás criterios que se establezcan. La parte más importante del diseño de una escala descriptiva es que el docente establece con anterioridad la importancia del criterio que se valorará. Si la valoración es parte de la planificación instruccional, entonces debe haber un vínculo entre lo que el docente planifica, lo que realmente enseña y lo que se valora. Las tablas 12.1, 12.2 y 12.3 son ejemplos de este tipo de escala, que van de una escala general a una escala específica.

Observación y valoración del compañero. Se han desarrollado experiencias de observación del desempeño del estudiante por parte de un compañero de estudios en aspectos tales como el registro del número de intentos ejecutados en una destreza, el suministro de información del desempeño al compañero con base a la demostración de una rutina, o la indicación de las características críticas ejecutadas en un movimiento. Para que este tipo de valoración sea efectiva, el observador debe conocer el criterio de valoración y lo que debe observar para considerar que la técnica o destreza haya sido ejecutada en forma apropiada. Además, se debe enseñar a los estudiantes acerca de cómo observar, cómo formarse juicios apropiados, y cómo ofrecer la información a sus compañeros. También es imperativo que los estudiantes comprendan que la valoración que hacen de sus compañeros tiene la intención de suministrarles información destinada al mejoramiento del desempeño.

El modelo de valoración del docente es el elemento apropiado para comenzar a entrenar a los estudiantes en la medición del desempeño, pudiendo ser enseñados a jugar un rol de mayor impacto en la valoración del desempeño en ambientes reales. Por ejemplo, muchos de ellos sirven como anotadores estadísticos en las competencias deportivas que sirven como información para el entrenador y el atleta, hecho que convierten los instrumentos en un registro formal del desempeño en el juego y en un tipo de herramienta útil para que los estudiantes valoren la actuación de sus compañeros.

Cualquier instrumento de anotación que el docente utilice puede ser modificado de acuerdo al nivel de los estudiantes, y ser adaptado a la destreza o combinación de destrezas que se están utilizando en la competencia. Esos instrumentos necesitarían ser enseñados, modelados y practicados hasta que los estudiantes – observadores se sientan cómodos recolectando las estadísticas de juego y que sus compañeros sientan que los datos recolectados son precisos y reflejan su desempeño.

VALORACIÓN	DESCRIPCIÓN
5 – Excelente	Demuestra dominio de las destrezas específicas del deporte y habilidad para ejecutarlas consistentemente con poco o ningún esfuerzo consciente, cometiendo muy pocos errores. Tiene un extenso conocimiento básico y comprensión del deporte. Emplea estrategias específicas efectivas dentro de una tarea o situación.
4 – Bueno	Demuestra competencia y habilidad para ejecutar destrezas básicas sin cometer demasiados errores. Tiene una completa comprensión de las reglas y estrategias del deporte específico. Por lo general selecciona la destreza y estrategia apropiada en una situación, y regularmente muestra un desempeño consistente.
3 – Satisfactorio	Muestra un conocimiento básico del deporte y habilidad para ejecutar adecuadamente las destrezas fundamentales requeridas para jugar. El desempeño es frecuentemente inconsistente, cometiendo numerosos errores. Comprende las destrezas básicas, pero le falta habilidad para emplearlas efectivamente.
2 – Bajo	Demuestra falta de habilidad para ejecutar algo más que las destrezas básicas. Más aún, tiene dificultad en la ejecución de ciertas destrezas básicas y frecuentemente comete errores muy serios durante el desempeño. Por lo general el desempeño es inconsistente y muestra sólo una mínima comprensión de las estrategias y reglas.
1 – Deficiente	Rara vez ejecuta las destrezas lo suficientemente bien como para ser capaz de jugar en forma efectiva. Demuestra poca comprensión del deporte y falta de habilidad para ejecutar las destrezas sin cometer errores frecuentes y significativos. Efectúa pocos intentos para ajustar el desempeño.

Tabla 12.1. Escala descriptiva para la valoración holística de destrezas deportivas

VALORACIÓN	DESCRIPCIÓN
5 – Excelente	Demuestra dominio de las destrezas en tenis y habilidad para ejecutar consistentemente los golpes con poco o ningún esfuerzo consciente y con pocos errores no forzados. Tiene un extenso conocimiento de base. Anticipa los tiros del oponente y emplea una efectiva estrategia específica en la situación.
4 – Bueno	Demuestra competencia y habilidad para ejecutar las destrezas básicas del tenis sin cometer demasiados errores. Tiene una completa comprensión de las reglas y estrategias del tenis. Por lo general selecciona la estrategia apropiada y el tiro adecuado a la situación, mostrando un desempeño consistente.
3 – Satisfactorio	Muestra una comprensión básica del tenis y es capaz de ejecutar las destrezas fundamentales adecuadamente como para participar en competencias. El desempeño es frecuentemente inconsistente, cometiendo numerosos errores. Comprende las estrategias básicas pero le falta habilidad para emplearlas efectivamente.
2 – Bajo	Demuestra falta de habilidad para ejecutar algo más que las destrezas básicas. Más aún, tiene dificultad en la ejecución de ciertas destrezas básicas y frecuentemente comete errores muy serios durante el desempeño. Por lo general el desempeño es inconsistente y muestra sólo una mínima comprensión de las estrategias y reglas.
1 – Deficiente	Rara vez ejecuta las destrezas lo suficientemente bien como para ser capaz de jugar un tenis efectivo. Demuestra poca comprensión del tenis y falta de habilidad para ejecutar las destrezas sin cometer errores frecuentes y significativos. Efectúa pocos intentos para ajustar el desempeño.

Tabla 12.2 Escala descriptiva para la valoración holística de la habilidad para jugar tenis

VALORACIÓN	DESCRIPCIÓN
5 – Excelente	Tiene un agarre apropiado, con buen balance y trabajo de los pies, y una técnica cercana a la perfección. Demuestra consistentemente en la mecánica correcta para golpear. Los tiros son golpeados con potencia, y colocados apropiadamente en forma consistente.
4 – Bueno	Agarre apropiado, buen balance, adecuado trabajo de los pies y una aceptable, aunque no perfecta técnica. Demuestra una consistencia por arriba del promedio en la mecánica del golpe. Al golpear tiene una moderada potencia y una colocación consistente dentro del área de la cancha.
3 – Satisfactorio	Agarre apropiado, balance aceptable, pero el trabajo de los pies es pobre. La técnica es errática e ineficiente, resultando en una inconsistencia en la colocación de los tiros. El estilo para golpear es más defensivo pero puede sostener un peloteo de corta duración.
2 – Bajo	Utiliza un agarre inapropiado en muchas ocasiones, un pobre trabajo de los pies, y una técnica básicamente incorrecta. La mecánica del golpe es inconsistente. Estilo de juego defensivo, únicamente tratando de mantener la pelota sobre la malla. Baja capacidad para sostener un peloteo.

1 – Deficiente	Agarre incorrecto. Falta de balance y pobre trabajo de los pies. La técnica es muy pobre y errática. Virtualmente no controla la colocación de la pelota. Experimenta dificultades para mantener la pelota sobre la malla.
----------------	--

Tabla 12.3 Escala descriptiva del golpe de derecha y de revés en tenis

La caja de puntuación

La caja de puntuación es un método de registro para valorar la actuación del estudiante, que contiene una lista de componentes con los que se le valora durante la clase o la unidad. La caja debe presentar un mínimo de tres áreas: los elementos críticos, la revisión de los conceptos y las destrezas de trabajo independiente.

Los elementos críticos. Se relacionan con el desempeño real en las partes más importantes de la destreza que se enseña en clase. El docente debe decidir cuáles elementos críticos son los más importantes que el estudiante domine. Los objetivos de la valoración y el enfoque de la enseñanza variarán de una clase a otra, con base en el dominio psicomotor. Los elementos críticos se transforman en tareas que el docente presenta con un propósito definido y de manera progresiva, para que en forma gradual y en secuencia se conduzca a los estudiantes hacia un mejor y variado desempeño. Las progresiones pueden tener cuatro niveles: precontrol, control, utilización y competencia. Si el estudiante demuestra los elementos críticos en una o más de las características o destrezas de cada nivel obtiene cinco puntos en el nivel de competencia, cuatro puntos en el nivel de utilización, tres puntos en el nivel de control y dos puntos en el nivel de precontrol, y los que no reúnan ni una característica en el nivel de precontrol, pero participan en la actividad obtienen un punto. Un ejemplo de la progresión de las tareas de acuerdo con estos niveles lo tenemos con la técnica del drible (Tabla 12.4).

Revisión de los conceptos. Además de la demostración de la habilidad para ejecutar los elementos críticos de una destreza, se debe valorar la comprensión de los conceptos relacionados con la destreza. Esto es importante por cuanto siempre hay la posibilidad de que un estudiante no pueda ejecutar el movimiento en el nivel de competencia, pero puede comprender sus conceptos. Este criterio se basa en el dominio cognitivo. Para la valoración de los conceptos pueden implementarse dos criterios: (a) Un punto por respuesta, si responde con claridad, presenta información precisa y no requiere interpretación por parte del evaluador; (b) Medio punto por respuesta, si las respuestas requieren de clarificación o explicación justificativa por parte del estudiante, o que requieran de interpretación por parte del evaluador.

Es recomendable que se valore sólo uno de los dominios, el psicomotor o el cognitivo, al inicio del proceso de enseñanza – aprendizaje, porque la valoración de ambos dominios a la vez sobrecargaría al docente y los estudiantes. Por otra parte, la puntuación del dominio cognitivo es más fácil porque no debe ocurrir en tiempo real, por lo que el docente puede basarse en las respuestas del estudiante cuando inicie su valoración.

Nivel de precontrol (2)

Impulsa el balón hacia abajo y lo agarra
Impulsa el balón continuamente hacia abajo con ambas manos
Dribla con una mano

Nivel de control (3)

Dribla a diferentes alturas
Dribla continuamente mientras cambia de mano
Dribla con el cuerpo en diferentes posiciones
Dribla alrededor del cuerpo y en forma estacionaria
Dribla y conduce

Nivel de utilización (4)

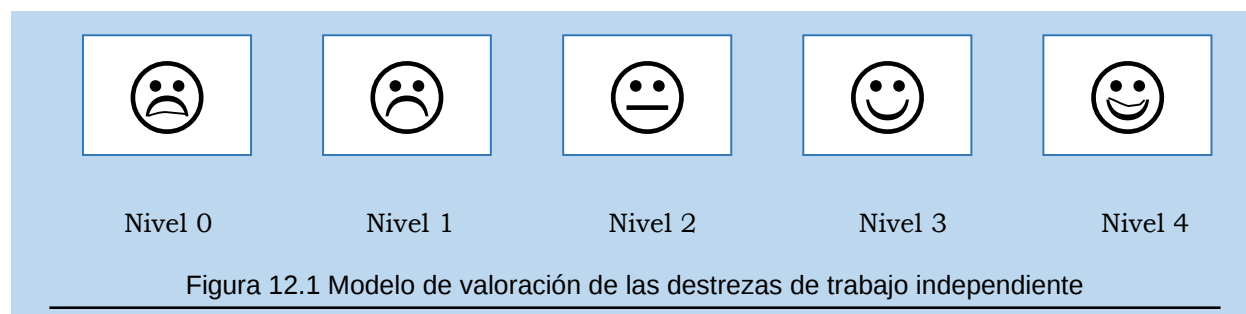
Dribla y cambia de velocidad al conducir
Dribla mientras cambia de dirección
Dribla por diferentes vías
Dribla alrededor de obstáculos
Dribla contra un oponente 1 vs. 1

Nivel de competencia (5)

Empieza y detiene el drible; cambia de dirección con drible rápido
Participa en juegos predeportivos driblando
Dribla y pasa en situaciones de competencia
Dribla y lanza hacia un blanco
Participa en juegos de pequeños grupos, utilizando el drible

Tabla 12.4 Desarrollo de las progresiones en la tarea de drible

Destrezas de trabajo independiente. Los estudiantes necesitan desarrollar las destrezas necesarias para aprender a lo largo de la vida. Quien aprende para toda la vida es capaz de trabajar solo y en compañía, buscar soluciones y permanecer concentrado en la tarea. Las destrezas de trabajo independiente pueden valorarse a través de la determinación del nivel en que se encuentra el estudiante, requiriéndose su observación continua (Figura 12.1).



Nivel 0: Doble ceño o enojo – El estudiante exhibe conductas que dificultan el aprendizaje de él y de otros. (Las conductas indisciplinadas entran en esta categoría.)

Nivel 1: Ceño o enojo simple – El estudiante exhibe conductas que le dificultan el aprendizaje, pero no interfiere con el aprendizaje de los demás estudiantes. Los que no participan o rehúyen participar entran en esta categoría.

Nivel 2: Más o menos – Ocasionalmente, el estudiante exhibe conductas que no se relacionan con la tarea. Necesita constantes recordatorios para que se concentre en su trabajo.

Nivel 3: Sonrisa simple – El estudiante exhibe conductas vinculadas a la tarea y trabaja continuamente sin intervención del docente.

Nivel 4: Doble sonrisa – El estudiante demuestra conductas relacionadas con la tarea, trabajando continuamente sin intervención del docente, y ayuda a sus compañeros de clase si lo necesitan.

El portafolio

El portafolio es la recolección sistemática, predeterminada y significativa del trabajo individual reunido a lo largo del período escolar. Es una poderosa herramienta que hace a los aprendices responsables de la demostración, monitoreo y exhibición de su propio aprendizaje, es decir, lo que ellos conocen y pueden hacer. En una clase que utilice el portafolio se notará que hay mucha colaboración, oportunidades para que los estudiantes revisen y/o reformulen su trabajo, los estudiantes y el docente reflejan el trabajo individual y del grupo como una totalidad, los estudiantes comprenden y aplican estándares claros para juzgar la calidad de su trabajo y de los otros, y muestran orgullo por lo realizado debido a su resultado, publicación o exhibición.

Desde esta perspectiva, la valoración a través del portafolio representa y fortalece el cambio de actitud hacia el rol de la evaluación en el aprendizaje. Los docentes que adoptan algún tipo de portafolio con su grupo cambian el énfasis de una evaluación basada en los resultados de pruebas prácticas, rangos percentiles, calificaciones o de rangos comparativos de logros, para dirigirse hacia el fortalecimiento del desempeño del estudiante por medio de la reflexión y la información.

Hay dos tipos básicos de portafolio: (a) El portafolio – modelo, consistente en la selección de muestras por medio de los cuales el estudiante documenta sus mejores trabajos; (b) El portafolio descriptivo, consistente en el trabajo general del estudiante, sin evaluar su producto. Se puede incluir otro tipo de información procesada durante el período, tal como los borradores de sus trabajos o el registro de su logro o progreso en el transcurso del lapso escolar.

El tipo, formato y contenido del portafolio por lo general es propuesto por el docente. Lo mejor es limitar el portafolio a un determinado número de trabajos, previendo

que no se convierta en una masiva colección de presentaciones con poco significado para el estudiante y que representa una monumental tarea de evaluación para el docente. Esto requiere que el estudiante tenga un buen juicio para determinar cuáles aspectos o tareas llenan los requisitos y demuestran su logro.

La valoración a través del portafolio es útil si: (a) En la evaluación se incluyen objetivos individuales y flexibles; por ejemplo, si la meta es el fortalecimiento de las destrezas sociales de los aprendices, se requeriría que algunos estudiantes sean menos agresivos mientras otros se hacen más asertivos; (b) El portafolio se ajusta a las necesidades y metas individuales; (c) Se reconoce que los estudiantes y sus programas están incluidos en decisiones para cambiar de actitud; (d) Se suministra información que le dé significado a las conductas y los cambios implícitos en ellas; (e) Se suministra una herramienta que asegure la responsabilidad y la comunicación hacia padres, comunidad, etc.; (f) Haya la posibilidad de valorar varias construcciones evaluativas y no sólo las que sean más fáciles de procesar.

La valoración a través del portafolio no será útil si el programa de evaluación se hace en función de metas rígidas, o se clasifica los participantes de una manera cuantitativa, o se compara a los estudiantes con normas estandarizadas.

La intención del portafolio es involucrar a los estudiantes en el proceso de valoración y que sientan su propiedad sobre el archivo o carpeta. El portafolio puede representar varios años de trabajo, un año escolar, una unidad de trabajo o una meta de aprendizaje. Por lo general, el docente establece la meta de aprendizaje y el estudiante decide lo que incluirá en el portafolio como evidencia de su trabajo hacia esa meta. También establecerá criterios claros acerca de cómo se valorará el portafolio. Para ello se puede diseñar una escala descriptiva flexible y que cada estudiante suministre las evidencias y la creatividad de su esfuerzo, limitando el número de ítems que se valorarán para que los estudiantes seleccionen aquello que mejor representan el logro de la meta.

Para seleccionar los materiales que reflejen su progreso, los estudiantes deben examinar sus trabajos, reflexionar lo que ellos representan, y valorarlos en función de la meta del portafolio. Este proceso de autoselección, reflexión y valoración hace del desarrollo del portafolio una herramienta instruccional. Sin embargo, el portafolio no es un instrumento de valoración hasta que no se haya considerado y negociado lo siguiente: (a) El propósito de la valoración; (b) La definición del cómo y el qué se seleccionará del material archivado; (c) Las decisiones sobre quién seleccionará los materiales del portafolio y cuándo puede efectuarse la selección y articulación del material; (d) Se identifique si los criterios de evaluación se aplicarán a la colección completa de materiales o a los aspectos particulares de la misma.

El portafolio en sí en un archivo que contiene la recolección del trabajo del estudiante. Los contenidos pueden incluir ítems sobre las siguientes formas de trabajo: diario de entrenamiento, periódico estudiantil, reportes escritos, fotografías, dibujos, escalas de estimación, resultados de la condición física, estadísticas de juego, planes de entrenamiento, reporte del análisis de la dieta alimenticia, videos, la puntuación en pruebas escri-

tas o proyectos que demuestren la comprensión de los conceptos, evidencias que demuestren la participación en actividad física fuera de clase, y diario sobre sus pensamientos, sentimientos y percepciones acerca de la actividad física. Los proyectos que podrían incluirse en el portafolio es casi ilimitado; una lista de algunos posibles proyectos de portafolios se presenta a continuación:

- Señale en forma escrita la autoevaluación de su actual nivel de destreza y de habilidad para jugar, con las respectivas metas individuales para su mejoramiento.
- Efectúe una progresiva autoevaluación y coevaluación del desempeño en las destrezas y en competencia, con listas de cotejo/chequeo del proceso y producto, escalas de estimación, estadísticas de juego, etc.
- Prepare un gráfico que muestre y explique el desempeño en las estrategias o destrezas en particular, a través del tiempo.
- Analice su desempeño en situaciones de competencia deportiva (aplicación tanto de destrezas como de las estrategias de juego), recolectando y analizando sus estadísticas de juego (por ejemplo, porcentaje de lanzamientos, asistencias, pases exitosos, etc.).
- Invente y ejecute un baile, una danza aeróbica, o una rutina gimnástica, aplicando sus conocimientos y destrezas. Entregue una rutina de ejercicios o una grabación en video.
- Documente su participación en la práctica deportiva, en competencias informales o en competencias organizadas, realizadas fuera de clase.
- Mantenga una relación diaria de la actividad física con las respectivas metas; registre su éxito, progreso y retroceso, y analice la situación para que se le hagan las recomendaciones para el trabajo presente y futuro.
- Con base en un autoanálisis o pre evaluación, seleccione o diseñe un apropiado programa práctico y el horario completo. Registre los resultados.
- Redacte un artículo periodístico como si fuese un periodista deportivo, reportando un torneo en la clase o una competencia deportiva. Se debe demostrar conocimiento del juego.
- Desarrolle y edite una revista sobre la clase de deportes o condición física.
- Programe, conduzca y participe en un torneo asignado al grupo. Lleve los registros y estadísticas individuales y del grupo.
- Entreviste un competidor exitoso acerca de su proceso de desarrollo como atleta y su técnica y programa actual de entrenamiento.
- Escriba un ensayo sobre el tópico “Qué he aprendido y logrado durante este período de actividad física y qué he aprendido de sí mismo durante el proceso”.

Los proyectos

Los proyectos se han usado en educación para juzgar la comprensión de un tema o tópico en particular, aunque su aplicación en educación física no está muy extendida. Aun cuando hay un infinito número de tareas que pudiesen ejecutarse con este propósito, por lo general los proyectos requieren que los estudiantes apliquen su conocimiento y destrezas para completar una tarea psicomotora.

Los proyectos requieren creatividad, pensamiento crítico, análisis y síntesis por parte del estudiante. Algunos ejemplos de proyectos efectuados en el ámbito del desempeño humano y la educación física son: redacción de un reporte de investigación sobre un tópico específico, entrevista a un atleta de la localidad o los árbitros, diseño de un programa personal de condición física, desarrollo de una grabación de audio o video, preparación de un póster para una presentación, análisis de las destrezas personales, creación de un nuevo juego o danza, y otros.

Los proyectos en grupo permiten que los estudiantes trabajen juntos en un problema complejo que requiere planificación, investigación, discusión interna y presentación. Aunque es difícil comprobar la contribución de cada miembro del grupo al producto final, los proyectos en grupo son atractivos porque estimulan y facilitan la cooperación, un valioso objetivo en el campo de la actividad física. Los proyectos se juzgan en concordancia con estándares de excelencia establecidos (los criterios para la puntuación) y que son conocidos por los participantes con anticipación.

El boletín

Los boletines (o cuaderno de notas) les permiten a los estudiantes reflejar y compartir sus pensamientos, sentimientos, impresiones, percepciones y actitudes acerca de su desempeño en eventos, asignaciones u otras experiencias de aprendizaje. Un boletín sirve como un medio para la descripción de una situación, reaccionando ante esa situación, reflexionando sobre sus acciones y utilizando dichas reflexiones para aprender y crecer. En este sentido, el boletín se convierte en una herramienta de aprendizaje útil para demostrar patrones, temas y relaciones que facilitan o inhiben su desarrollo. La redacción del boletín puede efectuarse en clase como una experiencia dentro de la unidad correspondiente y vinculada con asignaciones para el hogar.

Los boletines pueden tener gran variedad de formatos y propósitos, por ejemplo, ser muy generales y responder a “¿Cuáles fueron sus sentimientos acerca de...?” Los estudiantes, a través del boletín, pueden buscar respuestas a un conjunto de cuestiones específicas, o conformar un conjunto progresivo de reflexiones sobre un contenido o una experiencia. Para ayudar a desarrollar una sensación de confianza dentro de la actividad física, los estudiantes pueden ser estimulados a mantener durante el año escolar un boletín semanal en el cual señalen su desempeño en cada actividad e identifiquen aquella en la que se sintieron exitosos y porqué. En otro ejemplo, los estudiantes podrían esforzarse por lograr ciertas metas en cuanto a su condición física, anotando en un boletín su progreso, las barreras encontradas y sus sentimientos acerca de las metas alcanzadas. En todos los casos, el boletín implica reflexión: revisar los pensamientos, sentimientos y eventos de la experiencia de aprendizaje e interpretar lo que ello significa.

El contenido del boletín no se puede catalogar como correcto o incorrecto, y puede incluir ejemplos positivos y negativos. Dado que el boletín refleja sentimientos personales, pensamientos y percepciones, no pueden ser calificados por su contenido. El boletín suministra la visión del estudiante sobre su éxito o fracaso, así como información sobre cómo los estudiantes procesan la experiencia. Se valoran por estar completos o por utilizar

los criterios previstos para su redacción. Se debe tener cuidado con la puntuación del boletín porque puede haber la tendencia a reprimir la reflexión, honestidad y la expresión de los sentimientos personales y las percepciones.

Organización de eventos

La organización de un evento deportivo similar a la vida real permite que los problemas se aborden con múltiples soluciones y respuestas, lo cual es importante y relevante en el intento de estimular el interés del estudiante, y que puede ser completada en un tiempo de juego. Esto pudiese verificarse en un evento de culminación de la temporada de gimnasia; el éxito logrado por un equipo a través de actividades externas a la escuela; la organización, implementación y monitoreo de actividades para el bienestar de la comunidad, o que cada estudiante diseñe un juego basado en los conceptos de golpear objetos, velocidad y precisión. Aunque la organización de un evento puede ejecutarse en un período o lapso escolar, la planificación puede tomar mucho tiempo fuera de clase por lo que su organización se realizaría en forma individual o en grupo. Su valoración puede realizarse al momento de la presentación o de una grabación en video realizada por el docente, un compañero, e incluso a través de una autoevaluación. Asimismo, la organización de un evento puede valorarse por medio de una escala descriptiva lo cual significa que el docente debe tomar decisiones acerca de las dimensiones del desempeño que son importantes y luego establecer los diferentes niveles de desempeño para cada dimensión.

La organización de un evento bien diseñado presenta las siguientes características: refleja los productos del aprendizaje, suministra oportunidades para que se apliquen conocimientos y destrezas en una tarea significativa, permite que se demuestre el nivel de mejoramiento y logro, utiliza aspectos relacionados con la vida real e integra el conocimiento hasta donde sea posible. Suministrarles oportunidades a los estudiantes para que apliquen lo que han aprendido en clase no es difícil para un docente de educación física en cualquier nivel educativo.

Exhibiciones

Una exhibición es una muestra o desempeño público que constituye una vitrina de las competencias y aprendizajes que alcance el estudiante en alguna área en particular. La exhibición puede ser evaluada con una escala descriptiva desarrollada específicamente para este tipo de desempeño o muestra. Por otra parte, además de ser evaluado por el docente del curso, las decisiones que tomen los estudiantes también pueden ser valoradas por un panel de expertos adultos o por sus propios compañeros. Entre las exhibiciones pueden estar las siguientes:

- Ejecución de una rutina de ejercicios aeróbicos en una asamblea escolar.
- Organización y ejecución de una demostración de salto de cuerda en la mitad de un encuentro de baloncesto.

- Ejecución de una danza folklórica en la feria municipal.
- Demostración de algún tipo de arte marcial en un centro comercial.
- Diseño de un afiche sobre los beneficios de la actividad física para exhibirlo en una feria anual de condición física.
- Dibujar y pintar temas sobre varias actividades deportivas, para mostrarlos en locales comerciales y restaurantes

Diseño de la evaluación cualitativa

La cantidad y variedad de instrumentos mediante los cuales puede valorarse al estudiante en educación física es abrumadora, y la factibilidad de su aplicación es un requisito que debe considerarse en la evaluación cualitativa. Dada esta situación, en primer lugar, la valoración se facilitará si el docente puede describir como un buen desempeño del estudiante lo que se planificó como un resultado del aprendizaje. Por tanto, se deben redactar los criterios de desempeño y luego pensar en cómo valorarlos. El aspecto crítico es la cuidadosa adecuación de la valoración con respecto a los resultados que el docente previó al planificar el aprendizaje, suministrándoles a los estudiantes la oportunidad de demostrar su progreso hacia el logro de esos resultados. El docente debe responder una serie de interrogantes cuyas respuestas le guían hacia el desarrollo de unas efectivas tareas de valoración:

- ¿La valoración se adecua a los planteamientos realizados durante el proceso de enseñanza – aprendizaje?
- ¿La valoración representa apropiadamente el contenido y destrezas que se espera que los estudiantes logren?
- ¿La valoración permite que los estudiantes demuestren su progreso y capacidades?
- ¿La valoración utiliza tareas auténticas, propias del mundo real?
- ¿La valoración presenta un enfoque interdisciplinario?
- ¿La valoración puede ser estructurada de forma tal que facilite la medición de varios objetivos?

Es importante que los docentes generen sus propios instrumentos de evaluación cualitativa, sin que signifique que no se puedan adaptar/aplicar ideas obtenidas de otras fuentes, sino que los docentes necesitan crear instrumentos de evaluación cualitativa acordes con las necesidades específicas de su grupo. Una orientación para el desarrollo de instrumentos de evaluación cualitativa es que un instrumento constituye tanto una tarea como un criterio de desempeño. Con esto en la mente, el proceso de desarrollo de instrumentos significativos de evaluación cualitativa lleva los siguientes pasos:

Determinación del propósito de la valoración. El primer aspecto crítico del proceso de valoración es la determinación precisa de su propósito. Como ejemplos de pro-pósito tenemos: (a) Diagnosticar o determinar las deficiencias en el desempeño de un individuo o del producto que ha creado; (b) Comprobar el logro de los objetivos específicos para la aprobación de un grado; (c) Obtener información para la evaluación del programa; (d) Expresar los requisitos alcanzados por el estudiante, elaborando el informe respectivo.

Acá se reconoce la importancia de la evaluación y las decisiones que se tomarán con base en los resultados, precisándose quién utilizará los resultados.

Definición de los objetivos. El docente debe señalar lo que se espera del proceso de enseñanza – aprendizaje antes de iniciarlo, definiendo con claridad los resultados previstos: ¿Qué deben conocer los estudiantes y qué deben ser capaces de hacer al final del proceso?

Selección de la apropiada tarea de valoración. La clave de una valoración apropiada es la adecuación de la tarea y del instrumento de valoración a los objetivos previstos. Si no se adecuan a los resultados buscados, entonces no son útiles para los propósitos de la evaluación y los resultados no serán significativos para la toma de decisiones. El docente debe preguntarse si la valoración está acorde con el desempeño buscado, o si los estudiantes han tenido la oportunidad de adquirir el conocimiento y las destrezas exigidas por la tarea o instrumento de valoración. Si no es así, debe cambiarse la tarea o el instrumento de valoración.

Establecimiento del criterio de desempeño. El criterio de desempeño es tan necesario e importante como el instrumento en sí. La sola tarea o instrumento no constituye una valoración; también se debe tener un criterio mediante el cual ser capaz de juzgar el desempeño o producto. Muchos docentes tienden a concentrarse solo en la tarea, excluyendo el criterio. ¿Cómo se conocerá si un individuo ha logrado la meta u objetivo previsto? ¿Cómo se conocerá si un desempeño es lo suficientemente bueno? ¿Cómo determinará el docente si un portafolio es aceptable? ¿Con base en cuáles elementos se diagnosticará el desempeño de los estudiantes y se suministrará una información significativa sobre el desempeño? Cada una de estas preguntas es respondida por el criterio, no por la tarea o instrumento de valoración.

Determinación de la calidad de la valoración. Cuando se habla de la calidad de la valoración, nos debemos referir a la confiabilidad, validez e imparcialidad de los instrumentos y tareas de evaluación. En la evaluación cualitativa se debe ser particularmente cuidadoso y evitar las dificultades o trampas que amenacen la confiabilidad, validez e imparcialidad, o que lleven a tomar decisiones inapropiadas y consecuencias imprevistas.

Fiabilidad y validez de la evaluación cualitativa

Ya se ha mencionado que los instrumentos tradicionales de evaluación, tal como las pruebas de selección múltiple o las pruebas de destrezas deportivas, ya no son suficientes para realizar una valoración apropiada en el cambiante ámbito educacional, caracterizado por el énfasis en los productos del aprendizaje, en las destrezas cognitivas de orden superior y en el aprendizaje integrado. Esos instrumentos y técnicas tienden a medir unas características sin facilitar la integración de las destrezas en el proceso, aplicándolos en ambientes artificiales. Todo ello ha contribuido a la implementación de instrumentos de evaluación cualitativa. De allí que ahora se reconozca la destreza del estudiante como un desempeño de gran importancia para el aprendizaje, sin que se parezca a las pruebas estandarizadas tradicionales cuyo valor se deriva de su relación con los resultados de otros estudiantes ya valorados.

Por otro lado, si se utiliza algún instrumento de valoración alternativa para ofrecer información del desempeño a los aprendices, informar sobre el proceso de enseñanza – aprendizaje o reflejar el progreso del estudiante, la medición debe ser fiable y consistente. Esto quiere decir que el docente debe otorgar la misma puntuación en intentos diferentes y continuos o, si dos personas están observando el mismo desempeño, ambos deben dar el mismo puntaje. Esto se ha convertido en un gran problema para la aplicación de la valoración alternativa, por la confianza que se debe generar de los juicios emitidos por el docente, ya que se debe tratar equitativamente a todos los estudiantes.

Mientras se aboga en forma intuitiva por la aplicación de diversas formas de valoración alternativa, en realidad hay poca evidencia de fondo que justifique la superioridad de las técnicas de evaluación cualitativa ante las formas generalizadas de evaluación estandarizada. Aun así, la popularidad de los métodos de valoración alternativa se ha elevado dramáticamente en los años recientes. Antes de emitir un juicio apresurado acerca de la utilidad y propiedad de los instrumentos de valoración alternativa, es importante considerar su calidad técnica.

Se debe tomar en cuenta que el propósito de la valoración es suministrar información válida para la toma de decisiones. Un instrumento de valoración apropiado provee un indicador preciso del desempeño de los individuos y faculta al docente, al entrenador y a los padres para que emitan un juicio sobre el desempeño del estudiante, pero ¿qué constituye un buen instrumento de valoración? Cuando se habla de la calidad del instrumento, realmente se están formulando las siguientes preguntas: (a) ¿El instrumento de valoración ofrece una información precisa para la toma de decisiones? (b) ¿Los resultados permiten unas conclusiones precisas y justas acerca del desempeño del atleta o del estudiante? (c) ¿La utilización de los resultados contribuye a tomar decisiones acertadas? Para responder sí a estas preguntas, deben cumplirse tres criterios: fiabilidad, validez y equidad o imparcialidad.

Fiabilidad. A través de este criterio se comprueba la consistencia de los puntajes y de las observaciones. Un puntaje sin fiabilidad es inútil porque no le suministra información significativa al usuario. Por ejemplo, si un estudiante obtiene un puntaje un día, ¿se le dará el mismo puntaje por el mismo desempeño otro día? Si son dos personas las que observan y registran el desempeño del estudiante, ¿ellas le otorgarán el mismo puntaje al estudiante? La fiabilidad de una medición generalmente se obtiene mediante el procedimiento del test-retest; por lo tanto, ¿el estudiante obtendrá el mismo puntaje en el retest?

Cuando se recolectan los datos, la principal preocupación relacionada con la fiabilidad es el acuerdo entre los observadores. Lo que se ha llamado objetividad de los observadores refiere que, si dos personas están mirando el mismo desempeño, deben anotar el resultado de la misma manera. Cuando unos observadores no demuestran un elevado nivel de acuerdo en sus conclusiones, por lo general los criterios de puntuación o anotación no fueron claramente establecidos, o uno de los observadores es más eficiente al utilizar los criterios de valoración.

En la tabla 12.5 podemos ver un ejemplo de ello. Los observadores 1 y 2 están de acuerdo en la primera y segunda observación. Luego dieron una puntuación casi idéntica en la observación 3 pero muy diferente en la observación 4. En este caso, se podría asumir que están utilizando criterios diferentes para valorar el cuarto aspecto. El problema puede estar en los elementos que forman parte del instrumento utilizado para observar, o en la interpretación de esos elementos por parte de uno o ambos observadores.

	Observador 1	Observador 2
Observación 1	4	4
Observación 2	3	3
Observación 3	9	8
Observación 4	2	6

Tabla 12.5 Ejemplo de objetividad

La mayoría de las nuevas estrategias de valoración se basan en la observación del desempeño real de los estudiantes. La fiabilidad de estas mediciones es el aspecto clave para decidir qué hacer con la información recolectada con esos instrumentos de valoración, dados los problemas para la obtención de la fiabilidad al observarse conductas complejas. Mientras más sencilla sea la conducta a observar, más fácil es determinar su fiabilidad.

Puesto que la valoración alternativa depende en alto grado del juicio subjetivo del docente, éste debe preocuparse por determinar la objetividad o fiabilidad entre dos observadores, procurando que un observador juzgue en forma similar a otro observador. Considerando la naturaleza subjetiva de la valoración alternativa el docente debe esforzarse por minimizar la inconsistencia de su puntuación, para obtener confianza sobre su juicio acerca del resultado del desempeño real, sin elementos extraños que lo contaminen ya que las decisiones sobre la demostración de un individuo no son válidas a menos que se basen en información fiable. Para asegurar que durante la valoración alternativa se otorgue una puntuación fiable, se deben cumplir tres condiciones: (a) Un criterio claro y explícito para juzgar el desempeño o resultado obtenido; (b) Los que emiten un juicio (docentes, compañeros, entrenadores) comprenden ese criterio; (c) Los que emiten un juicio han aprendido a aplicar el criterio de puntuación de una manera consistente. La aplicación uniforme del criterio de puntuación es el mejor logro del entrenamiento y la práctica.

Validez. Es un indicador de cuán bien una valoración mide realmente lo que se supone que mide. Una valoración puede ser perfectamente fiable, pero si no se relaciona con las características fijadas para medir, puede comprometerse la precisión de las conclusiones acerca del desempeño de los individuos y las subsecuentes decisiones. Desafortunadamente, hay poca evidencia empírica acerca de la validez de los instrumentos de valoración alternativa. La forma primaria de validez que se ha reportado de las valoraciones alternativas, particularmente aquellas desarrolladas para su uso en el salón de clases, la validez de contenido, se basa en la relación entre la instrucción y la evaluación. Aunque este vínculo es importante, la sola validez de contenido no puede aceptarse como evidencia

suficiente para que se utilice un instrumento de valoración. Se necesita corroborar este tipo de evidencias para que las valoraciones alternativas produzcan conclusiones válidas.

Un ejemplo de validez es la relación entre una prueba de lanzamiento de tiro libre en baloncesto y la calidad de juego de los sujetos. Si un jugador es un excelente lanzador de tiro libre pero no es un buen jugador, entonces la prueba de tiro libre podría ser válida para valorar dicha técnica, pero no es válida para la determinación de cuán bien se juega al baloncesto. Por otra parte, para valorar si los estudiantes son capaces de cooperar en clase, el docente debe tener la certeza de lo que es y lo que no es una conducta cooperativa. La precisión de las conductas incluidas en esta categoría, la cooperación, debería reflejarse en la validez del instrumento de valoración.

Tradicionalmente, el tipo de validez más utilizado en educación física es la validez de contenido mediante el cual el docente define los elementos críticos de lo que se desea medir y luego los articula en el instrumento correspondiente, como se aprecia al utilizar una lista de chequeo para observar una destreza de movimiento o una conducta afectiva.

Otra forma es la validez de criterio, mediante la cual se determina si el puntaje obtenido en una prueba se corresponde con el puntaje en otra prueba que ya se ha mostrado como válida. Los estudiantes que salen bien en una prueba deberían hacerlo bien en otra, y quien lo haga mal en una debe salir mal en la segunda. Por ello se debe esperar que los resultados de una carrera de una milla y de una prueba de correr 12 minutos estén altamente correlacionados. Si los mejores jugadores de baloncesto (los que ganan) salen bien en una prueba de lanzamiento y los jugadores que no ganan fallan o no anotan tan bien como los primeros, entonces la prueba es una medida válida sobre el dominio del juego en baloncesto.

Sin embargo, puesto que muchos instrumentos de valoración alternativa son sustancialmente diferentes al tipo de pruebas estandarizadas, los procedimientos tradicionales de validación pueden ser inapropiados. A este respecto, especialistas en medición han sugerido algunos criterios para la determinación de la validez de los instrumentos de valoración alternativa entre los cuales encontramos:

Consecuencias: ¿La valoración conduce a las consecuencias previstas? ¿Cuáles son las consecuencias no previstas?

Imparcialidad: ¿La valoración permite que todos los estudiantes demuestren sus capacidades reales, o hay desventajas injustas para algunos estudiantes?

Transferencia y generalización: ¿Los resultados de la valoración son generalizables a otros problemas o situaciones del dominio psicomotor, afectivo o cognitivo?

Complejidad cognitiva: ¿La valoración considera adecuadamente los altos niveles de comprensión y el pensamiento complejo?

Calidad del contenido: ¿Se ha seleccionado un instrumento de valoración que mida un contenido de forma tal que merezca el tiempo y el esfuerzo del estudiante y sus asesores?

Cobertura del contenido: ¿La valoración prueba adecuadamente la amplitud del contenido del dominio psicomotor?

Significado: ¿El instrumento de valoración tiene significado para los estudiantes y los estimula a ejecutarlo mejor?

Costo y eficiencia: ¿El costo de la recolección y anotación de datos es razonable y eficiente?

Imparcialidad. La equidad o imparcialidad es una característica de importancia crítica en toda forma de evaluación, ya sea tradicional o alternativa. La imparcialidad significa que la valoración les ofrece a todos los estudiantes la misma oportunidad de salir bien, independientemente del sexo, raza o experiencia. La imparcialidad debe ser evidente tanto en el desarrollo o selección del instrumento de valoración como en el criterio utilizado para juzgar el desempeño, el resultado o producto.

¿El instrumento de valoración es transparente y libre de sesgos? ¿El instrumento de valoración favorece a niños o niñas, a los estudiantes de un grupo étnico en particular, a quienes viven en un lugar en particular, o aquellos cuya familia tiene mayores recursos financieros? Para que sea transparente, el instrumento debe reflejar el conocimiento, valores y experiencias que sean familiares y apropiadas para todos los estudiantes, y deben medir el conocimiento y destrezas que han tenido tiempo de adquirir. Además, es importante que los procedimientos de puntuación y el criterio para juzgar el desempeño o el producto creado estén libres de sesgos. Esto asegura que los puntajes alcanzados durante el desempeño reflejen una verdadera revisión de la capacidad y que no dependan de las tendencias y percepciones de la persona que juzga ese desempeño.

La evaluación tradicional tiende a demostrar mayor fiabilidad, objetividad y validez que la valoración alternativa, debido a la naturaleza más objetiva de la primera. Por ello se necesita repensar la calidad técnica utilizada para juzgar la validez de los instrumentos de valoración alternativa. La calidad técnica es tan importante en la valoración alternativa o evaluación cualitativa como en una evaluación estandarizada tradicional. Mientras se logra la construcción de instrumentos de valoración alternativa, el reto del docente que practica este tipo de valoración es reflexionar y considerar cuidadosamente los principios de validez, fiabilidad e imparcialidad, asegurándose que cualquier instrumento de valoración que se utilice conduzca a los juicios correctos y las decisiones apropiadas.

Técnicas estadísticas para la fiabilidad y validez. En la evaluación cualitativa es importante que se utilicen las técnicas estadísticas propias del tipo de evaluación referida a criterios. Dado que los instrumentos en la valoración cualitativa se basan especialmente en la observación que hace el docente del progreso del estudiante, la escala que se utiliza es de carácter nominal. Los datos de la observación se deben categorizar identificando los que dominan y los que no dominan el tema o destreza, es decir, los que están por arriba o por debajo del punto de corte; este punto de corte es el puntaje mínimo requerido para considerar que se domina el tema o destreza. El punto de corte es importante cuando la comparación del desempeño individual no es la preocupación principal, sino que lo fundamental es la determinación del nivel de desempeño logrado. Como un ejemplo, un punto de corte lo constituye el nivel de colesterol en la sangre. Si es mayor a 200 mg/dl hay un riesgo para la salud del sujeto, y si es menor a 200 existe un bajo riesgo de enfermedades coronarias; en este caso 200 mg/dl es el punto de corte establecido por los expertos en el área.

La fiabilidad referida a un criterio se define como la consistencia de la realización. Si un puntaje en un test referido a un criterio es fiable, una persona tendrá la misma realización en ocasiones repetidas, ya sea en dos intentos en un día o en días diferentes. Para determinar la fiabilidad y validez de los instrumentos de evaluación cualitativa se utiliza una tabla de contingencia 2 x 2, identificando y ubicando los que dominan y no dominan en la respectiva celda.

		Día 1		<i>Filas marginales</i>
		Domina	No domina	
Día 2	Domina	$n1$	$n2$	$n1 + n2$
	No domina	$n3$	$n4$	$n3 + n4$
<i>Columnas marginales</i>		$n1 + n3$	$n2 + n4$	

Tabla 12.6 Tabla de contingencia 2 x 2 para una prueba con base a criterios, ejecutada en dos días.

Dos técnicas sencillas, el coeficiente de la proporción del acuerdo (P) y el coeficiente Kappa (K), pueden ser interpretadas como coeficientes de correlación utilizadas para procesar los datos de la medición con una escala nominal. El valor de la proporción del acuerdo (P) se establece al sumar las proporciones o respuestas de las celdas que están correctamente clasificadas; así, P es igual al número de acuerdos ($n1 + n4$) dividida por la cantidad total de respuestas ($n1 + n2 + n3 + n4$), por lo cual se establece la siguiente fórmula:

$$P = (n1 + n4) / (n1 + n2 + n3 + n4)$$

El rango de P va de 0 a 1.00 y mientras más alto sea el valor, están los datos más cerca de haber sido correctamente asignados a las celdas. El problema con P es que los valores por arriba de 0.50 pudiesen producirse simplemente por azar.

El valor Kappa (K) es una técnica ampliamente utilizada por cuanto permite la corrección del azar de los acuerdos. K ha sido muy utilizado para valorar los acuerdos entre observadores, pero puede ser utilizado en situaciones de test – retest o pruebas repetidas, o para examinar el acuerdo entre criterios de escalas nominales. Aunque la proporción del acuerdo es un estimado aproximado del acuerdo o asociación entre dos variables nominales, el principal problema con esta técnica estadística es que no considera el hecho

de que algunos de esos acuerdos pudiesen haberse producido por azar. K toma en cuenta este azar y produce un estimado más conservador de la asociación entre dos variables nominales. La fórmula de K es:

$$K = (P_o - P_c) / (1 - P_c)$$

P_o es la proporción del acuerdo observada y P_c es la proporción del acuerdo debida al azar. En el siguiente ejemplo, 400 estudiantes de educación básica ejecutaron la prueba de correr 1 milla en dos oportunidades y días diferentes. El instructor desea conocer si la prueba mide la habilidad para lograr alcanzar el punto de corte o puntaje mínimo para aprobar el test. Los datos los tenemos en la siguiente tabla:

		Día 1		<i>Total</i>
		Alcanza- ron el punto de corte	No alcanza- ron el punto de corte	
Día 2	Alcanza- ron el punto de corte	250	20	270
	No al- canza- ron el punto de corte	50	80	130
<i>Total</i>		300	100	400

Tabla 12.7 Resultados de la aplicación de la tabla de contingencia

P_o se calcula con la fórmula ya señalada de la proporción del acuerdo (P), donde $n_1 = 250$; $n_2 = 20$; $n_3 = 50$; $n_4 = 80$. El total es 400. El cálculo se realiza de la siguiente manera:

$$P_o = (250 + 80) / 400 = 330 / 400 = 0,825$$

K se calcula para corregir el azar y da un resultado más conservador. De acuerdo con P_c de la fórmula de K , primero se multiplican ambos marginales del dominio y el resultado se divide entre el total al cuadrado (n^2), y luego se multiplican los marginales de no dominio e igualmente el resultado se divide entre n^2 ; luego se suman ambas proporciones para determinar P_c :

$$(270 \times 300) / (400 \times 400) = 0,506$$

$$(130 \times 100) / (400 \times 400) = 0,081$$

Entonces,

$$P_c = 0,506 + 0,081 = 0,587$$

Luego se procede a calcular K , conocidos ya los valores de P_o y P_c :

$$K = (0.825 - 0.587) / (1 - 0.587) = 0,283 / 0,413 = 0,576$$

Dado que el coeficiente de la proporción del acuerdo no toma en cuenta el azar, es lo que se corrige a través del coeficiente Kappa. En el ejemplo el coeficiente de la proporción del acuerdo es de 0,825 sin tomar en cuenta el azar. Al calcular el coeficiente Kappa y eliminar las actuaciones debidas al azar, tenemos un coeficiente de 0,576 observándose una amplia diferencia entre ambos coeficientes, por lo que las actuaciones del grupo dicen que hay una consistencia entre el primer intento y el segundo aproximado de un 58% y que la diferencia se debió al azar, lo cual es un resultado aceptable. Dado que P es afectado por el azar los resultados menores a 0,50 son inaceptables por lo cual debe estar entre 0,50 y 1,00 para ser aceptable. Igualmente, K debe estar más cercano a 1,00 que a 0 para ser aceptable.

ANEXOS

Tomado de: Baumgartner, T., Jackson, A., Mahar, M., & Rowe, D. (2007). *Measurement for evaluation in physical education & exercise science*. Human Kinetics. USA.

CAPÍTULO 8

La calificación: una evaluación sumativa

Tom Jones está listo para enseñar la unidad de baloncesto en la clase de Educación Física de bachillerato. Él quiere darle igual peso a los aspectos cognitivos y psicomotores de la unidad. Con base en su experiencia, Tom planifica hacer dos pruebas escritas durante la unidad para medir los objetivos cognitivos; el primer examen conformará el 20% de la calificación final y el examen final valdrá el 30% de la misma calificación. También planifica realizar dos pruebas de destrezas, lanzamiento y conducción, para medir los objetivos psicomotores de la unidad. La prueba de lanzamiento tendrá un peso del 35% de la calificación final, y la prueba de drible tendrá el 15%. ¿Cuáles pasos debe seguir Tom para asegurar que el puntaje compuesto final de cada estudiante refleje este criterio?

Este capítulo está dirigido a los estudiantes cuyo plan es formarse como docente de educación física, una profesión que requiere la valoración del desempeño del estudiante. La valoración del desempeño del estudiante en las escuelas públicas y privadas generalmente culmina con el reporte de una calificación. El proceso de calificar puede ser un asunto desafiante para muchos docentes principiantes, y aun veteranos, a los que no les gusta hacerlo. Aunque los docentes y estudiantes pudiesen disfrutar de dar y recibir buenas calificaciones, lo probable es que no les gusta dar y recibir bajas calificaciones. Sin embargo, los docentes prospectivos deben comprender que la asignación de calificaciones es una obligación profesional y un importante componente de la enseñanza de la educación física.

Después de estudiar este capítulo, usted será capaz de:

- *señalar los criterios apropiados para la asignación de calificaciones,*
- *ilustrar los métodos para la asignación de calificaciones, y*
- *utilizar los métodos para la asignación de calificaciones.*

Hay muchos usos de las calificaciones, incluyendo los siguientes:

- *motivación y orientación del aprendizaje y de los planes vocacionales y educacionales de los estudiantes, así como su desarrollo personal,*
- *comunicación con los estudiantes y los padres acerca del progreso del estudiante,*
- *condensación de la información para usarla en la determinación de admisión en las universidades, admisión en la escuela para graduados, y el recibimiento de becas,*
- *comunicación a posibles empleadores de las fortalezas y limitaciones de los estudiantes, y*
- *evaluar a los estudiantes para ayudar a la escuela a adaptar la educación a las necesidades e intereses de los estudiantes y evaluar la efectividad de los métodos de enseñanza.*

Además de las dificultades menores que ocurren durante el proceso de determinación de las calificaciones, es necesario considerar que las calificaciones, las cuales son juicios de una persona sobre otras, se basan en evidencias subjetivas. En ocasiones el volumen de subjetividad incluida en un conjunto de calificaciones es grande. Las calificaciones basadas en la valoración de las respuestas de un estudiante en un examen tipo ensayo y la participación en las discusiones en la clase son ejemplos de tales situaciones. Incluso las calificaciones obtenidas a través de métodos aparentemente precisos y objetivos a menudo implicaban más subjetividad de lo que se podría pensar. “Las pruebas objetivas” (verdadero – falso, selección múltiple, y otras) requieren de algunas decisiones subjetivas, tal como la determinación de cuáles ítems incluir en la prueba y cuál respuesta es la más correcta para cada ítem. A menudo los instructores crean una fórmula precisa para la combinación segura de los logros del estudiante, pero esa fórmula se basa generalmente en el peso subjetivo de la importancia de los objetivos del curso, similar a los que Tom Jones seleccionó en el escenario descrito al comienzo de este capítulo. ***Así, aunque es verdad que algunas calificaciones involucran una mayor subjetividad que otras, también es verdad que todas las calificaciones involucran algún grado de subjetividad.***

Generalmente, a mayor grado de subjetividad implicada en el sistema de calificación, mayor falta de fiabilidad. En otras palabras, un sistema subjetivo, si se repite, probablemente no asignaría las mismas calificaciones al mismo estudiante. La falta de objetividad y la consecuente pérdida de fiabilidad llevan a una segunda y más seria preocupación acerca de la calificación: la falta de una comprensión común de lo que representa una calificación en particular. ***Las calificaciones son afectadas por el tipo de anotación utilizado, el instructor particular que asigna la calificación, la composición de la clase, la institución donde se otorga la calificación, y muchos otros factores relacionados.*** Por ejemplo, una E podría parecer excelente en un sistema de calificación, pero un fracaso en otro; una calificación de B de un instructor que rara vez da una B tiene un significado diferente a la B que rara vez da una calificación baja; una A obtenida en una clase de estudiantes superiores puede representar un logro más significativo que una A recibida en una clase menos competitiva, y a cierto nivel de desempeño podría dársele una A en una institución, pero una calificación baja en otra.

Mientras el uso de evidencia objetiva insuficiente reduce la fiabilidad de las calificaciones, la falta de una clara y generalmente aceptada definición de lo que una calificación en particular representa afecta la validez de esas calificaciones. Las calificaciones que deberían reflejar el grado en el cual un individuo ha logrado los objetivos o metas del curso, a menudo son contaminadas por muchos otros factores.

Ninguna regla específica para calificar puede ser establecida ya que cualquier situación en la cual las calificaciones se asignen es diferente para algunos con respecto a cualquier otro. Las diferencias existen en áreas tales como técnicas de enseñanza, objetivos del curso, equipos y materiales disponibles, y tipos de estudiantes en la clase. Sin embargo, un efectivo y consistente proceso de calificación le permite a u instructor y sus estudiantes tener confianza en la validez de las calificaciones asignadas. Si un estudiante comprende y acepta el sistema de calificación del docente, no se sentirá feliz por una baja calificación, pero puede aceptar que está fracasando. ¿Te has sentido infeliz cuando recibes una calificación diferente a la que esperabas? ¿Fue más fácil aceptar una baja calificación si sentías que el método de calificación del docente fue justa para todos los estudiantes? El desafío de este capítulo es

ofrecerles los conocimientos y destrezas requeridas para desarrollar y utilizar buenas formas de calificación en su programa instruccional, si empleara esa práctica en un ambiente académico o en cualquier situación similar que requiera una valoración final del grado para el cual los objetivos han sido planteados.

Estándares y evaluaciones

En el capítulo 1 definimos varios términos importantes para el proceso de evaluación y medición. Diferenciamos entre una **evaluación formativa**, conducida durante un programa instruccional o de entrenamiento, y una **evaluación sumativa**, un juicio integral final, cerca de la culminación del programa instruccional o de entrenamiento. Las *evaluaciones* resultan del proceso de toma de decisiones mediante el cual se emite un juicio sobre la calidad de una medición. En la instrucción de la actividad física, los docentes hacen evaluaciones formativas en el inicio y durante el proceso instruccional, y las utilizan para detectar las debilidades del avance del estudiante y para dirigir futuras actividades de aprendizaje. Una evaluación formativa puede ser una actividad de medición formal, tal como un pretest previo a la instrucción, o una muy informal, una evaluación subjetiva dada por el instructor (tal como un feedback verbal durante la práctica del tenis). Este capítulo se enfoca en las evaluaciones sumativas formales incluyendo los pasos apropiados para la conducción de tales evaluaciones, lo cual resulta en calificación definitiva que es determinada por el desempeño general de los estudiantes en la unidad instruccional completa, o por el logro final de los temas y objetivos presentados.

Para hacer un juicio sobre la calidad del desempeño, este debe ser comparado con un estándar. En los capítulos 1, 6 y 7 discutimos los estándares para la evaluación referida a normas y referida a criterios. Para una breve revisión, un estándar referido a normas se establece comparando el desempeño individual con el desempeño de otros individuos de la misma edad y género. El establecimiento de este estándar usualmente requiere de algún tipo de análisis de datos. Por otro lado, un estándar referido a criterios es un nivel de desempeño predeterminado y específico que ha sido establecido – a partir de una base de datos anterior o de la opinión de expertos – antes que por la ejecución individual. En una evaluación referida a criterios logra el estándar (pasa o aprueba) o no lo logra (fracasa). En una evaluación referida a normas, los individuos son clasificados de excelente a bajo con base a su posición en la comparación entre los puntajes originales.

El proceso de calificación

¿En qué pueden ser calificados los estudiantes o las materias? En educación física y el desempeño humano hay tres grandes dominios de potenciales objetivos:

El **dominio psicomotor**, concerniente al desempeño físico.

El **dominio cognitivo**, concerniente al desempeño mental.

El **dominio afectivo**, concerniente a las actitudes y aspectos psicológicos (también llamado el dominio psicológico).

Los docentes de materias como matemáticas, ciencia, inglés o historia, han limitado los objetivos instruccionales mayoritariamente al dominio cognitivo. Así, en cierto sentido, su proceso de evaluación es más simple que el de un instructor de educación física. Para evaluar a los estudiantes efectivamente, un instructor debe tener una clara comprensión de los objetivos instruccionales que está enseñando. Se deben seleccionar y administrar las pruebas y mediciones que son relevantes para esos objetivos. El instructor entonces debe comparar los puntajes resultantes de la prueba con los estándares apropiados y, finalmente, determinar las calificaciones. La figura 8.1 ilustra este continuo.

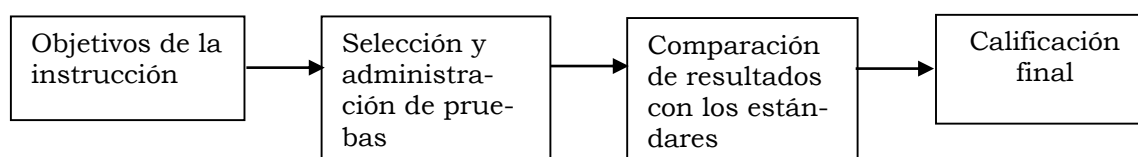


Figura 8.1. El proceso de calificación

Un efectivo y exitoso proceso de calificación requiere que los estudiantes comprendan los objetivos del curso, conozcan las pruebas y mediciones usadas para calificar, y conozca el método mediante el cual los puntajes de las pruebas se combinarán para determinar la calificación final. Informe a los estudiantes de estos factores al comienzo de la instrucción y emplee las técnicas de evaluación formativa para informarles de su progreso personal a lo largo del curso.

Determinación de los objetivos instruccionales

Hay tres cuestiones a considerar cuando se determine si un potencial objetivo debe ser parte del proceso instruccional de la unidad y, así, ser parte del proceso de calificación:

1. ¿El objetivo es defendible como un resultado educacional importante?
2. ¿Cada estudiante tiene igual oportunidad de demostrar su habilidad acerca del objetivo?
3. ¿El objetivo puede ser medido objetiva, fiable, relevante y válidamente?

Lo que no es calificación

A menudo los alumnos de educación física son calificados en aspectos como asistencia, uniformes, toma de la ducha, liderazgo, actitud, deportividad, participación, clasificación del equipo, o mejoramiento. La figura 8.2 resume los atributos reportados que se han utilizado para calificar; cuando usted hace las preguntas anteriores puede ver que la mayoría de los atributos calificados en la figura son ampliamente inapropiados como base para la calificación de los estudiantes y por ello deben ser eliminados. Es raro que uno o más de

estos elementos sean justificados como calificación. La investigación indica que muchos de esos parámetros son ampliamente sobre utilizados como objetivos en educación física (Hensley y East, 1989). Permítanos analizar brevemente el por qué ellos son inapropiados.

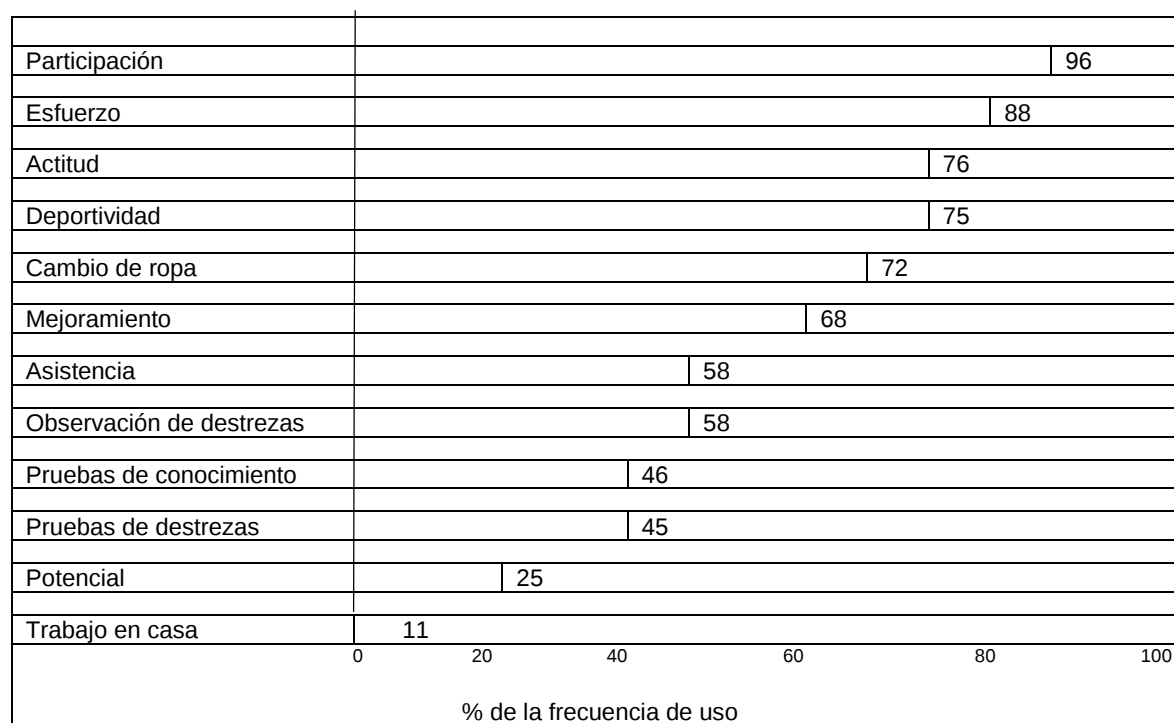


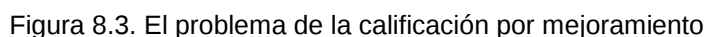
Figura 8.2 Atributos utilizados para calificar en educación física

Aunque la asistencia, el uniforme correcto y la toma de la ducha son obviamente valiosos y necesarios para la apropiada instrucción en educación física, ellos fracasan al tratar de alcanzar el primer criterio: Tener un importante resultado educacional. Los estudiantes en matemáticas pueden asistir a clase y llevar su libro de texto, pero rara vez son calificados en esos factores requeridos, ni deberían serlo.

Liderazgo, actitud, deportividad y participación son objetivos valiosos en cualquier programa atlético o de educación física. Sin embargo, para calificar esos factores fiable y válidamente, un instructor necesitaría implementar un programa sistemático y formal de medición y evaluación (tercera pregunta). La implementación de tal programa requiere de tiempo y experticia. Generalmente, el instructor de educación física no tiene el tiempo ni la experticia para ejecutar esta tarea. Cuando esos factores son utilizados para calificar (a no ser que algunas de las técnicas que se explican en el capítulo 9 sean usadas), ellos generalmente están basados en observaciones al azar del docente, las cuales tienden a ser subjetivas y posiblemente sesgadas. El resultado es una calificación con falta de objetividad, fiabilidad y validez, y por tanto son injustas para los estudiantes.

La calificación por la clasificación del equipo en la clase de deportes colectivos fracasa al tratar de alcanzar el segundo criterio, la provisión de igualdad de oportunidades para demostrar las habilidades. Un pobre ejecutante podría ser ubicado en un buen equipo y recibir una A, mientras un buen ejecutante podría ser ubicado en un pobre equipo y recibir una baja

El mejoramiento es uno de los objetivos más atractivos para usar en la calificación de la educación física. Como cualquier instructor, usted quiere que los estudiantes mejoren, pero la calificación del mejoramiento presenta algunos problemas difíciles. Uno de los problemas es ilustrado en la figura 8.3.



¿Por qué se supone que, a los educadores, especialmente los docentes de educación física, les gusta calificar el mejoramiento? Porque es probable que esta práctica les dé más

oportunidad a los estudiantes de obtener altas calificaciones, y aquellos estudiantes con un alto puntaje y los que muestran un gran mejoramiento reciben una A. Un gran número de docentes se oponen a dar bajas calificaciones ya que creen que desestimulan el esfuerzo, lo que a su vez incrementa la probabilidad de conseguir bajas calificaciones, resultando en un ciclo que continúa hasta que el estudiante rechaza la materia. La preocupación de los docentes de educación física es que cuando los estudiantes se “apagan” en la materia, es menos probable que ellos realicen actividad física a lo largo de la vida.

Sin embargo, hay poca, si hay alguna, evidencias que apoyen esta creencia. Es dudoso que los estudiantes se sientan premiados cuando se les da una alta calificación por su mejoramiento cuando su nivel real de desempeño es más bajo que muchos de sus compañeros. Por ejemplo, los nadadores de bajo nivel saben que ellos tienen ese nivel (sin que se les diga) así como sus compañeros; un estudiante honesto sabe que en una carrera larga es el nivel real de logro lo que es importante, no el nivel de su mejoramiento. Supóngase que sea cierto que las bajas calificaciones le causen un disgusto al estudiante con la escuela o, más específicamente, que la baja calificación en educación física destruye el incentivo o reduce la probabilidad de que un estudiante se comprometa con las actividades físicas. La solución de no otorgar bajas calificaciones es parecida al médico que trata los síntomas más que a la causa de la enfermedad. En su lugar, los docentes necesitan determinar por qué un individuo está desempeñándose a un bajo nivel y darle oportunidades para el éxito. Esto podría aplicarse a otras situaciones de la actividad física en clínicas y comunidades, para ajustarlas también; no se desea que se desanime a los clientes a asistir a clases o programas de ejercicios.

Primero, un instructor puede querer incrementar la cantidad de oportunidades de lograr el éxito ampliando el currículo de, por ejemplo, fútbol, baloncesto y softbol, para tener mayor variedad de actividades. Dado que las destrezas requeridas para el éxito en educación física son relativamente heterogéneas comparadas con las requeridas en otras materias. La variación en el currículo de educación física probablemente resultará en un mayor porcentaje de estudiantes logrando el éxito que en otras áreas. Por ejemplo, aprendiendo cómo controlar un golpe en particular no depende tanto de cómo efectuar un avance como, por ejemplo, aprender a derivar una raíz cuadrada es dependiente de saber cómo dividir. Afortunadamente, la actual tendencia en educación física es ofrecer más y más currículo variado.

Una segunda manera de estimular el éxito recae en la habilidad para agrupar y enganchar estudiantes en los cursos disponibles para su nivel de destreza. Si la diferencia con respecto a la habilidad inicial del participante es pequeña, la calificación del mejoramiento se hace innecesaria (porque se hace casi lo mismo como calificación del logro). Las materias de otras áreas como matemáticas, inglés o ciencias, están adelantadas en cuanto a la habilidad para agrupar; para adoptar este enfoque, el docente de educación física necesita desarrollar instrumentos de medición fiables y válidos para clasificar a los estudiantes por su habilidad.

Un tercer enfoque implica la educación de los estudiantes en el hecho de que una calificación es simplemente una expresión del nivel de logro de los objetivos del curso obtenido por el estudiante; no es un premio o un castigo. Además, no todos los estudiantes pueden sobresalir en cada curso. ***Una manera de ayudar a asegurar que la calificación sigue siendo un claro indicador del desempeño del estudiante es pensar en la calificación como mediciones más que como evaluaciones.*** En la calificación, una medición es una descripción

cuantitativa del logro del estudiante, mientras una evaluación es un juicio cualitativo de su logro. En otras palabras, una calificación, como medición, representa el grado de logro de estudiante con relación a los objetivos del curso. Si una calificación se considera una evaluación, debe indicar en algún grado cuán adecuado es el nivel de logro de un estudiante en particular.

Hay varias ventajas al darle tratamiento de mediciones a las calificaciones, más que como evaluaciones. Los juicios sobre la idoneidad del rendimiento de un estudiante no solo dependen de cuánto logró sino también de la oportunidad del logro y del esfuerzo. Esto hace difícil reportar evaluaciones precisamente en un sistema estándar de calificaciones. Adicionalmente, es probable que tenga más valor para un futuro docente o para un empleador saber que un estudiante fue excepcional, mediocre o pobre en un curso de estudios en particular, que saber que el estudiante hizo lo que se esperaba o que no cumplió con las expectativas de un maestro. Siendo válidas y fiables las mediciones del logro de un estudiante en varias áreas, el futuro docente o el empleador pueden hacer sus propias evaluaciones a la luz de las actuales circunstancias, las cuales probablemente serían más válidas que tener evaluaciones hechas por otros bajo condiciones totalmente diferentes. ***El valor de la educación física debe ser transmitido al estudiante a través de los procedimientos de enseñanza y del currículo, no a través de la premiación a través de altas calificaciones sin importar el nivel de eficiencia.***

Nótese que Tom Jones decidió otorgar la mitad de la calificación total con base en los aspectos cognitivos de la unidad de baloncesto. Así, él está ampliando las oportunidades de que los estudiantes tengan éxito, y el método de calificación no refleja los objetivos menos justificados de asistencia, mejoramiento, etc.

¿Qué calificar?

Fundamente las calificaciones en mediciones y pruebas válidas y fiables que sean representativas de importantes objetivos instruccionales. En las clases de deportes colectivos o individuales, las pruebas de destrezas deportivas, las escalas de clasificación, y el desempeño en torneos round-robin, pueden ser criterios efectivos de calificación. En los deportes individuales tal como el bowling, el golf o la arquería, se puede evaluar el logro académico de los objetivos a través del desempeño de los estudiantes. En la clase de condición física, las pruebas de campo suministran datos válidos y confiables para la evaluación. Dado que el dominio cognitivo es importante en educación física, la prueba de conocimiento debe ser parte de la calificación final. Los capítulos 10 al 14 ofrecen una selección de pruebas válidas y confiables para aplicarlas al desempeño humano.

En resumen, utilice los siguientes procedimientos para efectuar una calificación satisfactoria, fiable y válida:

- Determinar cuidadosamente unos objetivos justificables en cada curso, antes de comenzar.
- Agrupar a los estudiantes de acuerdo a la habilidad en las destrezas físicas necesarias para el curso.
- Construir pruebas y mediciones tan objetivamente como sea posible, consciente de que todas las pruebas son subjetivas en algún grado.

- Recuerde que no importa cuán bien construida esté una prueba, ninguna prueba es perfectamente fiable.
- Tenga en cuenta que la distribución de las calificaciones para cualquier grupo no necesariamente se ajusta a una curva en particular, pero que a largo plazo las destrezas físicas probablemente se distribuyan de una manera bastante normal.
- Determine las calificaciones que reflejen solo el nivel de logro en función de los objetivos del curso, y no de otros factores.
- Establezca las calificaciones con base en el estatus, no del mejoramiento.
- Evite el uso de las calificaciones para premiar un buen esfuerzo de alguien con poco dominio, o como castigo por el pobre esfuerzo de alguien que si domina.
- Considere las calificaciones como mediciones, no como evaluaciones.

Consistencia de la calificación

Una meta de cualquier sistema de calificación es lograr consistencia en la determinación de la calificación. Teóricamente, la calificación de un estudiante no puede depender de cualquiera de los siguientes factores:

- Una sección particular de un curso. Si los estudiantes están cursando PE-100 Tenis para principiantes, su calificación no puede variar si está en la sección de las 9:00 AM o la sección de las 2 PM.
- Un semestre particular de la clase. El nivel de desempeño de un estudiante debe obtener la misma calificación si toma el curso en otoño o si lo toma en verano.
- Otros estudiantes en la clase. La calificación de un estudiante depende solo de su desempeño y no puede ser influenciada por el desempeño de otros estudiantes en la clase.
- Un instructor particular. Los instructores X y Y deben dar la misma calificación al mismo nivel de desempeño en PE-100.

La meta de consistencia es extremadamente dura de lograr a causa de las diferencias entre el docente y los estudiantes, lo cual es un fenómeno natural en cualquier ambiente educacional. Cuando se estudie adelante en este capítulo, usted descubrirá que la falta de consistencia en la calificación es una de las debilidades de varios esquemas de calificación.

Mecánica de la calificación

Hay cuatro pasos involucrados en el proceso de la mecánica de alcanzar una calificación que represente el nivel de logro de cada estudiante con respecto a los objetivos del curso:

1. Determine y dele peso a los objetivos del curso.
2. Mida la cantidad de logros dirigidos a los objetivos del curso.
3. Combine las mediciones para obtener un puntaje compuesto para cada estudiante.

4. Convierta los puntajes compuestos en calificaciones.

Cada paso puede ser cumplido de varias maneras, debido a diferentes situaciones y filosofías.

Paso 1. Determinar y darles peso a los objetivos del curso

La determinación de los objetivos del curso y su significado es el más importante de los cuatro pasos, y requiere un considerable tiempo de reflexión. De hecho, ello es básico para cada aspecto de la enseñanza de un curso, no solo para la calificación. La determinación de los objetivos del curso precede a la planificación de la secuencia de presentación del material, el equipamiento necesario, los procedimientos de enseñanza, las asignaciones, y los procedimientos de calificación, y deben estar basados en tanto conocimiento de las habilidades futuras de los estudiantes como usted pueda obtener, el objetivo general de la educación física, y las consideraciones prácticas tales como la cantidad de estudiantes, las limitaciones de las instalaciones, la duración del curso o la unidad, y la cantidad, frecuencia y duración de los encuentros.

Los objetivos que finalmente se decidan para un curso de educación física pueden ser clasificados en los dominios psicomotor, cognitivo y afectivo (o psicológico) (ver también el capítulo 1, pp. 9 a 10). Expresando los objetivos en términos conductuales ayudará a dar el segundo paso (medición del logro) fácilmente. En otras palabras, enumerar los niveles de desempeño real, el conocimiento específico y la conducta social que se esperan que sean logrados por los estudiantes. Una muestra de los objetivos de la unidad de bádminton son los siguientes:

Muestra de objetivos de bádminton

- Objetivo cognitivo: Conocer las reglas establecidas cuando el juego se empatiza en ciertas puntuaciones.
- Objetivo psicomotor: Ser capaz de colocar al menos cuatro de cinco servicios cortos bajos en la parte apropiada de la cancha.
- Objetivo afectivo: Tener en cuenta la etiqueta adecuada al entregar el volante al oponente al final de una jugada.

Cómo darles peso a los objetivos del curso variará enormemente de una situación a otra, dependiendo de su filosofía, la edad y habilidad de los estudiantes, y así sucesivamente. Por ejemplo, usted probablemente les dé menos énfasis a los objetivos del dominio afectivo en una unidad de condición física para niños y niñas (sin distinción) de 10^{mo} grado que en una unidad de voleibol co-educacional del mismo grado. ***El peso real que se le dé a cada objetivo debe resultar en una lista de metas balanceada y justificada a ser lograda por cada miembro de la clase.***

Informando a los estudiantes al inicio de lo que se espera de ellos facilita la planificación de las experiencias académicas, los métodos de enseñanza, los procedimientos de calificación, y reduce la ansiedad del estudiante.

Paso 2. Medir el grado en que se logran los objetivos del curso

Recuerde que medición es el procedimiento de asignación de un número a cada integrante del grupo con base en alguna característica. En este caso, la característica involucrada es el grado de logro de cada objetivo del curso. A diferencia del paso 1, los pasos 2 al 4 ocurren después de que la enseñanza y, presumiblemente, el aprendizaje, han tenido lugar. Esto no es para implicar que todas las mediciones deben ser de naturaleza sumativa. Al contrario, hay mucho mérito en la obtención de mediciones en el transcurso de una unidad (evaluación formativa) porque hacerlo conduce al incremento de la conciencia tanto para el estudiante como para el docente, del progreso en procura de los objetivos del curso. Sin embargo, las pruebas o mediciones utilizadas para calificar el grado de logro de los estudiantes no deberían ocurrir, obviamente, antes de la completación de la instrucción y el aprendizaje. Los capítulos 10 al 13 ofrecen ideas para la construcción, evaluación, selección y administración de pruebas y otros instrumentos, para llegar a un valor numérico que permita asignarle a cada estudiante lo que con mayor precisión represente el nivel de logro de los objetivos del curso que están siendo medidos.

Paso 3. Obtener un puntaje compuesto

Rara vez un curso tiene un objetivo único. Por lo general se hace más de una medición para determinar la cantidad de logros por cada objetivo. ***Por estas razones, usualmente es necesario combinar varios puntajes para llegar a un valor único que represente el nivel total de logro de los objetivos del curso por parte del estudiante.*** Normalmente, este **puntaje compuesto** es luego convertido en cualquier formato de calificación que esté siendo utilizado (por ejemplo, A-B-C-D-F; aprueba/falla). El método correcto para obtener un puntaje compuesto depende de varios factores, de los cuales el más importante es la precisión de los puntajes que están siendo combinados.

En el caso de los puntajes de desempeño, ello es evidente porque un puntaje compuesto no se puede obtener con la simple totalización de varios puntajes originales. Usualmente las unidades difieren: los pies no pueden agregarse a los segundos, o la cantidad de repeticiones de un ejercicio no pueden sumarse a la distancia registrada en pulgadas. Como se mencionó brevemente en el capítulo 3 al discutir los puntajes estándar, los puntajes de distribuciones con diferentes cantidades de variabilidad, no pueden simplemente ser sumadas porque la variabilidad afecta el peso con el que cada puntaje contribuye al puntaje compuesto. El siguiente ejemplo ilustra el por qué sumar los puntajes originales de dos pruebas que tienen la misma unidad (como las pruebas escritas) no pueden resultar en un puntaje compuesto aspirado: Imagine que los puntajes de una clase de 25 estudiantes quedaron distribuidos como se muestra en la tabla 8-1, en tres pruebas con un valor de 9 puntos, 9 puntos y 27 puntos. Este ejemplo extremo fue seleccionado para considerar un punto en particular: que un puntaje dentro de un conjunto de puntajes que tienen una gran variabilidad, tendrá más peso en el puntaje compuesto que un puntaje dentro de un conjunto de puntajes con poca variabilidad, con respecto al valor absoluto de los puntajes. La variabilidad de los puntajes de las dos primeras pruebas es mayor que la variabilidad de la prueba 3. Podría parecer que, debido a que el número total de puntos posibles de la prueba 3 es el triple que el de las pruebas 1 y 2, el puntaje logrado en la prueba 3 tendría mayor influencia en el puntaje compuesto de los estudiantes.

Prueba 1 (9 puntos)		Prueba 2 (9 puntos)		Prueba 3 (27 puntos)	
Puntaje	Frecuencia	Puntaje	Frecuencia	Puntaje	Frecuencia
9	1	9	1	27	
8	2	8	2	26	
7	3	7	3	25	
6	4	6	4	24	7
5	5	5	5	23	11
4	4	4	4	22	7
3	3	3	3	21	
2	2	2	2	20	
1	1	1	1	19	
Estudiante	Puntaje en la prueba 1	Puntaje en la prueba 2	Puntaje en la prueba 3	Puntaje total	
A	5	5	24	34	
B	8	7	22	37	
C	9	5	23	37	
D	3	3	24	30	

Tabla 8-1 Distribuciones de los puntajes de las pruebas de una clase de 25 estudiantes

Nótese, sin embargo, algunas posibilidades:

El estudiante A fue uno de los 7 estudiantes que lograron el puntaje más alto alcanzado en la prueba 3; sus puntajes en las dos primeras pruebas fueron el promedio. El estudiante A recibió un puntaje compuesto original de 34. El puntaje compuesto original del estudiante B, que estuvo por arriba del promedio en las dos primeras pruebas, pero quedó entre los puntajes más bajos en la prueba 3, es de 37, más alto que el estudiante A. El estudiante C quedó en el promedio en dos de las pruebas y por arriba del promedio en una prueba, como lo hizo el estudiante A. Sin embargo, ya que el estudiante C alcanzó un desempeño por arriba del promedio en la prueba 1 (en la cual los puntajes fueron más variables que en la prueba en la que el estudiante A logró estar por arriba del promedio, la prueba 3), su puntaje compuesto original, 37, es también más alto que el del estudiante A. Finalmente, aun cuando el puntaje del estudiante D es más alto que cualquiera de la clase en la prueba 3, los puntajes bajos alcanzados en las primeras dos pruebas disminuyeron el puntaje compuesto original a 30, más bajo que los demás.

Al menos dos (o más) conjuntos de puntajes son similares en variabilidad, por lo que sumar los puntajes originales de los estudiantes para determinar un puntaje compuesto puede llevar a algunas conclusiones incorrectas.

Asuma que, en lugar de representar la distribución de los puntajes de las tres pruebas, los valores de la tabla 8-1 describan la distribución de las medidas de cuán bien los objetivos psicomotor, cognitivo y afectivo fueron alcanzados por los estudiantes. Además, asuma que ha decidido que el logro de los objetivos cognitivos, afectivos y psicomotor representan el 20%, 20% y 60% de la calificación final, respectivamente. Para lograr esta distribución, simplemente usted hizo que la cantidad de puntos posibles para cada uno de los tres objetivos fuese

*el porcentaje deseado de la cantidad total de puntos posibles (es decir, 9,9 y 27). Como se concluyó anteriormente, a menos que la variabilidad de los tres conjuntos de puntajes sea muy similar, el peso real será diferente a la originalmente planteada. **La solución es procesar la ponderación después de que se obtenga el puntaje de cada objetivo en lugar de intentar aplicar el factor de ponderación al sistema de puntuación, a menos que usted pueda suponer que se producirá una variabilidad igual o casi igual entre los conjuntos de puntajes.***

Si el cálculo de los puntajes compuestos combinando puntajes originales no es factible por la diferencia entre las unidades de medición o por la falta de igualdad en la variabilidad entre los conjuntos de puntajes, ¿cómo se pueden formar los puntajes compuestos? Como se describió en el capítulo 3 (p.42), convierta cada conjunto de puntajes a la misma distribución estándar estableciendo una base común para hacer posible la comparación, contraste, ponderación y suma de varios conjuntos de puntajes. Hay varios métodos para hacerlo (tres de los cuales se discuten pronto); la selección del mejor método depende de las mediciones involucradas y del supuesto de normalidad.

En el caso de la precisión de los puntajes, determine si los puntajes están en una escala ordinal o en una escala de intervalo o de razón. Como se describió en el capítulo 3, cuando se utiliza una escala ordinal (tal como las clasificaciones en un torneo round-robin), solo es posible decir que A es superior a B. Sin embargo, en las escalas de intervalo o de razón (como la cantidad de push-ups ejecutados), es posible afirmar cuán superior es A de B, porque estas escalas tienen unidades de igual tamaño.

La segunda consideración implica determinar si la distribución de los puntajes se aproxima a una distribución normal. Si no lo está, se debe a que el rasgo que está siendo medido, en sí, no está normalmente distribuido, o es porque, aun cuando el rasgo esté normalmente distribuido, la muestra que se tiene a mano, por alguna razón, no lo refleja.

Hay cinco posibles situaciones que involucran esas dos consideraciones (tabla 8-2). La razón por la que no hay seis posibles situaciones es que, si los puntajes están en una medida ordinal, la distribución de los puntajes – una simple clasificación de los estudiantes – no puede aproximarse a una distribución normal. Cada situación está asociada a uno de los tres métodos de conversión de puntajes a una distribución estándar: clasificación, normalización y puntaje estándar. Aunque hay pruebas estadísticas disponibles para determinar si una distribución de puntajes de intervalo o de razón es significativamente diferente a una distribución normal, tales pruebas están más allá del alcance de este libro. Sin embargo, una inspección visual de la distribución de las frecuencias de la muestra es por lo general suficiente para revelar su cercanía a una distribución normal. Si no hay seguridad de la proximidad de la distribución de los puntajes cercana a la normal, se puede utilizar el método de normalización (p. 139).

Forma de distribución de la Muestra	Escala de medición	
	Ordinal	Intervalo o razón
No es normal	Método de clasificación	Método de clasificación
No es normal pero los rasgos están normalmente distribuidos	Método de normalización	Método de normalización
Aproximadamente normal	(No es posible)	Método de puntajes estándar
Tabla 8-2 Métodos de obtención de un puntaje compuesto con base en la forma de la distribución y de la escala de medición		

Esté consciente de que el resultado final del método de clasificación es, simplemente, ubicar en una posición a los estudiantes, mientras el resultado final de los otros dos métodos es el establecimiento de los puntajes estándar. ***No es posible llegar a un puntaje compuesto si algún conjunto de puntajes es convertido a posiciones y otros son convertidos a puntajes estándar. Por tanto, si alguno de los puntajes que están siendo sumados para obtener un puntaje compuesto debe ser expresado en posiciones, entonces todos los puntajes deben ser expresados en posiciones.*** Por esta razón, se debe planificar por adelantado el tipo de medición que se utilizará durante la unidad de instrucción.

- Método de clasificación

El método más simple, que requiere una menor precisión de la medición, es la clasificación numérica del desempeño de los estudiantes en cada tarea realizada. Sin embargo, esta falta de precisión también hace este método menos fiable que los otros y, por ello, este método debe ser evitado en lo posible. En el método de clasificación, al mejor desempeño se le da la posición 1, al segundo mejor desempeño la posición 2, y así hasta el más bajo desempeño al que se le da una posición igual al número de estudiantes que está siendo medido. ***Para obtener el puntaje compuesto de cada estudiante en el sistema de clasificación numérica, simplemente se suman las posiciones que cada quien haya obtenido. El total más bajo representa el mejor logro total.*** En el caso de que una o más posiciones no estén sumadas por alguna razón, se puede utilizar la media de las posiciones la cual se obtiene dividiendo la suma de las clasificaciones por el número de valores que contribuyan a la suma. Si usted desea ponderar las clasificaciones, sume las que considere más importantes más de una vez para llegar al total. Por ejemplo, asuma que se lograron tres posiciones: la primera cuenta el 10%, la segunda el 40%, y la tercera el 50% de la calificación final. Para cada estudiante, un puntaje compuesto se obtendría sumando la primera posición, cuatro veces la segunda clasificación, y cinco veces la tercera clasificación. Como se dijo antes, el total más bajo representa el mejor logro total. Esto puede parecer contradictorio con lo que se estudió en el capítulo 3, página 32; sin embargo, la robustez de las clasificaciones en esta situación permite la ejecución de operaciones matemáticas de ellos aun cuando representan datos ordinales.

Una variación del sistema de clasificación numérico es calificar cada estudiante como perteneciente a una cantidad determinada de categorías. Por ejemplo, los cinco mejores logros pueden ser valorados como 1, los próximos cinco como 2, y así sucesivamente. Otra

variación va un paso más allá al darle a las categorías calificaciones en letras, en lugar de números. Este último procedimiento no requiere una cantidad particular de estudiantes por clasificar en cada categoría. Tiene algún mérito en que le da más información a los estudiantes que la mera clasificación numérica, pero los procedimientos utilizados para llegar al puntaje compuesto son los mismos. De hecho, en este sistema un puntaje compuesto se obtiene cambiando las calificaciones de letras a números y procediendo de una manera similar a la que se describe en el sistema de clasificación numérico. Por ejemplo, a las categorías A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, y F, se les dan los valores de 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, y 0, respectivamente. (Note que, en este método, un número alto es mejor que uno bajo). Un estudiante reciba una A+ en los objetivos afectivos (10%), una B- en los objetivos cognitivos (40%), y una C- en los objetivos psicomotor (50%). Para llegar a un puntaje compuesto, las clasificaciones literales son convertidas en sus equivalentes numéricos, multiplicados por el peso del objetivo correspondiente, y sumados:

$$(12 \times 1) + (7 \times 4) + (4 \times 5) = 12 + 28 + 20 = 60$$

Dividiendo esta suma por 10 (la suma de los pesos = 10) y comparando el valor resultante con las categorías, convierta el desempeño de este estudiante a una C+ ($60/10 = 6 = C+$).

- Método de normalización

Utilice el método de normalización cuando los puntajes obtenidos en la medición de un rasgo que se conoce o se cree que está normalmente distribuido, no parecen resultar en una curva normal aproximada. Por ejemplo, usted podría esperar que la distribución de la cantidad de tiros libres en baloncesto realizados por los niños de 10^{mo} grado en un período de tiempo dado se aproxime a la normalidad a medida que aumenta la cantidad de puntos como datos. Si la medición toma la forma de clasificación, la curva normal no será aproximada. Otras razones, tal como la falta de una muestra representativa, pueden generar una distribución diferente a la normal aun cuando el rasgo que está siendo medido esté normalmente distribuido. El método de normalización es muy parecido a la conversión de un conjunto de puntajes originales en una distribución de algún tipo de puntaje estándar como el puntaje T. La diferencia, sin embargo, es que en el presente caso los puntajes originales son primero convertidos a percentiles, los cuales a su vez son convertidos en una escala de puntajes estándar (en muchos casos se utilizan los puntajes T). Una descripción del procedimiento de conversión de un conjunto de puntajes originales en percentiles aparece en el capítulo 3 (p.34), y los percentiles resultantes pueden ser convertidos en puntajes T utilizando la tabla 8-3.

Percentil	Puntaje T	Percentil	Puntaje T	Percentil	Puntaje T
0,02	15	13,57	39	90,32	63
0,03	16	15,87	40	91,92	64
0,05	17	18,41	41	93,32	65
0,07	18	21,19	42	94,52	66
0,10	19	24,20	43	95,54	67
0,13	20	27,43	44	96,41	68
0,19	21	30,85	45	97,13	69
0,26	22	36,46	46	97,72	70
0,35	23	38,21	47	98,21	71
0,47	24	42,07	48	98,61	72
0,60	25	46,02	49	98,93	73
0,82	26	50,00	50	99,18	74
1,07	27	53,98	51	99,38	75
1,39	28	57,93	52	99,53	76
1,79	29	61,79	53	99,65	77
2,28	30	65,54	54	99,74	78
2,87	31	69,15	55	99,81	79
3,59	32	72,57	56	99,87	80
4,46	33	75,80	57	99,90	81
5,48	34	78,81	58	99,93	82
6,68	35	81,59	59	99,95	83
8,08	36	84,13	60	99,97	84
9,68	37	86,43	61	99,98	85
11,51	38	88,49	62		

Nota. Aunque la escala de puntajes T se extiende de 0 a 100, en la práctica los puntajes T más bajos de 15 o más altos de 85 son raros, y por ello no se incluyen en la tabla.

Tabla 8-3 Conversión de Percentiles a Puntajes T

Puntaje original	f	fa	fam	Percentil	Puntaje T de la tabla 8-3
85	1	6	5,5	91,7	64
74	1	5	4,5	75	57
63	1	4	3,5	58,4	52
59	1	3	2,5	41,7	48
53	1	2	1,5	25,0	43
47	1	1	0,5	8,4	36

Nota: En esta tabla la columna **f** es la frecuencia de los puntajes. La columna **fa** es la frecuencia acumulada asociada con cada puntaje (comenzando con el puntaje más bajo). Para obtener los valores de la columna **fam** (punto medio de la frecuencia acumulada) sume medio punto del valor **f** para cada intervalo al valor de **fa** desde el próximo puntaje más bajo. Por ejemplo, el valor de 3,5 en la columna **fa** al intervalo 63 se obtienen agregando 0,5 (1/2 de la frecuencia de 1 para este intervalo) a 3,0 (la **fa** del intervalo de abajo del intervalo de 63). Finalmente, se expresan los valores de **fam** como percentiles, se divide 100 por n (en este caso, 6) y se multiplica cada valor **fam** por el cociente resultante.

Tabla 8-4 Ejemplo del Método de Normalización

Para ilustrar cómo este procedimiento convierte una distribución no normal en una distribución normal, nótese que un puntaje, independientemente de la distancia de desviación estándar a la media del puntaje original, que cae al 34,13% por arriba de la media (1 desviación estándar por arriba de la media en una curva normal) es equivalente a un puntaje T de 60, el cual está a 1 desviación estándar por arriba de la media en una escala de puntajes T. Un ejemplo del método de normalización que deriva en el puntaje T correspondiente para cada puntaje original se muestra en la tabla 8-4. Una vez que cada conjunto de puntajes se convierte en puntajes T, a estos se les puede ponderar como se desee y luego combinarlo para llegar a un puntaje compuesto para cada estudiante. Estos puntajes compuestos pueden ser convertidos luego en una calificación apropiada de acuerdo al formato en particular que se esté utilizando.

- Método de puntaje estándar

Si usted considera que es seguro asumir que la distribución de los puntajes originales se aproxima bastante a una distribución normal, convierta los puntajes originales a una escala de puntajes estándar como la escala de puntajes T que se describe en el capítulo 3. El resultado neto de este procedimiento es que, como en el método de normalización, a cada puntaje original le corresponde un puntaje T. Si usted convierte todo el conjunto de puntajes a puntajes T (o en otra escala estándar), existe una base común para la comparación de los puntajes alcanzados en dos pruebas diferentes aun cuando los dos puntajes originales fuesen expresados en unidades diferentes o tuviesen diferente variabilidad. Además, ahora se pueden agregar los puntajes T que representen otros logros para obtener los puntajes compuestos significativos y que pueden ser utilizados en la determinación de la calificación final. Así como en el método de clasificación, también se puede ponderar varias pruebas (es decir, en los respectivos objetivos en los cuales ellas se basan) multiplicando el puntaje T por la ponderación apropiada que usted determinó previamente.

Una vez obtenido el puntaje compuesto para cada estudiante, el paso final es cambiar este puntaje a la calificación conforme al formato específico que esté siendo utilizado. Como en cada uno de los otros tres pasos, varios factores afectarán cómo usted hace esto. La decisión de cuál procedimiento seguir se basa en el formato del puntaje compuesto, las políticas del departamento o sistema escolar, y sus propias preferencias.

Paso 4. Convertir los puntajes compuestos a calificación

Al final del paso 3, los puntajes compuestos serán de una de las dos formas: cada estudiante tendrá una clasificación total o promedio, o un puntaje estándar total o promedio. En efecto, ambas formas conforman una ordenación de los estudiantes, aunque en el método de clasificación el total más bajo (o promedio más bajo) por lo general representa el mejor logro, mientras en los métodos de normalización y de puntaje estándar el total más alto (o promedio más alto) representa el mejor logro. Los procedimientos de conversión de un conjunto de puntajes compuestos a calificaciones son los mismos independientemente de la forma de los puntajes compuestos e implica respuestas a dos cuestiones relacionadas:

1. ¿Está la clase por ser calificada por debajo o por arriba del promedio en cuanto al logro de los objetivos del curso en comparación a clases similares?

2. ¿Qué porcentaje de estudiantes debe recibir cada calificación?

Si las pruebas y los instrumentos de medición son absolutamente fiables y válidos, y si los objetivos del curso permanecen constantes en el tiempo, los docentes no necesitarían responder a estas cuestiones antes de convertir los puntajes compuestos a calificaciones. Sin embargo, las mediciones no son perfectas, los objetivos cambian, ocurren eventos ni planificados ni esperados, las instalaciones y el equipamiento cambian con el tiempo, y otros factores adicionales hacen imposible comparar el logro de la clase actual con clases previas con base en los objetivos, estrictamente. ***La calificación es especialmente difícil para los nuevos docentes porque les falta la experiencia en la cual basar su respuesta a estas cuestiones.*** Después de tener algunas respuestas subjetivas, usted puede utilizar varios métodos para convertir los puntajes compuestos a calificaciones.

- Observación

La observación es uno de los métodos más simples para la determinación de las calificaciones. Haga una lista de los puntajes del mejor al inferior. Examine la lista para determinar las brechas naturales o discontinuidad de los puntajes. La tabla 8-5 presenta una lista de los puntajes de 15 jóvenes del bachillerato. Como se ve, dos brechas aparecen en la data. Las mismas se utilizan como puntos de corte para las calificaciones literales A, B, y C.

Este sistema usualmente trabaja bien con un pequeño número de puntajes, porque la brecha en la data aparece frecuentemente en ellos. Sin embargo, la observación no es útil con un gran número de puntajes, porque es probable que las brechas observables no se presenten. Además, este método no asegura consistencia. Las brechas naturales podrían ser muy diferentes en dos clases: una calificación de A en la clase del semestre de otoño podría bajar a una calificación de B en el grupo de verano.

Puntajes	Frecuencia	Calificación
150	1	A
140	2	
110	3	
100	2	B
90	2	
80	1	
70	1	
40	1	C
30	1	
20	1	

Tabla 8-5 Calificación de los Puntajes con el Método de Observación

- Porcentajes predeterminados

El método de porcentajes predeterminados puede ser utilizado con los puntajes compuestos en la forma de las clasificaciones o de los puntajes estándar, porque no es el valor del puntaje en esa materia sino su posición en la distribución. Una vez que se decida el porcentaje de

estudiantes a los cuales les serán asignadas las calificaciones, se necesita solo multiplicar el porcentaje de los valores por el número de estudiantes en la clase, como sigue:

$$k = n * \left(\frac{P}{100} \right)$$

donde k es la cantidad de estudiantes que reciben una calificación específica, n es la cantidad total de estudiantes, y P es el porcentaje de una calificación específica. El producto resultante es la cantidad que se asigna a cada calificación literal. La tabla 8-6 muestra este procedimiento para una clase de 45 estudiantes los cuales, sobre la base de los puntajes de la prueba y otras evidencias, han logrado sustancialmente llegar por encima del promedio de otras clases similares. El docente decide distribuir el 15% en A, 25% en B, 45% en C, 10% en D, y 5% en F. Si los puntajes compuestos están en la forma de puntajes estándar o en la suma (o promedio) de varias calificaciones (pero no si los puntajes compuestos están como una clasificación simple en la cual el mejor estudiante tiene el puesto 1, el segundo mejor estudiante tiene el puesto 2, etc.), usted puede modificar ligeramente el método de porcentaje predeterminado utilizando las divisiones naturales en la distribución de los puntajes compuestos. Como lo indica la última columna de la tabla, la cantidad real de puntajes para una calificación puede ser ligeramente diferente a la k calculada, pero los puntos de corte son seleccionados para tener como resultado la cantidad de estudiantes cercanos a la cantidad predeterminada.

Calificación a otorgar	Porcentaje preestablecido por el docente, de los estudiantes que recibirán la calificación	Cantidad de estudiantes que reciben la calificación (k)	Puntajes	Frecuencia	Cantidad total real de estudiantes que recibirán la calificación
A	15%	$45 \times 0,15 = 6,75 = 7$	120	2	6
			110	2	
			100	2	
B	25%	$45 \times 0,25 = 11,25 = 11$	90	3	12
			80	4	
			70	5	
C	45%	$45 \times 0,45 = 20,25 = 20$	60	8	21
			50	5	
			40	5	
D	10%	$45 \times 0,10 = 4,5 = 5$	30	3	4
			20	4	
F	5%	$45 \times 0,05 = 2,25 = 2$	10	2	2
Tabla 8-6 Calificaciones determinadas por los porcentajes preestablecidos					

Así como en el método de observación, el método calificación por porcentajes predeterminados no asegura consistencia de la calificación asignada de una clase a otra o de un semestre a semestre. Aun si los porcentajes predeterminados son consistentes, por cuanto los puntajes varían de una clase a otra y de un semestre a otro, el punto de corte real (en términos de puntaje) variará. Así, un puntaje que “se traduce” en una A en otoño podría ser una B en verano. Sin embargo, usted puede ajustar esto cambiando el porcentaje que se le da a cada calificación.

- Calificación en la curva

Aunque la *calificación en la curva* es una frase oída a menudo en el ambiente educativo, el proceso el proceso no ha sido bien entendido. Realmente, la calificación en la curva es una variación del método de porcentajes predeterminados, en el cual se parte del supuesto de que las diferencias en las habilidades de los estudiantes están distribuidas normalmente, o al menos casi normalmente, y por ello los porcentajes de cada calificación pueden ser determinados a través de la curva normal.

Cierta confusión acerca de la calificación en la curva se deriva del hecho de que es posible utilizar la curva normal de dos maneras. Los límites prácticos de la curva normal (± 3 desviaciones estándar) se pueden dividir por igual en el número de símbolos diferentes que se asignarán, o ciertas distancias de desviación estándar de la media simplemente se pueden seleccionar como los límites para cada símbolo asignado. Consulte la tabla 8-7 para ver los datos utilizados para ilustrar cada enfoque. En este ejemplo, los puntajes compuestos de 65 estudiantes han sido compilados, y usted se propone utilizar la curva normal para determinar los puntos de corte para la asignación de las calificaciones de A, B, C, D, y F. La media y la desviación estándar de los puntajes compuestos son 63,4 y 14,9, respectivamente.

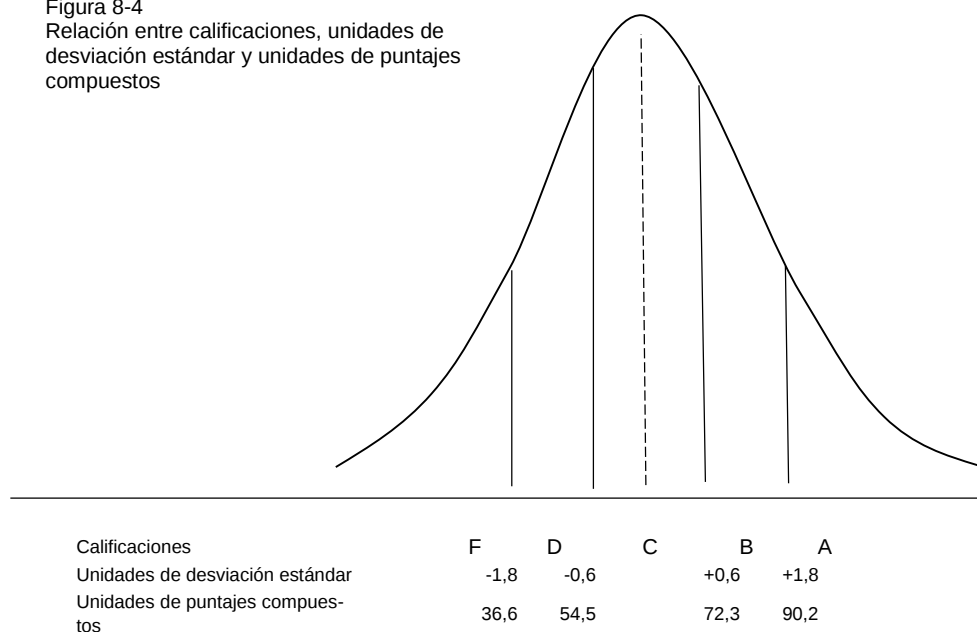
98	78	71	64	60	52	40
93	78	70	64	59	51	38
91	77	69	63	57	50	37
88	76	68	63	57	48	36
86	75	67	63	56	47	26
85	74	67	63	56	47	
83	73	66	62	55	46	
81	73	65	62	55	45	
81	72	65	61	54	44	
79	71	65	61	53	41	
Tabla 8-7 Puntajes compuestos de 65 estudiantes						

Primer enfoque

Para propósitos prácticos, se puede considerar que la curva normal se extiende ± 3 unidades de desviación estándar por arriba y por debajo de la media. (Recuerde que el 99,73% del área

bajo la curva normal se encuentra entre esos dos puntos). Esta amplitud total de 6 unidades de desviación estándar se divide igualmente en la misma cantidad de partes como calificaciones diferentes hayan de ser asignadas —en este caso, cinco (A, B, C, D, F). Cada calificación abarca un ancho de 1,2 unidades de desviación estándar ($6/5 = 1,2$) (figura 8.4). Ya que el formato de calificación utilizado en esta ilustración tiene un número impar de calificaciones, la mitad de la calificación media (C) cae a un lado y al otro de la media. Una calificación de C se asignará a aquellos estudiantes cuyos puntajes compuestos caen 0,6 unidades de desviación estándar por arriba de la media y 0,6 unidades de desviación estándar por debajo de la media. La calificación B será asignada a los estudiantes cuyos puntajes compuestos están entre 0,6 y 1,8 unidades de desviación estándar por arriba de la media; la calificación D será asignada a aquellos cuyos puntajes compuestos caen entre -0,6 y -1,8 unidades de desviación estándar por debajo de la media. Si este proceso continúa, resultará que las unidades de desviación estándar entre 1,8 y 3,0 por arriba de la media se califican como A, y las unidades de desviación estándar entre -1,8 y -3,0 se califican como F. Para tomar en cuenta los puntajes muy extremos, se puede especificar que los puntajes compuestos mayores a 3,0 unidades de desviación estándar por arriba de la media les corresponde una A, y que los puntajes compuestos más allá de -3,0 les corresponde una F.

Figura 8-4
Relación entre calificaciones, unidades de desviación estándar y unidades de puntajes compuestos



El procedimiento final implica la expresión de las unidades de desviación estándar en términos de los valores de los puntajes compuestos. Ya que en este ejemplo 1 desviación estándar es equivalente a 14,9 unidades de desviación estándar, el punto de corte entre C y B y entre C y D se sitúa en 8,94 unidades de puntaje compuesto por arriba y por debajo de la media, respectivamente. (El valor de 8,94 surge de la multiplicación de 14,9 por 0,6). Los dos puntos de corte son 72,3 y 8,94 ($63,4 \pm 8,94$). Los puntos de corte entre B y A, y entre y F se obtienen sumando y sustrayendo desde la media el producto de 14,9 y 1,8. El resultado

es 26,82 y los puntos de corte son 90,2 y 36,3 (ver figura 8-4). Para simplificar, usted probablemente redondearía los puntos de corte, y podría establecer también una carta de conversión para la asignación de calificaciones a los puntajes compuestos (tabla 8-8).

Si un formato de calificaciones que comprende cinco calificaciones es utilizado en conjunción con una curva normal, el docente está, en efecto, decidiendo asignar aproximadamente al 3,5% de la clase la calificación de A, 24% a B, 45% a C, 24% a D, y 3,5% a F. (Para verificar esto, revise cómo utilizar la tabla 3-3, p. 43.) Aplicando la tabla de conversión de la tabla 8-8 a los 65 puntajes compuestos, resulta en la asignación de 3 calificaciones A (4,6% de la clase), 15 calificaciones B (23,1%), 30 calificaciones C (46,1%), 14 calificaciones D (21,6%), y 3 calificaciones F (4,6%). Se puede producir una diferencia entre los porcentajes de la curva normal y los porcentajes reales al establecer la carta de conversión y al hecho de que el conjunto de los 65 puntajes no está exactamente distribuido normalmente.

Puntaje compuesto	Calificación
91 y más alto	A
73 – 90	B
55 – 72	C
38 – 54	D
37 y bajando	F

Tabla 8–8 Conversión de Puntajes Compuestos a Calificaciones: Ejemplo del Primer Enfoque

Segundo enfoque

El segundo enfoque de la calificación en la curva implica la selección apropiada de las distancias de la desviación estándar a partir de la media como los límites de cada calificación. En este enfoque, en lugar de dividir las seis desviaciones estándar (el límite práctico de la curva normal) en unidades iguales para cada calificación, se seleccionan las distancias que conforman su noción del porcentaje que le asigna a cada símbolo. Una posibilidad se representa en la tabla 8-9.

Calificación	Distancia de la desviación estándar para la calificación (predeterminada por el instructor)	Porcentaje de estudiantes que reciben la calificación (usar la tabla 3-3)	Basada en la tabla 8-7	
			Número de estudiantes que reciben la calificación (n = 65)	Puntaje compuesto que corresponde a la calificación
A	Arriba de + 1,48	7%	$0,07 \times 65 = 5$	≥ 86
B	+0,47 a + 1,47	25%	$0,25 \times 65 = 16$	71 – 85
C	-0,47 a +0,46	36%	$0,36 \times 65 = 23$	57 – 70
D	-1,47 a -0,46	25%	$0,25 \times 65 = 16$	41 – 56
F	Debajo de -1,48	7%	$0,07 \times 65 = 5$	≤ 40

Tabla 8-9 Calificación en la curva: Ejemplo del Segundo Enfoque

La selección de esas distancias específicas de desviación estándar resultaría en la asignación, aproximadamente, del 7% para A, 25% para B, 36% para C, 25% para D, y 7% para F, para un conjunto de puntajes normalmente distribuido. Como en el primer enfoque, esas desviaciones estándar deben ser convertidas a unidades de puntajes compuestos multiplicando la constante seleccionada por la desviación estándar de los puntajes compuestos. Para los datos señalados en la tabla 8-7 (media= 63,4; desviación estándar= 14,9), los puntos de corte resultantes se muestran en la tabla 8-9. (Como se dijo anteriormente, las ligeras diferencias entre los porcentajes de la curva normal y los porcentajes reales suceden por el redondeo y el hecho de que los 65 puntajes no están exactamente distribuidos normalmente).

Aunque se puede elegir cualquier conjunto de distancias, si el supuesto de normalidad es generalmente cierto, los valores seleccionados necesariamente resultarán en la asignación simétrica de las calificaciones. Si se plantea una distribución no simétrica de las calificaciones, el segundo enfoque de calificación en la curva se convierte en el método de porcentajes predeterminados descrito previamente. ***El método de la curva normal no asegura consistencia en la calificación de una clase a otra clase porque la media y la desviación estándar puede cambiar de una clase a otra clase.***

- Normas

Las normas se derivan de una gran cantidad de puntajes procedentes de las pruebas y mediciones de una población específicamente definida. Los puntajes son analizados estadísticamente para producir un análisis descriptivo estadístico, lo cual permite la producción de normas percentiles o normas de puntajes estándar. Hay normas disponibles para muchas pruebas de destrezas deportivas y de condición física (ver capítulos 11 al 13). El gobierno de USA realiza encuestas a gran escala sobre varias variables como la presión sanguínea y el colesterol. Esa data normativa es utilizada para desarrollar perfiles nacionales de salud, y las variables están relacionadas con la morbilidad y la mortalidad. El Estudio Nacional I y II de la Condición Física de Niños y Jóvenes produjo normas percentiles de varias pruebas de capacidad física de los jóvenes americanos (Pate, Ross, Dotson, y Gilbert 1985; Ross et al 1987).

Las normas pueden ser utilizadas para la asignación de calificaciones, o para generar sus propias normas de los puntajes compuestos y llegar a la calificación final. Por ejemplo, la tabla 8-10 proporciona las normas percentiles 25 y 75 de los puntajes de la prueba de caminar una milla (1,6 km) para 400 hombre y 426 mujeres con una edad entre 18 y 30 años. Usted podría asignarle una A al cuartil superior (por arriba de P_{75}), una B a los dos cuartiles medios ($P_{25} - P_{75}$), y una C al último cuartil (debajo de P_{25}). ***Las normas nacionales y otras publicadas pueden ser utilizadas para establecer escalas de calificación. Sin embargo, las normas más justas para la determinación de las calificaciones son las normas locales.*** Para ser utilizadas en las calificaciones, las normas deben ser representativas de la edad, sexo, entrenamiento y situación instruccional de los estudiantes. Las normas desarrolladas en el lugar de instrucción con estudiantes similares a los que se van a calificar, asegurará su representatividad.

Percentil	Masculino	Femenino
75	11:42	12:49
25	13:38	14:12
Tabla 8-10 Normas percentil para la prueba de caminar una milla		

Usted puede desarrollar normas locales para puntajes compuestos de acuerdo con su disposición a seguir los siguientes pasos:

1. Determinar los objetivos del programa instruccional o de entrenamiento.
2. Seleccionar las pruebas y mediciones para valorar esos objetivos.
3. Administrar las mismas pruebas con procedimientos estandarizados por varios años.
4. Recolectar suficientes datos para tener al menos 200 puntajes para cada la edad y el sexo que se intenta evaluar.
5. Conducir un análisis estadístico de los datos y establecer las normas percentiles o normas de puntajes estándar.

Si usted sigue estos pasos, usted será capaz de establecer normas que sean representativas de sus estudiantes y su situación de aprendizaje. Los estándares de calificación le proporcionarán consistencia a la asignación de calificaciones entre las clases, semestres, e instructores, si todos los instructores están comprometidos en el desarrollo y uso de las normas. Sin embargo, hasta poder obtener suficientes datos para establecer las normas, usted tendrá que utilizar cualquier otra técnica de calificación. Por último, usted tendrá que trabajar hacia el establecimiento de normas locales, porque este método reduce la mayoría de las inconsistencias de las calificaciones.

- Estándares arbitrarios

Las técnicas previas de conversión de puntajes a calificaciones han implicado el análisis de los datos de las mediciones observadas. El empleo de estándares arbitrarios para la asignación de calificaciones no requiere el análisis de la data de los puntajes de las pruebas. Utilizando este proceso, los estándares referidos a criterio se establecen para cada calificación. La tabla 8-11 proporciona un ejemplo de tales estándares en una prueba de conocimiento de 100 puntos.

Rango de los puntos	Calificación
100 – 90	A
89 – 80	B
79 – 70	C
69 – 60	D
59 – 0	F
Tabla 8-11 Estándares arbitrarios para la calificación	

La mayor ventaja en este tipo de sistema de calificación es que proporciona consistencia a la calificación. Un 90 en una prueba de conocimiento será un A en cualquier clase o semestre en la que el estudiante haga la prueba. El sistema es simple y fácil de comprender.

Sin embargo, el establecimiento de estándares en las pruebas de desempeño físico sin conocimiento de los niveles de desempeño esperado de los estudiantes resulta en un proceso de “mejor estimación” que puede generar una distribución indeseable de las calificaciones. Si los estándares que utiliza son reflejo de los niveles de rendimiento de los estudiantes, el sistema arbitrario puede utilizarse con buenos resultados, pero al igual que con el método de normas, se requiere un período relativamente largo de acumulación de datos.

Una versión específica del uso de estándares arbitrarios en la evaluación es la asignación de *pasar/fracasar*. El desempeño del estudiante es comparado con un estándar referido a criterios específico que representa el nivel mínimo de desempeño o habilidad aceptable. El estándar referido a criterios pasar/fracasar es utilizado extensamente en la evaluación de la condición física relacionada con la salud de los jóvenes. Tal estándar es un puntaje de la prueba que representa el valor de desempeño mínimo que está asociado a un reducido riesgo de enfermedad o una capacidad funcional aceptable.

Tomado de: Siedentop, Daryl & Tannehill, Deborah (2000). Developing Teaching Skills in Physical Education. Mountain View, CA. Mayfield Publishing Company.

Capítulo 16. Instrumentos de valoración de la efectividad de la enseñanza

Objetivo General

Utilizar un instrumento de observación existente o crear uno nuevo, para obtener datos fiables del comportamiento del docente y del estudiante.

Objetivos del Capítulo

- Definir fiabilidad y explicar su importancia en la observación sistemática
- Describir las fortalezas y debilidades de los métodos tradicionales de valoración de la enseñanza y sus resultados
- Seleccionar y suministrar ejemplos de los métodos de observación sistemática
- Combinar métodos de observación en un sistema para alcanzar un propósito específico
- Listar los pasos necesarios para desarrollar un sistema de observación
- Describir cómo los observadores son entrenados
- Calcular la fiabilidad de los datos observados con precisión
- Diferenciar entre los propósitos de varios sistemas de observación
- Calcular la fiabilidad de los datos observados con precisión
- Diferenciar entre los propósitos de varios sistemas de observación
- Observar fiablemente la conducta del docente y los estudiantes

Para recapitular, el registro es el evento terminal de una serie compleja que comienza con la definición de la clase de respuesta que interesa, continúa a través de la observación, y culmina con la creación de un registro permanente del comportamiento... El registro permanente que queda después de la definición, observación y registro que han tenido lugar es la única evidencia de que la medición realmente ocurrió, y la calidad del proceso completo no puede exceder las características de ese registro.

- J. JOHNSTON AND PENNYPACKER (1980)

Las destrezas de enseñanza mejorarán en la medida que los docentes en formación tengan la oportunidad de practicar destrezas específicas y recibir un feedback acerca de su progreso hacia los objetivos. Esperar tal mejoramiento sin feedback es una creencia basada en deseos más que en hechos. La literatura acerca de la investigación de la formación del docente no es muy útil cuando proviene de los resultados de las experiencias del pasante como estudiante para docente. No hay evidencias para defender la proposición de que únicamente colocando al docente en formación en un ambiente real mejorará automáticamente su enseñanza. Hay evidencias para apoyar la noción de que las destrezas de enseñanza del estudiante para docente (en términos de las destrezas enfatizadas por el programa de formación) realmente se deterioran.

Para mejorar las destrezas de enseñanza debe haber objetivos, feedback sobre una base regular, y la oportunidad de mejorar. Esto implica que la práctica de las experiencias de enseñanza requiere ser supervisada, al menos en el sentido de que las experiencias deben ser observadas y la información recolectada, para suministrarle al docente en formación feedback acerca de la experiencia. Pero la supervisión también debe incluir la recolección *sistemática* de información que sea de utilidad. La supervisión que es hecha intuitivamente, con un poco más que tomar notas a modo de recolección de datos, es improbable que sea lo suficientemente poderosa como para que contribuya al mejoramiento docente. Esto no solo es una opinión. Es una afirmación respaldada por un sustancial cuerpo de investigación. La investigación sobre las formas tradicionales de supervisión es tan triste que Mosher y Purpel (1972, p. 50) concluyeron su revisión con la sentencia de que “la conclusión inevitable de cualquier revisión de la literatura es que virtualmente no hay alguna investigación que sugiera que la supervisión de la enseñanza, independientemente de cómo se defina o realice, hace alguna diferencia”.

Para que se produzca el mejoramiento en la enseñanza, se requiere su observación sistemática. Frecuentemente, esta responsabilidad recae en el supervisor universitario con experiencia en el campo o con experiencia en formación docente. Sin embargo, existe una amplia evidencia de que el docente colaborador de la escuela puede hacer las observaciones igual de bien. También hay evidencias de que los estudiantes pueden observarse mutuamente si ellos son ubicados en parejas o tríadas en sus experiencias prácticas. Debemos subrayar, sin embargo, que las observaciones producen registros de lo que sucedió, en lugar de juicios sobre ello. Las mujeres y hombres que se preparan para enseñar necesitan los registros de sus actuaciones y las de los estudiantes a los que están enseñando. Los registros en forma de datos observacionales pueden ser utilizados de muchas maneras, incluyendo reuniones con el supervisor, reuniones de tríadas con el supervisor y el docente colaborador, y para auto-evaluación. Más que juicios, se necesitan los registros.

En los años recientes ha quedado claro que los docentes pueden mejorar su enseñanza, a menudo dramática y rápidamente (Metzler, 1990; Randall, 1992; Siedentop, 1986). Para hacerlo, sin embargo, se requiere que ellos tengan objetivos de enseñanza específicos, observaciones de su actuación y la actuación de los estudiantes que sean relevantes para esos objetivos, e información – feedback – a partir de las observaciones. La data observacional se convierte en un registro utilizado por el docente en formación, el supervisor y el docente colaborador, para revisar el desempeño y sugerir formas de mejorarlo. La cantidad de instrumentos de observación disponible es sustancial, tanto es así que se dispone de un libro de

texto centrado únicamente en instrumentos para diversos fines en educación física y deportes (Darst, Zakrajsek, y Mancini, 1989).

La estrategia apropiada para optimizar la influencia de la observación y la supervisión es (a) decidir sobre los objetivos específicos que pueden estar relacionados con las observaciones de la actuación del docente o del alumno, (b) crear un protocolo de observaciones que se relacione específicamente con esos objetivos, (c) garantizar observaciones fiables, y (d) utilizar el registro de observaciones para reflexionar sobre el logro de los objetivos y sugerir estrategias de mejoramiento. La caja 16.1 muestra un rango de variables que se ha demostrado claramente que están relacionadas con la enseñanza efectiva. Todas estas variables pueden ser observadas por medio de los instrumentos sugeridos en este capítulo.

Las investigaciones han demostrado que todas las variables enumeradas a continuación están relacionadas con la enseñanza efectiva. También revise "Evaluación de la enseñanza y el aprendizaje tal como ocurren" en el Capítulo 1.

Variables del proceso del profesor

Indicaciones gerenciales por episodio gerencial
Reacciones conductuales positivas / negativas
Distribución de la atención - Niño/niña; alto/medio/bajo dominio de las destrezas
Declaraciones de expectativas
Análisis del feedback a las destrezas
Tiempo dedicado a explicaciones de las destrezas y estrategias.
Número de indicaciones administrativas de rutina al comienzo del año escolar
Claridad de las indicaciones sobre las tareas instruccionales
Secuencia de las tareas de refinamiento, extensión y aplicación

Variables del proceso del estudiante

Tiempo de espera
Tiempo dedicado a la tarea durante el proceso gerencial, instruccional y la práctica.
Respuestas a las tareas instruccionales (congruentes, modificadas, fuera de la tarea)
Análisis de las respuestas a las destrezas (adecuadas, con éxito, etc.)
Casos de comportamiento disruptivo
Ayuda y apoyo adecuados para los compañeros de clase

Unidades didácticas

Episodios gerenciales
Precisión del ciclo feedback del maestro-respuesta del estudiante
Latencia y congruencia de las respuestas iniciales a las tareas por parte del alumno
Relación entre los patrones de supervisión y el comportamiento en la tarea

Variables de criterio del proceso

Tiempo de aprendizaje académico en educación física
Oportunidad de responder (idoneidad y éxito de las respuestas)

Variables de resultado de los estudiantes a corto plazo

Puntuaciones en la prueba de condición física
Estadísticas de juego en torneos interclase de final de unidad
Puntajes en pruebas de conocimiento

Caja 16.1 Qué observar: Valoración de la enseñanza y sus resultados

La fiabilidad de la data observacional

Es de suma importancia la fiabilidad al recolectar los datos de observación. El término fiabilidad tiene muchos significados en la literatura científica; nosotros lo definimos como el grado en que dos personas, utilizando las mismas definiciones, mirando a la misma persona, al mismo momento, registran el mismo comportamiento. ¿Por qué la fiabilidad es importante? Suponga que su compañero o el instructor observa su enseñanza durante la primera semana de una experiencia docente y registra las interacciones en el comportamiento y el tipo de feedback que les da a los estudiantes. Esto se convierte en su desempeño de referencia. Suponga que se hace otra observación la semana siguiente y que sus valores en la interacción y el feedback sean considerablemente más bajos. Ahora, la pregunta es, “¿Quién cambió? ¿Usted o el observador?” Si el observador cambió (tal vez interpretando las definiciones en forma diferente, o tal vez no fue tan preciso), entonces tendrá una información errónea y quizá incluso una pobre evaluación. Esto es por lo que es importante que la data recolectada sea confiable, esto es, que razonablemente se pueda suponer que los datos reflejan fielmente lo que realmente sucedió durante el episodio de enseñanza. Este capítulo discute las técnicas de recolección de información confiable sobre la enseñanza.

La mayor parte de la recopilación de datos en las ciencias se realiza mediante sistemas automáticos de registro. Los fisiólogos del ejercicio registran automáticamente el ritmo cardíaco de los sujetos antes, durante y después del ejercicio. En este caso el desempeño o comportamiento está en la frecuencia cardíaca y se hace observable con la colocación de electrodos en la piel y la transmisión del impulso a una máquina que continua y automáticamente imprime las fluctuaciones del ritmo cardíaco. Los kinesiólogos recopilan la acción de los músculos de la misma manera. Estos datos son precisos y fiables, y proporcionan un conveniente registro permanente de algunos aspectos cruciales de la conducta humana. Pero los comportamientos importantes de la enseñanza no pueden ser registrados con la transmisión de impulsos por electrodos. Muy a menudo, los comportamientos en cuestión deben ser observados por otro ser humano. El comportamiento socialmente más significativo rara vez es conveniente ya que no es fácilmente observable, y esta falta de conveniencia crea problemas de medición que por lo general demanda el empleo de observadores humanos.

Cuando un humano es utilizado para observar la actuación de otro humano, se deben dar unos pasos para garantizar la fiabilidad de las observaciones. Si la psicología nos ha dicho algo en los últimos 50 años, es que los hechos y nuestra percepción de los hechos pueden diferir considerablemente. Si yo observo las destrezas de enseñanzas de alguien, yo puedo estar viéndolas desde mi experiencia histórica e interpretar las cosas de manera diferente a otra persona. Si durante las sesiones de observación subsecuentes detecto algún cambio en su desempeño docente, debo asegurarme que el cambio ocurrió en su desempeño y no en mis observaciones. Todos tendemos a ver lo que queremos ver, y somos particularmente susceptibles a la influencia de la sugestión. Si el docente colaborador le dice a su supervisor que usted ha “mejorado realmente” en algún aspecto de su enseñanza, todas las evidencias indican que su supervisor tenderá a ver que usted ha mejorado, así cualquier cambio haya ocurrido o no. Por ello, es importante recolectar datos que ofrezcan evidencias fiables de su progreso y que no sean susceptibles a los caprichos de la sugestión o a las distorsiones de la percepción que comúnmente afecta a los sistemas inapropiados de observación. Los métodos

de valoración de la fiabilidad de las observaciones se presentan más adelante, en este capítulo.

Métodos tradicionales de valoración de la enseñanza

Por muchos años, los formadores de los docentes y los investigadores de la enseñanza han intentado valorar la enseñanza y sus resultados a través de una variedad de métodos como el juicio intuitivo, “eyeballing”¹, el registro anecdótico, la lista de chequeo, o las escalas de calificación. En la investigación de la enseñanza, estos métodos fueron abandonados hace tiempo al demostrarse que no eran fiables ni válidos como medidas de la enseñanza. Pero por alguna razón, estos métodos aún son usados ampliamente como métodos para obtener datos sobre la enseñanza con propósitos de supervisión. Aunque los expertos y los textos constantemente advierten contra el uso de tales métodos, todavía se emplean con más frecuencia que la observación sistemática.

Juicio Intuitivo

Lo que nosotros llamamos *juicio intuitivo* implica que un supervisor o docente colaborador observa su enseñanza y luego emite un juicio completo acerca de lo que observó. Si el observador está experimentado tanto en la enseñanza como en la observación, tal compendio de juicios es a menudo preciso. La cuestión es: “¿Son valiosos para el mejoramiento de la enseñanza?” Nosotros pensamos que no. Los docentes en formación no necesitan estimaciones generales tanto como requieren de información específica de su desempeño. También creemos que el juicio intuitivo es el resultado de un enfoque concentrado en el docente y una falta de enfoque en lo que los estudiantes hacen en clase. Esto no significa que los observadores altamente experimentados y reflexivos no puedan exponer buenos juicios globales, sino que tales juicios son más útiles sólo cuando se emplean en conjunto con un registro desarrollado a partir de la observación sistemática.

Registro anecdótico

Si un observador toma notas sobre lo que sucede durante la observación de una clase y utiliza esas notas para discutir la sesión con usted, el método es referido como un *registro anecdótico*. Los registros anecdóticos pueden ser muy útiles si el observador describe cuidadosamente los eventos y evita su interpretación. Un buen registro anecdótico es sistemático, pero de modo cualitativo más que cuantitativo. Los registros anecdóticos que caen por debajo de este estándar son poco útiles y simplemente se convierten en una versión más sistemática del “eyeballing”.

El registro anecdótico cualitativo es particularmente útil en las últimas etapas de un programa de formación docente, cuando los docentes en formación ya han recibido muchos comentarios sobre sus destrezas de enseñanza y de las destrezas de los estudiantes a los que ellos enseñan. El registro anecdótico cualitativo, hecho por un profesional experimentado, puede revelar los elementos sutiles y complejos que contribuyen a las lecciones exitosas. Esta información suele ser valiosa para los estudiantes que ya han mejorado sus destrezas básicas

¹ Al ojo por ciento

de enseñanza y están listos para recibir información más compleja acerca de los eventos en clase. El registro anecdótico, sin embargo, debe estar relacionado con los objetivos del estudiante para una lección en particular, excepto que los objetivos puedan ser más amplios y complejos que para una tarea específica o para el comportamiento de los estudiantes.

“Eyeballing”

La forma más común de feedback utilizada en la preparación de los docentes es a lo que nos referimos como “*eyeballing*”. El supervisor o el docente colaborador simplemente observa lo que usted enseña en un período determinado. No se toman notas, no se utiliza una lista de chequeo, ningún dato es registrado. Después de la lección el supervisor discute con usted su desempeño en la enseñanza. Se pueden mencionar algunos incidentes muy específicos. Alguna información muy valiosa puede atraer su atención, pero es poco probable que cualquier información que se le transmita le ayude a mejorar sistemáticamente sus destrezas de enseñanza. Como cualquier método de observación, “*eyeballing*” es muy susceptible a errores de percepción debido a conceptos erróneos, historias previas o la sugestión. Con demasiada frecuencia, el feedback generado por “*eyeballing*” es insignificante e inútil para el mejoramiento de las destrezas de enseñanza.

“*Eyeballing*” puede ser una valiosa técnica si es usada en adición a un método sistemático de observación y registro del comportamiento. Tiene un valor potencial porque el observador por lo general es un profesional entrenado, un docente guía, que puede ver la complejidad de la interacción durante la enseñanza, que son tan sutiles como para que puedan ser recogidas en un formato del programa de observación sistemática. Pero si “*eyeballing*” es la fuente de observación principal mediante el cual se genera el feedback, entonces una gran cantidad de información fiable probablemente no estará disponible para usted.

Listas de Verificación y Escalas de Calificación

En el pasado, el método más común de observación sistemática fue el uso de listas de verificación. Una lista de verificación es una lista de enunciados o características acerca de las cuales un observador hace un juicio. Por lo general, el juicio es una decisión de si o no. Algunas veces ello implica el uso de una escala para que el espacio entre los puntos si-no sea calificado permitiendo un rango de posibles respuestas como *a menudo*, *algunas veces*, *poco frecuente*, *nunca*.

El método de lista de verificación tiene una dudosa ventaja y una serie de inconvenientes muy serios. La ventaja es que se presenta con la apariencia de un verdadero enfoque basado en datos para el mejoramiento de las destrezas de enseñanza. Al utilizar una lista de verificación como un instrumento de evaluación terminal, el supervisor le está dando una envoltura pseudocientífica a un enfoque de evaluación muy informal. Si la lista de verificación se utiliza en sesiones de observaciones sucesivas como una herramienta de aprendizaje más que, estrictamente, como una herramienta de valoración final, se obtiene la misma ventaja. El beneficio obtenido del uso de la lista de verificación se deriva principalmente del supervisor; si proporciona una falsa sensación de seguridad, la ilusión que el feedback le da al pasante se basa en lo que parece una evidencia dura, más que en un simple *eyeballing*”.

Las listas de verificación son notoriamente poco fiables. Los enunciados o características de las listas de verificación no están definidas suficientemente como para garantizar observaciones fiables. Dar una calificación acerca de la iniciativa mostrada por un pasante es prácticamente imposible a menos que la característica etiquetada como iniciativa se defina de manera que el pasante, el supervisor y otras partes interesadas tengan una comprensión común de su significado y algunos ejemplos sobre la iniciativa o la falta de ella.

Las escalas de calificación por lo general son ideadas por ser más precisas y sofisticadas al presentar una gran cantidad de puntos de selección. La escala de calificación mostrada aquí tiene nueve puntos de selección en un rango de siempre a nunca.

Siempre	1	2	3	4	5	6	7	8	9	Nunca
---------	---	---	---	---	---	---	---	---	---	-------

Este tipo de dispositivo de calificación es muy poco fiable. La ilusión de mayor precisión y calificación se obtiene a costa de fiabilidad. El hecho es que cuanto menos sean los puntos de selección, más fiables son las calificaciones. Sin embargo, esto se equilibra por el hecho de que pocos puntos de selección proporcionan información menos precisa. Uno se queda con un dilema: una elección entre información fiable e imprecisa, o información más precisa pero menos fiable. Ninguno de los dos es aceptable como formato primario de recolección de datos para un programa que se toma en serio ayudar a los pasantes a mejorar sus destrezas de enseñanza.

Las escalas de calificación son útiles cuando la información generada por simples puntos de selección es de suficiente valor para ayudar a mejorar las destrezas de enseñanza. En este caso, las escalas de calificación son rápidas, eficientes y fiables. Las listas de verificación son útiles para verificar las tareas completadas. Sirven como buenos recordatorios del número de tareas a completar, la naturaleza de estas tareas, y el momento en que se completan. Pero este uso de las listas de verificación solo es un poco más que el mantenimiento de registros y no debe verse como un sustituto de la verdadera recopilación de datos.

Métodos de observación sistemática

La observación sistemática de la enseñanza de los docentes ha revolucionado la investigación de la enseñanza y ha llegado a importantes descubrimientos acerca de la naturaleza de la enseñanza efectiva (ver particularmente el Capítulo 2). La observación sistemática es la base sobre la cual se ha construido la investigación de la enseñanza. También debe ser la base sobre la cual las destrezas de enseñanza se desarrollen. La observación sistemática es simple de hacer: solo requiere una comprensión básica y un poco de práctica. La data producida mediante la observación sistemática se convierte en la información utilizada para ayudar a los docentes a mejorar. En muchos casos, un simple resumen de los datos originales es todo lo que se necesita. Ningún análisis estadístico sofisticado se necesita. Suma, resta y división son todas las operaciones que se necesitan para desarrollar información muy útil y significativa sobre la enseñanza. Las principales técnicas de la observación sistemática son el registro

de eventos, registro de duración, registro de intervalos, muestreo del tiempo grupal, y recopilación de datos auto-registrados e integrados en la lección.

Estos métodos de observación y registro del comportamiento han sido utilizados extensamente en muchas áreas de la investigación que se relacionan con el comportamiento humano. Ya que se han utilizado extensamente, su fiabilidad está bien demostrada. Estos métodos también son incluidos porque son fáciles de aprender y utilizar. No requieren equipos más sofisticados que un cronómetro o una grabadora. Las técnicas han sido utilizadas fiablemente para investigadores, docentes y estudiantes (Siedentop, 1981).

El uso fiable de estos métodos depende de cuán bien se definan las diferentes categorías de desempeño. Conseguidas las definiciones adecuadas, los métodos son fáciles y fiables. Usualmente, se pueden aprender en una o dos sesiones prácticas. La mayoría de las dificultades del uso de los métodos surgen más de los problemas de definición de las categorías de desempeño que con los errores técnicos asociados a los sistemas de observación.

Registro de eventos

Una vez que una categoría de desempeño ha sido definida adecuadamente, se le puede observar de la manera más simple llevando un registro acumulativo de la cantidad de situaciones diferentes que ocurren dentro de un período específico de tiempo. Esto da como resultado un recuento de la frecuencia de los eventos a medida que ocurren (Hall, 1971). Su supervisor puede registrar la cantidad de interacciones positivas con los estudiantes. Su docente colaborador puede contar el número de veces que los estudiantes rompen las reglas específicas de la clase. Usted puede contar la cantidad de intentos que dos estudiantes tienen en una destreza durante una sesión de clase. El registro de eventos produce un resultado numérico que se puede convertir fácilmente en frecuencia por minuto. El valor de la conversión a frecuencia por minuto es que el desempeño en diferentes ocasiones puede ser comparado porque se clasifican en unidades comunes, esto es, frecuencias por minuto.

El *registro de eventos* es uno de los métodos más útiles de recolección de datos significativos; cualquier acción o reacción de un estudiante o un docente, o cualquier aspecto de la interacción entre ambos que pueda ser definido, puede ser medido por el recuento del número de veces que ocurre. A conceptos tales como cooperación, competición, esfuerzo competitivo, deportividad y agresividad, se les puede dar un nuevo significado definiéndolos en términos tales que puedan ser observados y luego contándolos tal como ocurren.

El registro de eventos se puede hacer continuamente; esto es, varias categorías del comportamiento docente pueden ser observadas mediante un registro de eventos en una sesión completa de enseñanza. La duración de una sesión es fácilmente determinada, y los datos pueden ser convertidos a una medida de frecuencia por minuto. A menudo es demasiado lento y fatigante hacer el registro de eventos durante una sesión completa. También, puede ser necesario hacer otras observaciones. Una medida válida del comportamiento docente se puede obtener efectuando un registro de los eventos en un corto período de tiempo y repitiéndolo en intervalos durante toda la sesión de enseñanza. Por ejemplo, es usualmente muy satisfactorio registrar los eventos en períodos de 3 minutos implementando cinco de esos intervalos en un período de enseñanza. Si los cinco intervalos de registro son distribuidos a

lo largo del período, se obtiene una muestra válida del comportamiento del docente aun cuando solo 15 minutos de la sesión fuese dedicada a la recolección de datos. Este concepto de muestreo del comportamiento más que su registro continuo es importante para cualquier enfoque basado en datos relacionado con el mejoramiento de la enseñanza.

Registro de duración

El registro de eventos es útil si se puede obtener una comprensión más significativa de un comportamiento al tener alguna idea de la frecuencia con la cual ocurre. Por ejemplo, una forma de comprender mejor la eficiencia de un ambiente de aprendizaje es tener alguna idea del número de intentos que los estudiantes necesitan para practicar una destreza. Los intervalos colocados periódicamente en el registro de eventos a lo largo de una sesión generan una información fiable y válida que puede ser vista como el número de intentos por minuto.

Sin embargo, algunas veces la frecuencia de un comportamiento no produce la información más útil. Suponga que usted quiere conseguir una medida del grado de participación de un estudiante en clase. La primera tarea debería ser la definición de participación y no participación. No sería apropiado utilizar el registro de eventos para estudiar la participación. Una sola participación (un evento) podría durar mucho tiempo, y otra participación (otro evento) podría durar un menor tiempo. Conocer que ocurrieron dos eventos no ayudaría a comprender la frecuencia de la participación de un estudiante. Sería mucho mejor registrar la cantidad de tiempo que un estudiante pasa en la actividad que usted ha definido como participación. Un cronómetro podría encenderse y apagarse de acuerdo con la participación del estudiante; la medida más precisa de participación sería el tiempo resultante acumulado.

El *registro de duración* utiliza el tiempo como una medida del comportamiento. Los datos originales derivados del registro de duración se expresan en minutos y segundos. Un estudiante podría participar un total de 21 minutos y 30 segundos en una clase de 30 minutos. Estos datos sin procesar se pueden convertir a una cifra porcentual que permita la comparación entre estudiantes y entre varias sesiones. Los datos se convierten dividiendo el tiempo total del registro en una sesión por el tiempo que se tiene para realizar el registro. La medida resultante se expresa como un porcentaje del tiempo total empleado en participar.

Al igual que en el registro de eventos, es a menudo ineficiente efectuar el registro de duración en forma continua, en una sesión completa de enseñanza; en el registro de duración también se deben tomar muestras. Tres muestras espaciadas regularmente de 5 minutos empleando el registro de duración en una sesión de enseñanza, proporcionan una información válida acerca del porcentaje de tiempo empleado en cualquier comportamiento definido. En este caso, la cifra porcentual se deriva del tiempo total de los intervalos registrados antes que del tiempo total de la clase. El resultado permanece como un porcentaje del tiempo empleado en una categoría definida de comportamiento.

El registro de duración es útil para cualquier categoría de comportamiento en el que el tiempo dedicado a ese comportamiento proporciona la mejor estimación de su importancia. Esto es cierto tanto para el comportamiento del docente por la cantidad de tiempo dándole instrucciones a la clase como para el comportamiento del estudiante por la cantidad de tiempo empleado en actividades administrativas o en participación.

Registro de intervalo

El término *registro de intervalo* se refiere a la observación del comportamiento durante cortos períodos de tiempo (intervalos) y a la decisión de cuál comportamiento caracteriza mejor ese período de tiempo. Por ejemplo, el tiempo total del período podría ser dividido en intervalos de 10 segundos. En el primer intervalo se observa al docente. En el segundo intervalo de 10 segundos el observador registra la categoría de comportamiento que mejor representa lo que acaba de observar. En el registro de intervalos, los intervalos consecutivos son utilizados primero para observar y luego para registrar. Los intervalos deben ser cortos, usualmente no mayores de 20 segundos y algunas veces tan pequeños como 6 segundos. El intervalo de observación no tiene que ser de la misma duración que el intervalo de registro. Usualmente, el intervalo de registro puede ser más corto, especialmente a medida que los observadores se hacen más diestros o cuando se utiliza un pequeño sistema de categorías en el cual se necesitan unas pocas decisiones.

Los datos generados por el registro de intervalos se expresan como un porcentaje de los intervalos en los que ocurre cada comportamiento. Sin embargo, ya que el intervalo representa una media precisa del tiempo, la técnica del intervalo también puede ser utilizada para estimar las implicaciones del tiempo. El registro de los intervalos tiene la ventaja de poseer una alta fiabilidad. Las instrucciones tal como cuándo observar y cuándo registrar, pueden ser pre-programadas con la tecnología del caso en equipos destinados a tal fin.

El registro de intervalo ha sido utilizado exitosamente para observar el comportamiento del docente, el comportamiento del estudiante, y en medidas como el tiempo de aprendizaje académico. Los observadores deben esforzarse en utilizar unos intervalos tan cortos como les sea posible, para obtener unos datos fiables. Los únicos problemas al utilizar la técnica del intervalo son generalmente encontrados cuando los intervalos son tan largos que pueden ocurrir varios comportamientos y el observador tiene dificultades con el tiempo para decidir cuál comportamiento registrar. Un intervalo corto (6 – 12 segundos) normalmente evita este problema. Si se utiliza un sistema de intervalos de 10 segundos para observar y 10 segundos para registrar, habrá una data puntual reunida cada 20 segundos, 3 por minuto, y 90 por 30 minutos en un episodio de enseñanza. El total de 90 datos puntuales es generalmente suficiente para garantizar la validez del comportamiento que está siendo observado; esto es, lo que está registrado representa fielmente lo que realmente sucedió en el contexto de la sesión.

Muestreo del Tiempo Grupal

Una técnica utilizada para reunir datos periódicos de todos los miembros de un grupo (típicamente una clase o un subconjunto específico de una clase) es el *muestreo del tiempo grupal*. El muestreo del tiempo grupal también ha sido referido como el *registro “placheck”* (planned activity check = chequeo de la actividad planificada). A intervalos regulares durante la observación de la sesión, el observador rápidamente escanea al grupo y cuenta el número de estudiantes participando en la categoría de comportamiento que le interesa. Este escaneo típicamente no le toma más de 10 segundos, incluso con una clase bastante numerosa. Un pequeño grupo podría ser escaneado en 5 segundos. Una vez que un estudiante es escaneado, el observador no retorna a ese estudiante aún si ha cambiado de comportamiento. La meta es

observar a cada individuo en un momento dado y registrar la cantidad del grupo total participando en una categoría de comportamiento en particular. Las observaciones del comportamiento por categorías como esfuerzo, participación, productividad y comportamiento apropiado se prestan bien a la técnica de muestreo de tiempo grupal. Medidas periódicas de las variables de criterio como el tiempo de aprendizaje académico, podrían tomarse de esta manera.

El método de muestreo del tiempo grupal se utiliza como sigue. El observador siempre escanea en una dirección específica, generalmente de izquierda a derecha. Para el escaneo se toma un tiempo de duración específico (usualmente 10 segundos). Sin embargo, ello depende de la cantidad total del grupo ya que con grupos grandes se requiere progresivamente más tiempo para completar el escaneo. Se cuenta la cantidad de individuos activos en la categoría del comportamiento observada. Siempre es más fácil calificar la categoría de comportamiento en la que participan menos personas. Por ejemplo, si usted está calificando el comportamiento productivo e improductivo, y muchos estudiantes están participando en comportamientos productivos, es más fácil contar los que están demostrando comportamientos improductivos. La cantidad de participantes en comportamientos productivos pueden ser calculados simplemente restando la cantidad de participantes en comportamientos improductivos del total de participantes. De nuevo, es mejor convertir esos datos originales en cifras porcentuales. Esto se hace anotando el tamaño del grupo total y dividiendo el total de la cifra correspondiente². Así, puede calcularse fácilmente el porcentaje de estudiantes participando en comportamientos productivos dividiendo su cantidad por la cantidad total de estudiantes.

La muestra del tiempo grupal debe ser espaciada regularmente a través de la sesión de clase. Ya que les toma solo 10 segundos realizarlo, ello no requiere evitar que el observador haga otras observaciones durante el tiempo entre intervalos. Ocho muestras del tiempo grupal espaciadas de manera uniforme durante una sesión de 40 minutos, tomará solo 1 minuto y 20 segundos de tiempo de observación y, sin embargo, proporcionará información válida sobre el comportamiento del grupo.

Auto-registro y Recolección de Datos Incorporados a la Lección

La información acerca del docente y sus estudiantes puede ser recolectada como una parte regular del proceso de enseñanza. Esto es particularmente cierto para la estimación de la participación del estudiante. Los estudiantes que registran sus propias oportunidades o las de sus compañeros para responder (OTRs³) con relación a las tareas instruccionales producen un registro útil. Los modelos de enseñanza recíproca y tutoría del compañero (ver Capítulo 14) crean situaciones en las cuales los estudiantes pueden registrar en esfuerzo de su compañero en la completación de una tarea. Los formatos de pequeños grupos a menudo incluyen un registrador quien podría funcionar con ese propósito. Los formatos de Educación Deportiva incluyen la recolección regular del desempeño en competencias que podrían ser utilizados con propósitos de supervisión. Dado que hemos defendido a lo largo de este libro que la enseñanza efectiva se analiza mejor con referencia a la participación del estudiante en la

² Cantidad de comportamientos productivos.

³ Opportunities To Respond

lección, estas fuentes de datos integradas en la lección pueden ser indicadores poderosos del mejoramiento de la enseñanza.

Los docentes también pueden auto-registrar su comportamiento, y hacerlo de manera que no distraiga la clase. Un equipo pequeño de grabación puede ser utilizado para registrar su comportamiento verbal durante la clase. Más tarde se puede codificar la grabación para producir un registro de las destrezas como descripción de las tareas, indicaciones, tipo de feedback, o los comentarios de apoyo. Muchas escuelas tienen equipos de grabación de videos. Ocasionalmente, la producción de un video de la clase puede ayudarle a aprender de sí mismo y de sus estudiantes. Cuando se use un sistema de video, es mejor utilizar un micrófono inalámbrico para capturar en la grabación su comportamiento verbal. Los cronógrafos de muñeca se pueden utilizar para cronometrar los segmentos más importantes de la clase, permitiendo un registro completo del tiempo administrativo, instruccional y de práctica. Se puede llevar el conteo de los eventos en un portapapeles o en un pequeño cuaderno, Algunos docentes utilizan contadores de muñeca para contabilizar los eventos.

Combinando las técnicas de observación en un sistema

La selección de una técnica de observación debe tener sentido en términos del comportamiento que está siendo observado. El feedback de un docente a un estudiante es mejor observarlo con el registro de eventos –la medida más significativa de feedback es como una tasa (número de feedback por minuto o por lección de enseñanza de 30 minutos), o como una proporción (porcentaje del total de feedback que se entregaron con precisión o porcentaje del total de feedback que tenían contenido con información específica). La información útil acerca del feedback también puede ser desarrollada mediante el formato de intervalo. Pero la duración del registro no da información útil sobre el feedback. Conocer la duración de una interacción de feedback no nos dice mucho al respecto.

Las oportunidades de aprendizaje de un estudiante pueden ser ponderadas como una medida de tiempo utilizando el registro de duración o el registro de intervalo. Así, la cantidad total de tiempo empleado en el aprendizaje activo es una parte muy significativa de la información. También lo es el porcentaje de intervalos en los que un estudiante participa en el tiempo de aprendizaje académico. Las oportunidades de aprendizaje del estudiante también pueden ser observadas mediante el registro de eventos, acumuladas al contabilizar el número de intentos que suceden entre los miembros de una clase durante 30 minutos.

Las variables como el comportamiento del estudiante en la tarea pueden ser observadas de varias maneras, a través del registro de eventos (número de ocasiones de un comportamiento fuera de la tarea), registro de duración (porcentaje del tiempo total empleado en la tarea), registro de intervalos (porcentaje del total de intervalos empleado en la tarea), o la muestra del tiempo grupal (porcentaje de estudiantes fuera de la tarea). La elección de cuál usar estará, con frecuencia, dictada por consideraciones como la fiabilidad y el uso económico del tiempo del observador. La meta debe ser obtener la mayor cantidad de data fiable posible con el más eficiente uso del tiempo del observador (utilizando un menor tiempo para

obtener una data fiable sobre el comportamiento en la tarea permite tener más tiempo para observar otras cosas).

Muchos sistemas de información incorporan más de un tipo de técnica de observación. Esto puede ser hecho muy fácilmente a través del muestreo de varios comportamientos en cuestión. La meta debe ser muestrearlos regularmente y tener las muestras distribuidas a lo largo del período establecido. Si toda la data del comportamiento docente fue recolectada en los primeros 15 minutos de un episodio de enseñanza y toda la data del comportamiento de los estudiantes en los 15 minutos subsiguientes, los datos no arrojarían una imagen real de lo que ocurre en todo el período. Es mucho mejor muestrear una pequeña cantidad del comportamiento docente, luego pasar al comportamiento del estudiante, luego regresar al comportamiento docente, y así sucesivamente hasta completar el período de observación. Por ejemplo, el siguiente ciclo de 4 minutos de observación utiliza el registro de eventos para el comportamiento docente, y tanto el muestreo del tiempo grupal en la tarea como el tiempo de aprendizaje del estudiante:

Tiempo	Técnicas
10 segundos	Muestra del tiempo grupal, número de estudiantes fuera de la tarea
10 segundos	Registro
10 segundos	Muestra del tiempo grupal, número de estudiantes en ALT-PE ⁴
10 segundos	Registro
20 segundos	Descanso
60 segundos	Registro de eventos, comportamiento del docente
10 segundos	Muestra del tiempo grupal, número de estudiantes fuera de la tarea
10 segundos	Registro
10 segundos	Muestra del tiempo grupal, número de estudiantes en ALT-PE
10 segundos	Registro
20 segundos	Descanso
60 segundos	Registro de eventos, comportamiento del docente
4 minutos	Tiempo total

Con este formato, el ciclo de 4 minutos podría repetirse siete veces en un lapso de observación de 30 minutos, dejando 2 minutos extra para el descanso. Los siete ciclos podrían producir 14 muestras de tiempo grupal para el comportamiento fuera de la tarea, 14 muestras de tiempo de aprendizaje académico en educación física, y 14 muestras de registro de eventos con datos del comportamiento docente. Ya que cada ciclo fue adecuadamente distribuido a lo largo del tiempo total, todos juntos representan una imagen muy fiel de lo que realmente sucedió.

Todo esto es cumplido solo con un observador. A ese observador se le ayuda enormemente (y los datos se vuelven más fiables) si los patrones particulares de observación y registro son pre-programados con claves simples como “Observe los estudiantes fuera de la tarea”, “Registre los estudiantes fuera de la tarea”, “Inicie el registro de evento de 1 minuto del comportamiento del docente”, “Finalice el registro del evento del comportamiento docente” y “Comience 1 minuto de descanso”. Se deben utilizar unos audífonos para que las claves sean escuchadas solo por el observador y no sean un distractor en el sitio donde se

⁴ Tiempo de Aprendizaje Académico en Educación Física (ALT-PE)

están realizando las observaciones. Idealmente, un equipo pequeño se coloca en el bolsillo y tan discreto como sea posible.

Con el muestreo del comportamiento en lugar de observarlo continuamente y con el desarrollo de sistemas de observación de múltiples técnicas, la cantidad de información recolectada por un observador puede ser sustancial. Este observador puede ser un instructor, un compañero pasante, un estudiante ayudante, un docente colaborador, o el propio docente si se hace un video de la sesión.

Decisiones importantes para desarrollar una estrategia de observación

Un sistema de observación debe servir a un propósito específico. A menudo un docente, o los estudiantes y profesores de un programa de formación docente, podrían querer producir información de unas pocas variables específicas del docente o del estudiante. Si es así, un sistema corto y específico debe adaptarse a los objetivos de esa experiencia. En otra ocasión, se necesita un sistema de observación más completo, pero también debe ser desarrollado para que produzca una información relativa a los objetivos más completos. El punto está claro: usted debe desarrollar un sistema de información con base en los objetivos establecidos, antes que seleccionar un sistema existente, y luego permitir que los objetivos se desarrollen a partir de esa selección.

Una vez que sean comprendidas y dominadas las técnicas observacionales de evento, duración, intervalo y muestreo del tiempo, la adaptación de un sistema de observación existente o el desarrollo de nuevos sistemas se hace muy fácil. Si se presta la debida atención al muestreo del comportamiento adecuado y si las categorías están bien definidas, los sistemas de observación pueden ser desarrollados y refinados a partir de lo más básico a lo más complejo. Estos son los pasos que deben ser seguidos para el desarrollo de sistemas observacionales:

1. Defina los objetivos para los que se utilizarán los datos de la observación. Mientras más cuidadosa y específicamente sea hecho, más fáciles serán los pasos de este proceso. Suponga que usted quiere obtener información acerca del feedback docente. ¿Puede usted intentar valorar la precisión del feedback (por ejemplo, el grado en el cual la información transmitida al estudiante se basó en un diagnóstico preciso del desempeño)? ¿Son suficientes las subcategorías de positivo general, positivo específico, y correctivo? ¿Quiere usted subdividirlas aún más para descubrir cuántas expresiones de feedback están dirigidas a niñas y niños, o a estudiantes de poca y mucha destreza?
2. Decida cuáles comportamientos de los docentes y/o estudiantes le dará más información válida con relación a sus objetivos. Este paso es bastante fácil si el objetivo es producir información relativa al tiempo de clase empleado en actividades administrativas y transicionales. Es más difícil si usted quiere producir una medida del clima de la clase o del entusiasmo del docente. En esta parte del proceso, muy explícito, las definiciones del comportamiento son extremadamente importantes. Los problemas de

la producción de observaciones fiables casi siempre están relacionados con definiciones incompletas o difusas de las categorías de comportamiento.

3. Decida cuál técnica de observación utilizar. Algunos comportamientos como las indicaciones del docente claramente conducen, por si mismas, al registro de eventos; otros comportamientos como el estudiante en espera, conducen al registro de duración. El registro de intervalos puede ser utilizado en casi todas las categorías de comportamiento. Por lo tanto, a medida que aumenta el número de objetivos a alcanzar a través de un sistema de observación, deben tomarse decisiones acerca de la combinación de técnicas de observación que mejor produzcan un registro válido de comportamientos relacionados con esos objetivos. El formato de codificación para las destrezas gerenciales avanzadas (ver Figura 16.7) y el instrumento de supervisión general mostrado adelante en este capítulo (ver Figura 16.10) utilizan varias técnicas de observación combinadas para producir una cantidad de conjuntos diferentes de datos.
4. Decida cuánto puede ser observado fiablemente. La complejidad de la observación está relacionada con la cantidad de decisiones que el observador tiene que tomar de acuerdo con su experiencia en la observación. Observar el sistema de comportamiento de un docente de categoría 5⁵ es bastante fácil, incluso para un observador principiante. El sistema de comportamiento de un docente de categoría 24 es mucho más difícil porque cada vez que el docente hace algo, el observador tiene que decidir, entre muchas categorías, cuál categoría utilizar para registrar ese comportamiento. Si el registro de eventos y el registro de duración están siendo utilizados, la complejidad se incrementa porque la atención del observador puede estar enfocada en dos cosas a la vez. Generalmente, es más importante producir menos datos que sean altamente fiables que producir más datos con baja fiabilidad. La convención generalmente aceptada en la investigación aplicada es que los observadores independientes deben lograr un 80% de acuerdo para que los datos sean considerados fiables.

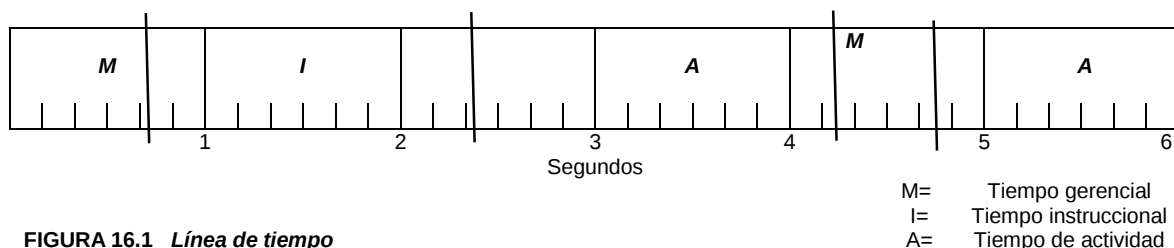
Construcción del Sistema de Observación

El primer paso para la construcción de un sistema de observación es seleccionar las variables a ser observadas. Este proceso fue explicado en la sección precedente. El segundo paso es seleccionar la técnica de observación, esto es, decidir si las variables seleccionadas serían mejor observadas a través del registro de eventos, el registro de duración el registro de intervalos o la muestra del tiempo grupal. Esta decisión se toma sobre la base de dos factores: (a) el ajuste entre la técnica y la variable, y (b) la integración de varias técnicas en un sistema total. Si el sistema de observación es pequeño, limitado a solo una o a unas pocas variables, entonces el ajuste entre la técnica y las variables debe guiar la decisión. Esto significa, primero, que la mayoría de los comportamientos docentes se observan mejor a través del registro de eventos. Segundo, ciertas unidades analíticas como los episodios gerenciales se observan de manera más significativa a través del registro de duración (simplemente contar la frecuencia de los episodios gerenciales no revela mucho de lo que interesa; lo que cuenta es su duración). Tercero, un criterio general de una variable de proceso como el tiempo de

⁵ Según la clasificación de los docentes en USA

Pero si hay varias variables que observar, entonces el sistema debe ser construido de manera que permita al observador hacer varias cosas. En este caso, las medidas del ALT-PE podrían obtenerse mejor con el muestreo del tiempo grupal que requiere menos tiempo del observador que efectúa un registro de intervalos. El tiempo ahorrado puede ser utilizado observando otras variables mediante otras técnicas.

Una porción de la línea de tiempo se muestra en la Figura 16.1. Esta línea de tiempo tiene una duración de 6 minutos y cada minuto se divide en unidades de 10 segundos. El observador simplemente traza una línea que cruza la línea de tiempo cuando una actividad, como un episodio gerencial, finaliza y se inicia una actividad instruccional, o cuando una actividad de transición termina y comienza una actividad práctica. Posteriormente se puede revisar la línea de tiempo para calcular el tiempo total empleado en esas actividades. El método es simple y eficiente. También es un formato útil para proporcionar feedback ya que el docente puede ver realmente el flujo de actividades tal como ocurrieron en clase.



Al igual que los instrumentos de observación, las hojas de codificación pueden ser desarrolladas con propósitos específicos, locales. Es improbable que una hoja de codificación desarrollada en un lugar sea exactamente lo que se requiere en otro lugar. La clave es desarrollar un instrumento que refleje con tanta precisión como sea posible, las decisiones tomadas con respecto a los objetivos del episodio de enseñanza y las técnicas a través de las cuales las variables serán observadas.

Las hojas de codificación como el instrumento de supervisión general mostrado en la Figura 16.10 proporciona una gran cantidad de información en una hoja. Se utilizan varias técnicas y se provee un espacio en el cual se resume la información. Esto le suministra un feedback útil al docente y una manera conveniente de almacenar la data recolectada sobre la enseñanza como referencia para el futuro. Siempre debe haber un área en la propia hoja de codificación (ya sea en la parte superior o en la posterior) para facilitar información relevante tal como la concerniente al docente, ambiente, estudiantes, y la duración del período total de observación. Si esta información se incluye, la hoja de codificación completada se convierte en un valioso registro que puede ser utilizado para feedback, investigación y, eventualmente, para el establecimiento de objetivos y expectativas realistas de futuras experiencias.

Efectuando las observaciones

La selección de variables relevantes, ajustando las técnicas de observación a las variables, y desarrollando la hoja de codificación para transferir las observaciones a un registro permanente, representa la verdadera construcción de un instrumento de observación. Lo que queda es ponerlo en práctica para garantizar que se puede usar de manera fiable.

La pre-programación de un audio con indicaciones para el observador es una importante contribución a la observación sistemática, especialmente para sistemas que utilizan diferentes técnicas, muestras del comportamiento, o sean sistemas de intervalo. Por supuesto, si usted está utilizando un instrumento de registro de eventos en una categoría cinco para monitorear el feedback del docente, la función del audio es innecesaria. Pero si los observadores lo necesitan para una muestra del comportamiento, cambiando la observación de una variable a otra, o cambiando del docente a los estudiantes, las indicaciones pre-programadas pueden simplificar enormemente la tarea del observador y, por ende, incrementar las oportunidades de obtener una data fiable.

El observador debe ser tan discreto como le sea posible y, aun así, estar en posición de obtener los datos necesarios. Si el comportamiento del docente está siendo observado, entonces el observador puede planificar cómo moverse alrededor, manteniendo una proximidad cercana al docente, sin llamar la atención tanto como le sea posible. Los estudiantes reaccionarán ante la presencia de un observador, pero reaccionarán cada vez menos cuánto más tiempo esté el observador en el ambiente. Los docentes también reaccionarán y al principio tenderán a demostrar su “mejor comportamiento”. Sin embargo, a través del tiempo las demandas de la enseñanza son tales que los docentes propenderán a comportarse con más normalidad, atendiendo las necesidades de sus estudiantes y la situación inmediata, más que pensando constantemente en que está siendo observado. Aun así, debe lograrse una observación discreta.

El observador debe tener todo lo que se necesita para completar las observaciones incluyendo, (a) Cantidad extra de hojas de codificación, (b) cantidad extra de lápices, (c) tabla de anotación u otra superficie firme para colocar la hoja de codificación, (d) baterías recargadas para el equipo de reproducción del audio, y (e) copias de todas las definiciones de las categorías bajo observación. El observador debe llegar con tiempo de antelación y estar preparado antes del comienzo de la sesión de enseñanza. Los datos se pueden resumir mientras el profesor termina la clase y se encarga de supervisar el grupo después de la clase.

Posteriormente, los datos pueden ser utilizados como feedback cuando el docente tenga el tiempo disponible para estudiarlos y reaccionar ante ellos. La conferencia de supervisión debe estar basada en los datos, utilizándolos como la base para examinar e interpretar lo que sucedió en la sesión.

Entrenamiento de los observadores

Hemos enfatizado en que las destrezas observacionales son fácilmente adquiridas y que las observaciones fiables pueden realizarlas colegas, instructores, docentes colaboradores, supervisores, o los propios docentes si la sesión de enseñanza es video-grabada. En cada caso el observador, sin importar quien sea, necesita tener suficiente entrenamiento para garantizar una recolección de datos con estándares mínimos de fiabilidad. La mayoría de las técnicas de observación descritas en este capítulo y los sistemas de observación mostrados como ejemplos pueden ser aprendidos con un adecuado grado de fiabilidad en un corto período de tiempo, a menudo tan corto como de 2 a 4 horas. Los pasos del entrenamiento de los observadores son claros y son ampliamente aceptados en la literatura concerniente a la observación sistemática:

1. Los observadores se aprenden las definiciones de los materiales impresos. Las definiciones deben tener suficientes ejemplos para que las distinciones entre las categorías estén claras. Las buenas definiciones son el único y más importante ingrediente de la recolección de data fiable. Cuando los observadores tienen problemas, la dificultad casi siempre está localizada en los problemas definicionales.
2. Los observadores estudian una transcripción escrita de la lección de enseñanza y categorizan los comportamientos a partir de esa transcripción. Este proceso está concebido como un “trabajo para la casa” y tiende a muchos errores y a corregir los conceptos erróneos.
3. Si es útil, los observadores discuten las categorías, los ejemplos, las transcripciones, y demás elementos.
4. Los observadores practican haciendo observaciones a un video. Este debe haber sido codificado por un observador experimentado para que el observador en entrenamiento pueda comparar sus datos con los del observador experimentado. Este proceso ayuda a establecer la precisión del observador. El término *precisión del observador* se refiere al grado con el cual un observador concuerda con el estándar pre-codificado.
5. Los observadores practican en el campo. Los observadores siempre deben practicar en pareja para que la fiabilidad inter-observadores pueda ser calculada. El término *fiabilidad inter-observadores* se refiere al grado con el cual dos observadores independientes trabajan con las mismas definiciones, visualizan los mismos sujetos al mismo tiempo y registran datos similares. Los dos observadores (o más) luego pueden discutir las discrepancias y resolver los problemas.
6. A lo largo del proceso de capacitación, se debe mantener un registro de las decisiones. Un *registro de decisiones* es un registro de los desacuerdos entre los observadores y cómo se resolvieron. Periódicamente, las definiciones deben ser revisadas a la luz de la información desarrollada mediante el registro de decisiones y los cambios realizados para ajustar las decisiones concernientes a cómo manejar diversas situaciones.

7. Los observadores practican hasta que cumplan con un estándar mínimo de fiabilidad. En la investigación de la observación del comportamiento, utilizando técnicas de cálculo del acuerdo inter-observador (ver próxima sección), se requiere que el observador alcance un criterio del 80 por ciento antes de que pueda comenzar a recopilar datos que se utilizan para investigar. Un requerimiento un poco menos estricto sin duda es aceptable si el propósito de la observación es generar feedback para los docentes en formación.
8. A menudo, la fiabilidad debe ser chequeada para garantizar que los observadores están aplicando los códigos con precisión. Esto es análogo a la calibración de la escala de peso regularmente. El observador debe ser “calibrado” también.

Cálculo de la fiabilidad de los datos de observación

Hay varias razones por las que es importante garantizar que las observaciones sean fiables. Primero, las observaciones fiables indican si la definición de la categoría de desempeño de un docente o de un estudiante es suficientemente clara y adecuada, o no. Una categoría de desempeño del docente o del estudiante con mala definición, casi garantiza unas observaciones poco fiables. Cuando un chequeo de la fiabilidad indica que hay baja fiabilidad, la situación casi siempre se soluciona clarificando la definición de las categorías de desempeño.

Una segunda razón para la estimación de la fiabilidad es sentirse seguro de que los cambios observados en el docente o en los estudiantes se deben a ellos mismos y no al observador. Los observadores tienden a menudo a ver lo que ellos quieren ver ya sea consciente o inconscientemente, sesgando las observaciones por sus sentimientos acerca de cómo quieren que la experiencia resulte. Esto no significa que un observador sea poco profesional o incompetente. Esto significa que todos estamos sujetos a esta tendencia.

Una tercera razón para chequear la fiabilidad es garantizar que los cambios vistos por medio de las observaciones reflejan lo que está sucediendo en la clase. Si las observaciones indican que un pasante disminuye el tiempo gerencial de una semana a la otra, puede sentirse orgulloso de su mejoramiento en la medida en que las observaciones sean fiables. Si no son fiables, entonces los resultados apenas difieren de los generados por “eyeballing”.

Hemos enfatizado en un enfoque basado en datos para el mejoramiento de las destrezas de enseñanza. Si este modelo es utilizado y ocurren ciertas mejorías en sus destrezas de enseñanza, es crucial que usted, la escuela donde usted trabaja y el colegio o universidad en la que usted estudia tengan confianza en que esas mejorías son reales. Tal confianza está directamente relacionada con la fiabilidad de los registros hechos durante la experiencia de la pasantía. El término fiabilidad se refiere al grado mediante el cual observadores independientes concuerdan en lo que ven y registran. En este sentido, el significado del término *observadores independientes* puede entenderse como la imposibilidad de que un observador pueda detectar los registros que está haciendo otro observador. Por lo general este criterio se satisface ubicando a los observadores suficientemente separados para que ni señales visuales ni auditivas puedan ser utilizadas para detectar las observaciones que están haciendo. Si se utiliza una grabación del formato de codificación, se pueden realizar verificaciones de la

fiabilidad con audífonos adicionales y suficiente cable para que los observadores se sienten aproximadamente a 10 pies (3 metros) de distancia.

La fórmula general de computación de la fiabilidad es:

$$\frac{\text{Acuerdos}}{\text{Acuerdos} + \text{desacuerdos}} * 100 = \% \text{ del acuerdo}$$

Para el registro de eventos, el registro de duración, o la medición permanente de un producto, la fiabilidad puede ser calculada mediante la división de los datos del observador que tiene el menor número de casos o tiempo por los datos del observador que tiene el número más alto de casos o tiempo.

Si el registro de eventos se utiliza para juzgar el número de interacciones sociales que un estudiante tiene durante un período, y un observador registra 14 mientras un segundo observador registra 12, la fiabilidad podría ser computada como sigue:

$$\frac{12}{14} * 100 = 86 \% \text{ de fiabilidad}$$

Si se está utilizando el registro de duración para medir la cantidad de tiempo que un docente emplea hablándole a la clase, y un observador registra 12:30 mientras un segundo observador registra 13:10, la fiabilidad puede ser computada como sigue:

$$\begin{aligned} 12:30 &= 750 \text{ segundos} \\ 13:10 &= 790 \text{ segundos} \end{aligned}$$

$$\frac{750}{790} * 100 = 95\% \text{ de fiabilidad}$$

Si los estudiantes auto-registran su asistencia al ingresar a la clase, el docente, el estudiante auxiliar o el observador pueden verificar discretamente cómo ellos registran su asistencia. Si un observador estuvo de acuerdo con cada estudiante que se registró, la fiabilidad sería del 100 por ciento. Si en una clase de 30 estudiantes el monitor no estuvo de acuerdo con el registro de 2 estudiantes, la fiabilidad se calcula de la siguiente manera:

$$\frac{28}{30} * 100 = 93\%$$

Para los registros del intervalo y del muestreo del tiempo grupal, la fiabilidad está determinada por la estimación del grado en el cual los observadores independientes están de acuerdo o no en los registros de cada intervalo o del muestreo. Suponga que el registro de intervalo para calificar el grado en el cual dos estudiantes estuvieron comprometidos en el comportamiento de aprendizaje productivo durante la clase de educación física. Los datos originales de la muestra del tiempo aparecen como en la Tabla 16.1.

Para computar la fiabilidad, las observaciones por intervalo para cada estudiante deben ser comparadas. Cualquier intervalo en el cual los observadores hayan registrado la misma puntuación indica un acuerdo. Cualquier intervalo en el cual los observadores hayan registrado diferente puntuación indica un desacuerdo. Si los datos originales son reorganizados para que las observaciones de cada estudiante por parte de los dos observadores puedan ser comparadas, los acuerdos y los desacuerdos se hacen evidentes de inmediato.

TABLA 16.1 Muestra de los Datos de Observadores Independientes de dos Estudiantes

Intervalo	Observador A		Observador B	
	Estudiante 1	Estudiante 2	Estudiante 1	Estudiante 2
1	I	P	I	P
2	I	I	I	P
3	P	P	P	P
4	I	P	P	P
5	P	P	I	I
6	P	P	I	P
7	P	I	P	I
8	P	P	I	P
9	U	I	P	I
10	P	P	P	P
11	P	P	P	P
12	I	P	P	P

Nota: P= Productivo
I= Improductivo

La Tabla 16.2 muestra que las observaciones sombreadas están claramente en desacuerdo. Para el estudiante 1 hay siete intervalos en desacuerdo. Para el estudiante 2 se indican dos desacuerdos. Con 12 intervalos en total, la fiabilidad se calcularía como sigue:

TABLA 16.2 Muestra de los datos registrados con los desacuerdos entre los observadores

Intervalo	Estudiante 1		Estudiante 2	
	Observador A	Observador B	Observador A	Observador B
1	I	I	P	P
2	I	I	I	P
3	P	P	P	P
4	I	P	P	P
5	P	I	P	I
6	P	I	P	P
7	P	I	I	I
8	P	I	P	P
9	I	P	I	I
10	P	P	P	P
11	P	P	P	P
12	I	P	P	P

Fiabilidad del estudiante 1

Fiabilidad del estudiante 2

$$\frac{5}{5 + 7} * 100 = 42\%$$

$$\frac{10}{10 + 2} * 100 = 83\%$$

Hay claramente un problema con estos datos. Generalmente se requiere una fiabilidad del 80 por ciento con propósitos de investigación. Con una baja cantidad de intervalos (12), una fiabilidad del 75 por ciento probablemente sería suficiente. Pero estos datos indican una discrepancia sustancial y deben ser considerados como poco fiables. Los observadores deberían intentar clarificar las definiciones del comportamiento productivo e improductivo, utilizando ejemplos de cada uno para generar un mayor acuerdo en la categoría de desempeño que ellos están observando.

La fiabilidad de la muestra del tiempo grupal (GTS) está determinada por el cálculo de cuánto acuerdo registraron los observadores independientes por cada muestra del tiempo grupal. Suponga que usted quiere verificar el grado en el cual sus estudiantes están participando en el aprendizaje activo. Después de definir lo que usted considera como estar en aprendizaje activo, podría probar oportunamente esta categoría usando el GTS. Suponga que usted hace una muestra del tiempo grupal cada 3 minutos durante un período de 30 minutos. Esto le suministraría 10 muestras por período y le daría una buena idea del grado en el que sus estudiantes estuvieron involucrados en el aprendizaje activo. Permítanos asumir que usted tiene 24 estudiantes en su clase y que su docente colaborador está verificando la fiabilidad. Los datos originales de los registros de la muestra del tiempo grupal aparecen como en la Tabla 16.3.

TABLA 16.3 Datos de la Muestra del Tiempo Grupal (GTS)

	Sus observaciones	Docente colaborador
GTS 1	12/24	14/24
GTS 2	18/24	19/24
GTS 3	17/24	17/24
GTS 4	14/24	14/24
GTS 5	10/24	12/24
GTS 6	12/24	10/24
GTS 7	14/24	14/24
GTS 8	20/24	21/24
GTS 9	22/24	22/24
GTS 10	20/24	20/24

La fiabilidad es calculada muy fácilmente con el conteo de los desacuerdos por cada GTS. Como usted registró 12 estudiantes activos en el primer GTS y el docente colaborador registró 14, hubo 2 desacuerdos; ya que hay 24 estudiantes, hubo 22 acuerdos. En los 10 GTS mostrados, hay un total de 8 desacuerdos que al sustraerlos de un total posible de 240 (10 x 24 integrantes de la clase) muestra 232 acuerdos. La fiabilidad se calcula como es usual:

$$\frac{232}{232 + 8} * 100 = 97\%$$

Esto demuestra una alta fiabilidad, lo cual le daría a usted confianza de que sus observaciones del grado en el que sus estudiantes están participando en el aprendizaje activo son precisas. Incidentalmente, esos datos hipotéticos muestran que, durante la porción media de la sesión de clase, el porcentaje de estudiantes participando en el aprendizaje activo fue del 50 por ciento apenas, un hecho que podría estimularle a que usted examine qué en la

organización de la clase causó que la mitad de los estudiantes estuviesen inactivos en la parte sustancial del tiempo instruccional.

Ejemplos se sistemas de observación

Lo que sigue son varios ejemplos de instrumentos de observación sistemática. La codificación de estos instrumentos será *evento*, *duración*, *intervalo*, *muestra del tiempo*, o alguna combinación de los mismos. Los protocolos de observación han sido seleccionados para mostrar una variedad de esas técnicas de registro, así como para mostrar sistemas que se centran en unas pocas categorías de comportamiento básicos, así como en sistemas más completos.

1. *Análisis del tiempo del estudiante*. La información con relación a cómo emplean el tiempo los estudiantes puede ser útil para formarse un juicio acerca de la efectividad general de una clase. La Figura 16.2 muestra un sencillo análisis del tiempo de tres categorías, en el cual los bloques de tiempo utilizados son listados en las columnas apropiadas de *habla el docente*, *gerencia y aprendizaje activo*. Datos como estos están se expresan mejor como un porcentaje del tiempo total de la clase, permitiendo la comparación con las clases de diferente duración. Usted notará en este ejemplo que la totalidad de habla el docente y gerencia suman más que el tiempo empleado en el aprendizaje activo.

Registro de la Porción del Tiempo de la Clase		
Clase: <i>9no grado / Señor Allen</i>	Fecha: <i>1/9</i>	Tiempo: <i>2:00 – 2:40</i>
Habla el docente (demostración e instrucción)	Gerencial	Aprendizaje Activo
<i>3:06</i>	<i>1:17</i>	<i>6:18</i>
<i>1:08</i>	<i>1:24</i>	<i>5:20</i>
<i>4:30</i>	<i>0:46</i>	<i>3:50</i>
<i>2:06</i>	<i>0:40</i>	<i>4:25</i>
<i>1:10</i>	<i>2:50</i>	<i>19:53</i>
<i>12:00</i>	<i>1:50</i>	
	<i>1:00</i>	
	<i>0:20</i>	
	<i>9:01</i>	

FIGURA 16.2 Análisis del tiempo del estudiante en columna

2. *Análisis del tiempo del estudiante utilizando una línea de tiempo*. El ejemplo de la Figura 16.3 emplea cinco categorías, permitiendo así un análisis más detallado del tiempo de lo que fue posible en el ejemplo anterior. También, el empleo de la línea de tiempo preserva un registro visual de la secuencia de varios bloques del tiempo a

través de la lección. Los docentes pueden mirar una línea de tiempo completa y ver inmediatamente donde ocurrió el bloque gerencial y dónde, durante la clase, los estudiantes estaban en tiempo de espera. Pero los datos de la línea de tiempo son ligeramente más difíciles de contar y resumir que en el enfoque de columna simple mostrada previamente.

Docente		Escuela		Actividad		Fecha	
Grado	Hora inicial		Finalización		Nº en clase	Nº participantes	
Análisis del tiempo							
Espera (E)			Período sin actividad y sin movimientos entre actividades				
Transición (T)			Períodos de cambio de una actividad a otra (incluyendo alinearse o tranquilizarse para la próxima actividad)				
Gerencial (G)			Tiempo relacionado con la gerencia de la clase y no relacionado con la actividad educativa.				
Actividad (A)			Estudiantes que participan en prácticas de destrezas, juegos de práctica, juegos recreativos, condición física u otras prácticas relacionadas con los objetivos de la lección.				
Recepción de información (I)			Los estudiantes atienden las instrucciones o demostraciones del docente, u otra información relacionada con la clase				

	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16	17	18
	19	20	21	22	23	24
	25	26	27	28	29	30
	31	32	33	34	35	36
	37	38	39	40	41	42
	43	44	45	46	47	48

FIGURA 16.3 Análisis del tiempo del estudiante utilizando una línea de tiempo

3. *Análisis de la reacción del docente empleando el registro de eventos.* Los docentes responden a los estudiantes a menudo durante la clase y de muchas maneras diferentes. Estos eventos ocurren con tanta frecuencia y a menudo tan rápidamente que los docentes rara vez tienen una idea clara del patrón general de sus reacciones. Por lo tanto, es útil registrar estos eventos para suministrarle feedback a los docentes. Hemos encontrado que es beneficioso hacer distinción entre las reacciones a los comportamientos sociales/gerenciales y las reacciones a lo sustantivo, al comportamiento relacionado con la materia como los intentos en una destreza o en juegos deportivos. Ejemplos de estos pueden encontrarse en la Caja 5.3 y en la Caja 13.4. El sistema de

registro de eventos puede ser desarrollado para observar cualquier combinación de esos comportamientos. El ejemplo de la Figura 16.4 utiliza cuatro categorías importantes, dos para el feedback a las destrezas y dos para el feedback al comportamiento. Estas observaciones sistemáticas de las reacciones del docente podrían ser más categorizadas para mostrar la orientación del feedback (niño / niña, estudiantes con destrezas altas / media / baja, y así sucesivamente).

CAJA 5.3 Motivación del Comportamiento Apropriado a través de Interacciones Positivas			
Interacción general positiva (sin información específica del contenido):			
Si	Bueno	Buen trabajo	Excelente
Bello	Estupendo	Camino a seguir	Buena situación
Gracias	Esa es la manera	Bien hecho	A todos les fue bien
Genial	Extraordinario	Lo estás haciendo mejor	Fantástico
Interacción positiva no verbal:			
Sonrisa	Si, con la cabeza	Palmadas	Pulgar arriba
Signo ok	Chocar las cinco	Guiño	Aplausos
Interacciones positivas específicas (puede ser combinada con acciones no verbales):			
¡Equipo 2, hizo un gran trabajo al organizarse rápidamente esta vez!			
¡Gracias por prestar atención, Jack!			
¿Vieron ustedes la manera como William ayudó a Deborah?			
¡Toda la clase trabajó duro en ese ejercicio!			
¡Este grupo estuvo tranquilo justo a la señal!			
¡Gran trabajo de preparación, solo les tomó 12 segundos!			
¡Te agradezco tu cooperación con ese equipo, Mary!			
Interacciones positivas con valor de contenido (el valor de contenido describe porqué el comportamiento es importante):			
Esa es la manera de tranquilizarse. ¡Ahora podemos comenzar el juego más rápido!			
Gracias Bill. ¡Cuando haces un esfuerzo, los otros chicos también lo parecen!			
Ann, hiciste un gran trabajo esta semana, ¡y estoy segura de que te divertiste más!			
El equipo 3 hizo un gran trabajo con el equipo. ¡Ahora podemos comenzar antes!			
¡Bien hecho! ¡Cuando trabajas tanto en los ejercicios, juegas mejor!			

BOX 13.4 Feedback sobre la Materia: Tipos y Ejemplos		
<i>Feedback positivo general:</i> Su propósito es apoyar el esfuerzo del estudiante y construir un clima de aprendizaje positivo		
Buen tiro.	Buen esfuerzo.	Esa es la idea correcta, Mary.
Defensa fuerte, Jim.	A-1, Roberto.	Tremendo pase, Bob.
Eso está mejor, Jill	Muy bien.	El equipo 1 lo hizo realmente bien.
<i>Feedback positivo no verbal:</i> El mismo propósito anterior. Puede acompañarse de expresiones verbales.		
Palmadas	Pulgares arriba	Palmada en la espalda
El signo "OK"	Revolver el cabello	Puño elevado y apretado
<i>Feedback positivo específico:</i> El propósito es proporcionar información específica acerca de lo que fue hecho certeramente.		
Buena transmisión, Bill, era exactamente la velocidad y altura correctas. Eso estuvo bien. Tu círculo era diferente al de cualquiera. ¡Hermosa! Realmente tenías las rodillas flexionadas esta vez. Todos en este círculo hicieron el cambio al paso-salto exactamente al compás. Mucho mejor. Tu brazo delantero estaba recto durante todo el balanceo atrás. Bien hecho, equipo 2. La sincronización con los cortes fue buena.		
<i>Feedback correctivo:</i> El propósito es corregir los errores con información específica.		
Denise, debes mantener tu posición por más tiempo antes de moverte. Empieza por las piernas, Joe. Solo estás disparando con los brazos. Ok, pero lo hiciste la última vez. Intenta encontrar una forma diferente ahora. El equipo 3 necesita cubrir mejor las alas de la zona. ¡Anticipar! Jane, tenías un callejón abierto y deberías haber intentado un lanzamiento y pase.		
<i>Feedback específico con valor de contenido:</i> El propósito no solo es proporcionar información, sino también conectar el desempeño con el resultado.		
Es la manera de ayudar en defensa, Pat. Cuando cubres así, podemos arriesgarnos.		
Eso es mejor. Cuando tienes la cabeza levantada, puedes ver a tus compañeros de equipo.		
Buen esfuerzo, Jane. Cuando trabajas duro así, mejorarás muy rápidamente.		
Eso es Wanda. Cuando le brindas ese tipo de ayuda, Jesse aprenderá más.		

Docente: <i>Longlin</i>	Fecha: <i>3/9</i>	Escuela: <i>Desert H.E.</i>	
Actividad: <i>Pista</i>	Hora de inicio: <i>9:05</i>	Hora final: <i>9:40</i>	
Duración de la observación: <i>35</i>	Observador: <i>Cusimano</i>		
Definiciones:			
1. <i>Proporcionando información exacta sobre el desempeño (motor)</i> 2. <i>Palabras que apoyan la respuesta motora de los estudiantes</i> 3. <i>Proporcionar reconocimientos recomendables sobre el comportamiento, que no sea motor</i> 4. <i>Comentario del docente para terminar un comportamiento</i>			
<i>Feedback positivo específico a las destrezas</i>	<i>Feedback positivo general a las destrezas</i>	<i>Elogio al comportamiento</i>	<i>Desiste</i>
III III III I	III III III III III III III III III III III III	III II	III III III III III
Totales: <i>16</i>	<i>58</i>	<i>7</i>	<i>23</i>
Resumen de los datos:			
Comportamientos	Frecuencia total	Promedio por minuto	
1. <i>Feedback positivo a las destrezas (E)</i>	<i>16</i>	<i>0,45</i>	
2. <i>Feedback positivo a las destrezas (G)</i>	<i>58</i>	<i>1,65</i>	
3. <i>Elogios</i>	<i>7</i>	<i>0,20</i>	
4. <i>Desiste</i>	<i>23</i>	<i>0,65</i>	
Comentarios:			
★ <i>Pareces ser más específico hacia los estudiantes masculinos</i>			
★ <i>¡Permítanos trabajar en el comportamiento del elogio! (Crucial en este momento del año)</i>			
★ <i>¡Sé firme cuando desistas!</i>			

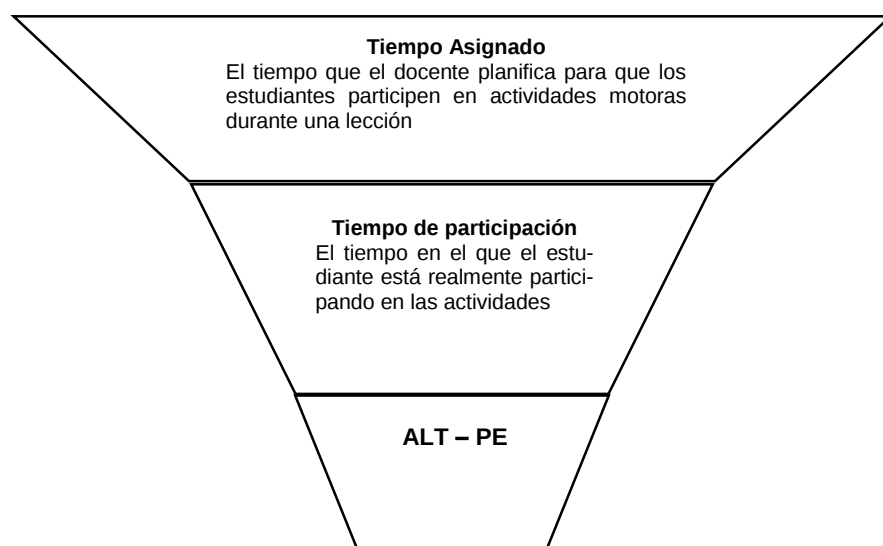
FIGURA 16.4 Análisis de la reacción del docente utilizando el registro de eventos (Fuente: "Basic Recording Tactics", de H. van der Mars, en *Analyzing Physical Education and Sport Instruction*, 2^{da} ed., pgs. 19 – 51, por P.W. Darst, D.B. Zakrajsek, V.H. Mancini, eds. 1989, Champaign, IL: Human Kinetics. © 1989 Paul W. Darst, Dorothy B. Zakrajsek, Victor Mancini. Reimpreso con permiso).

4. *Análisis del estudiante o de la clase utilizando el muestreo del tiempo grupal.*
 Cuando se desarrolle un formato de observación del comportamiento del estudiante, usted debe decidir si observa un estudiante, varios estudiantes, o la clase completa. La decisión debe ser tomada sobre la base de qué tipo de información servirá mejor al objetivo de la observación. En ocasiones, es útil obtener información relacionada con una clase llena de estudiantes. El ejemplo de la Figura 16.5 utiliza el muestreo del tiempo grupal para valorar el grado en que los estudiantes están comportándose apropiadamente (en la tarea, por ejemplo, independientemente de si el enfoque es gerencial o instruccional), participando en una actividad temática o en una actividad temática con una tasa alta de éxito, como se requiere para ALT-PE. Cuando usted revise la Caja 2.1 usted verá que las tres medidas de la participación utilizadas en este sistema proporcionarán información muy importante para el docente. La estrategia de observación es el muestreo del tiempo grupal, donde la clase es escaneada periódicamente y registrada la cantidad de estudiantes participando de acuerdo con las tres definiciones de las categorías. En la

Figura 16.5 el muestreo del tiempo se hace cada 4 minutos. Los datos de este ejemplo muestran una clase que se comporta bien (98 por ciento apropiado), participando en las actividades temáticas con una tasa alta (69 por ciento) pero no siempre con éxito (38 por ciento).

CAJA 2.1 El Efecto Túnel en el Tiempo de Actividad del Estudiante

A menudo, los docentes planifican una cierta cantidad de tiempo para que los estudiantes participen en las actividades de la clase. Sin duda, ellos también esperan que la mayor parte del tiempo de participación funcionará para que aprendan. Pero el tiempo planificado por el docente no siempre se convierte en un tiempo de aprendizaje funcional para el estudiante. Esto puede verse en los tres tipos de tiempo descritos abajo y cómo el tiempo se pierde en lo que Metzler (1980) llamó el “efecto embudo”.



Lo que es importante para el aprendizaje del estudiante es la cantidad de tiempo al final del túnel. Cuando el tiempo de participación es mucho menor que el tiempo asignado, el maestro generalmente tiene un problema para gerenciar y organizar a los estudiantes. Cuando ALT-PE es mucho menor que el tiempo de participación, el maestro tiene problemas para diseñar actividades en las que los estudiantes puedan experimentar el éxito.

Fuente: Adaptado de Metzler, 1980.

Análisis del Comportamiento del Estudiante											
Clase: <i>Período 5^o / Voleibol</i>				Docente: <i>Brown</i>				Nº en clase: <i>30</i>			
Hora de inicio: <i>1:30</i>				Hora de finalización: <i>2:10</i>				Duración de la observación: <i>40 min</i>			
Comportamiento del estudiante	Apropiado	<i>20</i>	<i>30</i>	<i>28</i>	<i>26</i>	<i>20</i>	<i>30</i>	<i>30</i>	<i>30</i>	<i>30</i>	<i>20</i>
	Participando	<i>4</i>	<i>26</i>	<i>16</i>	<i>18</i>	<i>20</i>	<i>30</i>	<i>28</i>	<i>26</i>	<i>24</i>	<i>16</i>
	ALT-PE	<i>0</i>	<i>0</i>	<i>14</i>	<i>18</i>	<i>0</i>	<i>24</i>	<i>26</i>	<i>22</i>	<i>0</i>	<i>10</i>
	Apropiada: <i>78%</i>			Participando: <i>69%</i>				ALT-PE: <i>38%</i>			

Figura 16.5 Formato de observación para el análisis del comportamiento del estudiante

5. *Análisis de episodios gerenciales.* Los episodios gerenciales son importantes unidades de enseñanza (ver Capítulo 1). El tiempo total empleado en la gerencia es crucial, como lo es la dirección efectiva de cada episodio gerencial. El sistema de observación ilustrado en la Figura 16.6 fue desarrollado para enfocarse exclusivamente en los episodios gerenciales; por ello, la estructura del formato de observación es por episodios. La Figura 16.6 utiliza el registro de eventos por la cantidad de interacciones positivas y negativas del docente dentro de cada episodio, el registro de duración por la duración de cada episodio, el registro de eventos por el número de comportamientos gerenciales en cada episodio (tal como indicaciones, orientación) y el GTS por una medida del comportamiento apropiado de la clase durante el episodio. Este ejemplo es de la medición de referencia, por lo que es anterior a un intento de cambio, lo cual es bueno porque el docente claramente es demasiado negativo y los episodios son demasiado largos.

Comportamientos en las Interacciones													
Período de Registro de Eventos de 3 o 5 minutos		+	-	+	-	+	-	+	-	+	-	+	-
		II	III I	I	III	III	III III	I	III	II	III		
Episodios Gerenciales	Duración	2:47		0:58		3:16		1:42		1:36		2:30	
	Cantidad de comportamientos gerenciales por episodio	III		III		III II		III		III		III	
GTS para el comportamiento apropiado por episodio		22 / 28		25 / 28		24 / 28		21 / 28		26 / 28		22 / 28	
Clase: Jones 8º grado Fecha: 11 / 1 Hora de inicio: 9:30 Hora de final: 10:10													
Resumen de datos													
Tasa de + reacciones por minuto= 0,27				Tiempo total en gerencia= 12:49									
Tasa de + reacciones por minuto= 1,67				Tiempo promedio por episodio= 769 / 6= 128 segundos= 2:08									
Tasa de +/- reacciones= 4 / 15 = 1 / 4				Porcentaje de comportamiento apropiado= 184 / 224 = 82%									
Comentarios													
REFERENCIA - 2ª OBSERVACIÓN													
LA GERENCIA DE INICIO Y FINAL DEMASIADO LARGO													
EL ENFOQUE CASI TOTALMENTE EN LA INTERACCIÓN NEGATIVA													
PARECE QUE LOS EPISODIOS GERENCIALES LARGOS CONTRIBUYEN A COMPORTAMIENTOS INAPROPIADOS													

FIGURA 16.6 Análisis de episodios gerenciales

6. *Análisis de destrezas gerenciales.* La Figura 16.7 muestra un sistema de observaciones que puede ser usado para valorar y mejorar las destrezas gerenciales. Tiene algunas de las características mostradas en la Figura 16.6, como la extensión de cada episodio gerencial y un muestreo del tiempo grupal del comportamiento apropiado del estudiante por episodio. También incluye un análisis mucho más detallado de las interacciones conductuales en cuatro principales categorías (general, específica, de valor general, de valor específico), con el registro adicional más categorizado de acuerdo a la orientación del feedback (a individuos, pequeños grupos, o al grupo total), e incluyendo la codificación separada de las interacciones no verbales (facial, gestual, o de contacto). Hay espacios específicos para que el observador califique el

modelado, el uso de la extinción, la orientación adecuada de las reacciones y su momento, así como un espacio para comentar el uso de la voz por parte del docente y la variedad de las interacciones. El sistema requiere el uso del registro de eventos (hecho en este caso con 5 minutos de intervalo por 20 minutos), registro de la duración y muestra del tiempo del grupo.

Interacciones conductuales																	
Duración del tiempo en el Registro de Eventos	Cuatro intervalos de 5 minutos= 20 minutos	Positivo(+)						Iniciado por + comportamiento	Negativo (-)						Comentarios		
		Orientación			No verbal				Orientación			No verbal					
		I	S	G	F	G	C		I	S	G	F	G	C			
	General	III		III	II	I	III	III I	I		III	III	I		Mirando hacia abajo		
	Específico	III III I					III	III No hay suficiente aquí	III		III I						
Valor general	I	I	III														
Valor específico	III	III	III I				II	II		III			II				
Voz y variedad		Buen trabajo IIII-11 demasiada voz, buena en grupo dirigido															
	Duración	0:32			0:47			0:16	1:02			0:18					
Episodios gerenciales	Cantidad de comportamientos gerenciales	II			III			II	III			II					
GTS de comportamiento apropiado		22/24		24/24		21/24		22/24	20/24		23/24		24/24				
Uso de modelaje III	Mucho mejor, el uso de John para modelar siguiendo las instrucciones bien hecho, la mayoría de los comentarios grupales tienen información y contenido de valor	Valoración por focalización						Más positivo específico para organización rápida, más positivo para Billy y Walter									
Uso de extinción III	Aún tiene problemas para ignorar a Nancy	Valoración del tiempo						Mucho mejor, observa la reacción cuando mira hacia abajo, desiste bien en clase durante el cambio de ejercicio									
Clase: Chapman 4º		Fecha: 3/16, 3ª observación después de la referencia						Hora de inicio: 1:10				Hora de finalización: 1:52					
Resumen de datos																	
+ interacción/minuto= 2.4								% de interacciones no verbales= 15/74 = 20%									
- interacción/minuto= 1.3								Contacto no verbal/sin contacto= 8/15 = 50%									
Relación +/- = 48/26= 2/1								% de interacciones específicas= 43/74 = 58%									
% de interacciones iniciadas por + = 23%								% de interacciones de valor= 26/74 = 35%									
Total de tiempo gerencial= 2,55																	
Promedio de tiempo por episodio= 0.35																	
Comportamiento gerencial por episodio= 2.4																	
% de comportamiento apropiado= 156/168= 93%																	

FIGURA 16.7 Análisis de destrezas gerenciales

7. *Análisis de ALT-PE.* El ALT-PE es uno de los criterios de las variables del proceso más utilizadas para juzgar la efectividad de la enseñanza en educación física. El formato de observación más común es el sistema de registro de intervalos que se muestra en la Figura 16.8. El formato de codificación es dividido en intervalos, con cada caja de intervalo conformada un nivel superior y uno inferior. El nivel superior se utiliza para describir el contexto del intervalo, y las diez opciones son de contenido general,

contenido de conocimiento de la materia y contenido motor de la materia. Esta decisión se toma sobre la base de lo que está haciendo la clase como totalidad; por ejemplo, ¿están participando en el calentamiento, en una lectura sobre estrategia, o en practicar las destrezas? El nivel bajo de la caja de intervalo se utiliza para describir la participación de un estudiante, con la selección de las categorías descritas como “en compromiso motor” o “sin compromiso motor”. La letra del código para la categoría correcta se coloca en la parte indicada de la caja de intervalo. Este sistema aporta una visión total de lo que la clase hace a través de la lección y una imagen finamente capturada de la participación de varios estudiantes. Por lo general, la sugerencia es observar tres estudiantes de diferentes niveles de destreza y alternar la observación entre ellos en cada intervalo. Aquellas cajas de intervalo marcadas como apropiados para motor (MA) son los intervalos ALT-PE, y su total revela el total de ALT-PE para ese estudiante durante la clase.

Tiempo de Aprendizaje Académico en Educación Física (ALT-PE)																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
S__	C																										
	LI																										
S__	C																										
	LI																										
S__	C																										
	LI																										
S__	C																										
	LI																										
S__	C																										
	LI																										
S__	C																										
	LI																										

Contenido general	Nivel de contexto (C)		Nivel de participación del aprendiz (LI)	
	Contenido de conocimiento de materia	Contenido de conocimiento motor	Sin compromiso motor	En compromiso motor
Transición (T)	Técnica (TN)	Práctica de destreza (P)	Parcialmente (PA)	Movimiento apropiado (MA)
Gerencial (G)	Estrategia (E)	Simulación de juegos / Rutina (S)	Espera (E)	Movimiento inapropiado (MI)
Alto (A)	Reglas (R)	Competencia (CO)	Fuera de la tarea (FT)	Apoyo (AP)
Calentar (C)	Comportamiento social (CS)	Condición Física (CF)	En la tarea (ET)	
	Antecedentes (AN)		Cognitivo (CG)	

FIGURA 16.8 Análisis de ALT-PE utilizando el registro de intervalos (Fuente: Siedentop, Tousignant, y Parker, 1982)

8. *Análisis de OTR*⁶. La oportunidad para responder es otro criterio de la variable del proceso que está relacionada con la enseñanza efectiva. En el análisis OTR las respuestas del estudiante se evalúan a medida que ocurren usando un formato de registro de eventos. En el ejemplo de la Figura 16.9 la forma topográfica de la respuesta es evaluada como aceptable (A) o inaceptable (IA), y el éxito de la respuesta dentro del contexto es evaluado como exitoso (E) o no exitoso (NE). Este ejemplo muestra los códigos de la destreza en fútbol y voleibol, para que se pueda efectuar un registro que sea específico de acuerdo con la destreza que intenta el

⁶ Opportunity to respond, Oportunidad para responder

estudiante. El ejemplo también muestra el momento de la clase en que se produjo la respuesta. Obviamente, este tipo de observación no puede ser hecha a una clase completa. El procedimiento típico es observar dos o tres estudiantes de diferentes niveles de destreza. La hoja de codificación de la Figura 16.9 muestra la observación de dos estudiantes en una clase de voleibol.

Fútbol de campo						Voleibol				Inicio de conteo		0268
Trampa-T		Lanzamiento-L		Servicio-S		"SPIKE"-I		Fin de conteo				1396
Golpe-G		Arquero-A		Pase-P		Bloqueo-BQ		Se detiene reloj				16:41
Drible-D		No responde-N		"BUMP"-B		Clavada-C		Inicia el reloj				1:25
		No codificable-X		"DIG"-D		No responde-N		Tiempo total				15:15
Voleibol / Soccer – Clase # 1 de 20						Fecha de la clase: 4 / 26 Modificación 4						
Observador: Brown						Fecha de la observación: 2/17						
Sujeto: James						Sujeto: Sally						

Secuencia de la res- puesta	Código de des- treza	Topografía		Resultados			Secuencia de la res- puesta	Código de des- treza	Topografía		Resultados		
		A	IA	E	NE				A	IA	E	NE	
1	1:29	P	1		1		1	1:35	P		1	1	
2	2:13	P		1	1		2	1:46	B		1		1
3	2:25	N					3	1:54	N				
4	2:33	N					4	2:08	N				
5	4:20	P		1	1		5	2:12	P		1	1	
6	6:59	P		1		1	6	2:14	P		1		1
7	8:24	P		1		1	7	2:25	N				
8	8:42	P		1		1	8	3:47	S		1	1	
9	9:23	P		1		1	9	4:02	S		1	1	
10	9:45	P	1		1		10	4:06	P		1	1	
11	12:12	P		1		1	11	5:37	P		1		1
12	14:14	S	1			1	12	5:52	P		1	1	
13	15:01	P		1		1	13	6:43	S		1		1
14							14	8:21	S		1	1	
15							15	9:20	S		1	1	
16							16	9:45	P		1	1	
17							17	10:42	S	1			1
18							18	11:12	P		1		1
19							19	11:58	S		1		1
20							20	14:11	N				
21							21	15:20	P		1	1	
Totales			3	8	4	7	Totales			1	16	10	7

FIGURA 16.9 Análisis de OTR usando el registro de eventos (Fuente: “Systematic Observation of Student Opportunities to Respond (SOSOR)” por Will Brown, en Analyzing Physical Education and Sport Instruction. 2nd ed., pp. 189-193, P.W. Darst, D.B. Zakrajsek, V.H. Mancini, eds., 1989, Champaign, IL: Human Kinetics. © por Paul W. Darst, Dorothy B. Zakrajsek, Victor H. Mancini. Reprinted by permission).

9. *Un instrumento de supervisión general.* Durante las experiencias de campo y de formación del estudiante, a menudo es útil para un programa la adopción de un sistema de supervisión general que se utilizará con todos los pasantes. El sistema sirve para varias funciones: proporciona el énfasis necesario para lograr los principales objetivos del programa, un medio de evaluación del nivel del progreso del programa hacia esas metas, un medio de evaluación individual de los pasantes, y un medio de comparación del desempeño relativo de los diferentes pasantes. El sistema de observación

mostrado en la Figura 16.10 utiliza el registro de eventos para las reacciones docentes (feedback a las destrezas y al comportamiento en las interacciones), una línea de tiempo para el registro de la duración de los aspectos más importantes sobre cómo se usa el tiempo en la clase (instrucción, gerencia y actividad), y muestreo del tiempo del grupo cada 3 minutos para valorar dos importantes características del comportamiento en clase (el número de estudiantes comportándose apropiadamente y el número de estudiantes en ALT-PE). El resumen apropiado de las estadísticas se presenta en la parte inferior de la hoja de codificación. Este sistema de codificación de la supervisión puede ser utilizado por un observador – el supervisor universitario, el docente colaborador de la escuela, o un compañero pasante. Los datos desarrollados son importantes indicadores de la efectividad, y pueden ser codificados fiablemente y resumidos rápidamente en reunión con el pasante que fue observado.

10. *Un instrumento de observación basado en las tareas de la lección.* Una alternativa al instrumento basado en el tiempo mostrado en la Figura 16.10 es un instrumento basado en la progresión de las tareas durante una lección. Ya que los docentes a menudo planifican por tareas, este protocolo de observación es más cercano a cómo las lecciones están realmente planificadas y ejecutadas. La Figura 16.11 muestra el formato de codificación el cual está organizado por tareas (gerencial, transición, instrucción, práctica). Se registra el momento inicial de cada tarea por lo que el tiempo total en esa tarea es fácilmente determinado en referencia al momento inicial de la próxima tarea. Se registra el tipo de tarea (información, refinamiento, extensión, aplicación). En cada tarea se selecciona un estudiante diferente para ser observado, rotándolos por sexo y nivel de destreza. Luego se toma una decisión considerando si el estudiante estaba realizando la tarea, o la modificó, o la abandonó. Cada respuesta del estudiante se codifica por su adecuación y éxito. El instrumento proporciona un claro registro de las tareas en cualquier lección, cuánto tiempo hizo cada uno, y qué hicieron los estudiantes durante las tareas.

Registro de asignación de tiempo del estudiante en clase

Observador	Fecha	Escuela	Grado	Ambiente
Hora de inicio	Hora final	Total de minutos observados		Experimental
Nº d estudiantes en clase	Actividad	Pasante		Generalización

(Códigos para el Análisis del Tiempo: 1= Instrucción; A= Actividad; G= Gerencia)

1	2	3	4	5	6	7	8	9	10
		MTG1			MTG2			MTG3	
11	12	13	14	15	16	17	18	19	20
	MTG4			MTG5			MTG6		
21	22	23	24	25	26	27	28	29	30
MTG7			MTG8			MTG9			MTG10
31	32	33	34	35	36	37	38	39	40
		MTG11			MTG12			MTG13	
41	42	43	44	45	46	47	48	49	50
	MTG14			MTG15			MTG16		

Análisis del tiempo		Informe del feedback a las destrezas					
		Positivo		Correctivo		Negativo	
		General	Específico	General	Específico	General	Específico
Tiempo total gerencial							
% del tiempo gerencial							
Tiempo total instruccional							
% del tiempo instruccional							
Tiempo total de actividad							
% del tiempo de actividad							
Total							
Promedio P/M							

Comportamiento en interacciones			
Positivo		Negativo	
General	Específico	General	Específico
Total			
Promedio P/M			

Análisis de la muestra del tiempo grupal			
% de comportamiento apropiado			
% ALT-PE			

FIGURA 16.10 Instrumento de supervisión general

Fecha: 2/6	Escuela: McVey
Lección: 2 de 18 N° en clase:	Página: 1 de 4
Hora: 10:02	
Tarea del episodio: M	<i>Entrada en la rutina bien realizada!</i>
Tipo de tarea:	
Estudiante observado: MH	
Participación en la tarea: ST	
Hora: 10:03	
Tarea del episodio: T	<i>Se realiza la rutina rápido!</i>
Tipo de tarea:	
Estudiante observado: FH	
Participación en la tarea: ST	
Hora: 10:04	
Tarea del episodio: I	
Tipo de tarea: I	<i>Explica y demuestra el pase con antebrazos en forma clara y precisa!</i>
Estudiante observado: ML	
Participación en la tarea: ST	
Hora: 10:09	<i>En pareja, pase con antebrazo</i>
Tarea del episodio: P	IU AS AS
Tipo de tarea: I	IU AS
Estudiante observado: FL	AU AS
Participación en la tarea: ST	AU AS
Episodio: (M) Gerencial, (T) Transición, (I) Instrucción, (P) Práctica	
Tipo: (I) Informar, (R) Refinar, (A) Aplicar	
Estudiante observado: (M) Masculino, (F) Femenino, (H) Altamente adiestrado, (L) Poco adiestrado	
Participación: (ST) En la tarea, (M+) Modificada con mejoría, (M-) Modificada con desmejora, (OT) Fuera de la tarea	
Respuestas: (A) Apropia, (I) Inapropia, (S) Exitosa, (U) No exitosa	

FIGURA 16.11 Instrumento basado en la progresión de la tarea

RESUMEN

1. Para que la enseñanza mejore deben definirse objetivos específicos, proporcionar prácticas apropiadas y ofrecer un feedback apropiado y relevante.
2. Los datos son confiables cuando dos observadores independientes, utilizando las mismas definiciones y observando a mismo docente al mismo tiempo, producen registros similares.
3. Los métodos tradicionales de observación de la enseñanza han sido el juicio intuitivo, “eyeballing”, registro anecdótico, lista de verificación o escala de calificación. Cada uno de ellos tiene problemas de sesgo del observador y tendencia a la falta de fiabilidad.
4. Las técnicas de observación sistemática incluyen el registro de eventos, el registro de duración, el registro de intervalos y el muestreo del tiempo del grupo.
5. Los docentes pueden auto-registrar su propio comportamiento en la enseñanza y los comportamientos relacionados con los aprendizajes más importantes de sus estudiantes.
6. Las técnicas de observación pueden ser combinadas para producir un sistema de observación multidimensional que se concentre en ambos, en el comportamiento del docente y el de los estudiantes.
7. Las decisiones en el desarrollo de un sistema de observación incluyen cuáles objetivos se buscan en los episodios de enseñanza, cuáles comportamientos del docente y cuáles de los estudiantes son indicadores válidos de esos objetivos, cuáles técnicas de

observación son las más adecuadas para medir esos indicadores, y cuántos pueden ser observados fiablemente.

8. Los sistemas de observación se construyen combinando la técnica de observación con la variable observada, e integrando las diversas técnicas de observación en un sistema viable.
9. Los observadores necesitan tener todos los materiales necesarios para completar las observaciones, así como ayuda tecnológica para señalar sus observaciones. Ellos deben ser tan discretos como sea posible.
10. Los observadores necesitan ser entrenados cuidadosa y sistemáticamente, yendo de las definiciones de las categorías a la video práctica, a la práctica en vivo en el campo y, finalmente, estableciendo la fiabilidad con un observador entrenado.
11. La fiabilidad de varias técnicas de observación se calcula comparando los registros de observación de los observadores independientes y calculando un puntaje porcentual del acuerdo.

Referencias

- Baumgartner, T., Jackson, A., Mahar, M., & Rowe, D. (2007). *Measurement for evaluation in physical education & exercise science*. Human Kinetics. USA.
- Blair, RC., y Taylor, R. (2008). *Bioestadística*. Pearson Educación. México.
- Burton, A. & Miller, D. (1998). *Movement skill assessment*. Human Kinetics. USA.
- Departamento de Salud y Asistencia Pública de los Estados Unidos. (2006). *Manual de Evaluación de la Actividad Física*. Atlanta, Georgia.
- Fernández, Miguel. (1994). *Las tareas de la profesión de enseñar*. Siglo XXI Editores S.A. Madrid
- George, J., Fisher, A., y Vehrs, P. (1996). *Tests y pruebas físicas*. Paidotribo. España.
- Hernández, R., Fernández, C., y Baptista, P. (1991). *Metodología de la investigación*. Mc Graw Hill. México.
- Heyward, V., & Wagner, D. (2004). *Applied body composition assessment*. Human Kinetics. USA.
- Lind, D., Marchall, W., y Wathen, S. (2008) *Estadística aplicada a los negocios y economía*. McGraw Hill Interamericana. México.
- Martínez, Emilio. (2002). *Pruebas de aptitud física*. Paidotribo. España.
- Maureira, Fernando. (2017). *Estadística Básica para Educación Física*. Bubok Publishing SL. España.
- Mendenhall, W., Beaver, R., Beaver, B. (2010). *Introducción a la probabilidad y estadística*. Cengage learning. México.
- Morrow, J., Jackson, A., Disch, J., & Mood, D. (2000). *Measurement and evaluation in human performance*. McGraw Hill. USA.
- Norton, K., y Olds, T. (2000). *Antropométrica*. Biosystem Servicio Educativo. Argentina.
- Reiman, M., & Manske, R. (2009). *Functional testing in human performance*. Human Kinetics. USA.
- Tanner, R., & Gore, C. Editors. (2013) *Physiological Tests for Elite Athletes*. Human Kinetics. Australian Institute of Sport.

Thomas, J., Nelson, J., & Silverman, S. (2005). *Research methods in physical activity*. Human Kinetics. USA.

Triola, Mario. (2009). *Estadística*. Pearson Educación. México.

Vincent, William. (2005). *Statistics in kinesiology*. Human kinetics. USA.

Welk, Gregory (Editor). (2002). *Physical activity assessments for health – related research*. Human Kinetics. USA.

Presentación de la Colección: Premio del Libro Universitario EDILUZ 60 Aniversario. 2021

Como director de la Editorial de la Universidad del Zulia (EDILUZ) tomé la decisión de convocar el Concurso del Libro Universitario EDILUZ 60 Aniversario en el año 2021. Este concurso se abrió para la participación de todos los profesores de la Universidad del Zulia, con el fin de promover y difundir el conocimiento humanístico y científico de nuestra Alma Mater, porque existía una deuda de convertir en libros digitales tantos años de investigación a los que sean dedicado la mayoría de nuestros docentes.

Alrededor de este premio debemos resaltar dos sucesos históricos importantes. En primer lugar, el reconcomiendo a la trayectoria de una editorial universitaria que alcanza sus 60 años de existencia cumpliendo denodadamente sus compromisos y labores; luego, está la aparición de la pandemia de la COVID-19, que nos llevó a cambiar nuestros hábitos cotidianos, y que a pesar de ello no detuvo el funcionamiento de nuestra editorial en pleno confinamiento y restricciones de todo tipo, creándose propuestas alternativas como fue la convocatoria de este concurso del libro universitario.

En este Concurso de Libros Universitarios EDILUZ 60 Aniversario, se evidenció el compromiso de los profesores y el alto nivel académico de la universidad, resaltando la excelencia investigativa, a pesar de la tremenda crisis a la que ha sido sometida la universidad pública en Venezuela durante más de dos décadas.

De los 57 libros participantes en el Concurso de Libros Universitarios EDILUZ 60 Aniversario, los jurados prestigiosos seleccionaron 16 obras para declararlas ganadoras, y cuyo premio consiste en la publicación digital de las mismas, y que hoy es una realidad cumplida. Con el éxito de esta iniciativa editorial se demuestra que EDILUZ está a tono con los tiempos y su forma de gerenciar responde a las exigencias de innovación y transformación impuestas a nivel mundial, con el fin de derribar barreras y ampliar los horizontes que ofrece la globalización de la digitalidad y las nuevas tecnologías en beneficio del desarrollo universitario.

Reciban todos los participantes y ganadores nuestras felicitaciones. Sigamos construyendo el futuro de nuestra Universidad del Zulia.

Poeta Carlos Ildemar Pérez
Director de la Editorial



UNIVERSIDAD
DEL ZULIA

