9

Comparación de puntuaciones

Las puntuaciones de las personas en los tests se obtienen, generalmente, para ser comparadas.

Pueden establecerse diversos tipos de comparaciones:

1. Comparaciones entre las puntuaciones de grupos.

2. Comparaciones entre la puntuación de una persona y las de un grupo.

3. Comparaciones entre puntuaciones de personas.

4. Comparaciones entre la puntuación de una persona y un nivel prefijado.

Cada uno de estos tipos de comparaciones tiene un valor distinto.

Las puntuaciones adquieren su significado por comparación. De uno u otro modo, es la comparación de una puntuación con otra u otras lo que pretende dotarla de significado psicológico. Generalmente, una puntuación cobra sentido cuando se conoce su comparación con otra u otras.

1. Comparaciones entre las puntuaciones de grupos

Las comparaciones entre las puntuaciones de grupos, colectivos o muestras pueden realizarse con diversos *propósitos*:

1. Para contribuir a determinar la validez del test al contrastar si colectivos que por hipótesis debían diferir en el test efectivamente difieren (enfoque de validez).

2. Para estudiar aspectos diferenciales del funcionamiento del test en diversos colectivos humanos distintos en alguna o

algunas variables conocidas (enfoque diferencial).

1.1. Enfoque de validez

El enfoque de validez de la comparación entre grupos pretende poner a prueba el test y contrastar si efectivamente discrimina entre grupos que *se sabe a priori* que son distintos en la variable que mide el test.

Ejemplo. Supongamos que hemos diseñado un test cuyo propósito es distinguir entre "buenos conductores" y "conductores peligrosos", con la intención de, una vez demostrado su poder predictivo, proponer su incorporación a las pruebas iniciales o periódicas a que se han de someter los conductores.

Se selecciona una muestra suficiente de buenos conductores. (Definimos a los "buenos" conductores de acuerdo a criterios claros ajenos al test. Por ejemplo podríamos considerar "buen conductor" a aquel que conduce por encima de determinado número de horas semanales, no ha tenido ninguna infracción, accidente ni incidente en un periodo determinado de años y ha superado con puntuaciones muy altas una prueba de conocimiento del código de la circulación y otra de buena conducción a juicio de dos examinadores). Paralelamente, se selecciona una muestra suficiente de "malos conductores" (También habría que definir con exactitud y con criterios claros ajenos al test que es un "mal conductor". Por ejemplo podría ser una muestra por debajo de determinados niveles de las mismas condiciones que hemos usado para definir un "buen conductor").

Se aplica el nuevo test a ambos grupos. Si el test es valido ambos grupos habrán de mostrar diferencias entre sí en sus puntuaciones en el test. Si los dos grupos no difieren en sus puntuaciones difícilmente podrá decirse que las puntuaciones en el test están relacionadas con lo "buen" o "mal" conductor que se sea.

También puede aplicarse el enfoque de validez para contrastar si un test es sensible al cambio de los sujetos. Supongamos que *sabemos con certeza,* por medios independientes del test, que determinado colectivo ha variado en la variable X después de cierto tiempo. Si ello es así un test que mida X y que sea sensible al cambio debería manifestar esas diferencias entre las puntuaciones en el tiempo 1 y en el tiempo 2.

1.2. Enfoque Diferencial

En otras ocasiones estamos interesados en comparar grupos mediante un test ya contrastado, con el propósito de conocer en qué grado difieren en la variable que mide el test o para poner a prueba hipótesis relacionadas con esa variable aunque no con la validez del test. Denominamos a este caso enfoque diferencial de la comparación entre grupos.

Ejemplo 1. Uso de un test para describir diferencialmente dos colectivos. Un claro ejemplo de este uso se produce cuando formulamos cuestiones del tipo ¿difieren hombres y mujeres de cierta población en su fluidez verbal? o ¿difieren hombres y mujeres en su actitud ante la pena de muerte? El método para resolver estas preguntas consiste en obtener una muestra adecuada de hombres y mujeres, administrarles el test y estudiar las diferencias entre sus puntuaciones.

Ejemplo 2. Uso de un test para poner a prueba una hipótesis Supongamos que determinada teoría postula que los directivos tienden a ser autoritarios en el sector secundario y participativos en el sector servicios. Si disponemos de un test adecuado cuya capacidad para medir el grado de autoritarismo de los directivos está bien contrastada, entonces si disponemos de una muestra adecuada de directivos de ambos sectores podemos administrarles el test y comprobar si la hipótesis es o no cierta comparando sus puntuaciones.

Un contraste de hipótesis particularmente relevante es el que se refiere a comprobar si las puntuaciones de un mismo colectivo de sujetos han cambiado significativamente después de determinado periodo de tiempo en el que se ha introducido algún tipo de tratamiento o cambio ambiental o bien se espera cualquier tipo de cambios evolutivos.

1.3. Método de Trabajo y Técnicas estadísticas a emplear

En los dos enfoques el método de trabajo es el mismo:

1. Obtener una muestra adecuada que represente adecuadamente cada uno de los colectivos que se desea comparar.

2. Administrarles el test.

3. Comparar los colectivos en sus puntuaciones en el test.

¿Cómo se efectúa la comparación entre las puntuaciones de las muestras de cada colectivo? (Podemos denominar muestra de cada colectivo a la fracción de la muestra general que comprende cada colectivo).

A efectos estadísticos hay que distinguir dos casos según el número de muestras (colectivos distintos) a comparar:

1. Cuando se desea comparar en el test a sólo dos colectivos corresponde aplicar una prueba t.

Si las muestras son independientes entre si (p.e. hombres y mujeres; ocupados y desocupados, enfermos y sanos, etc.) procede aplicar una ***prueba t clásica para dos muestras independientes*** cuya fórmula es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \cdot \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

aplicada con $N_1 + N_2 - 2$ grados de libertad. En esta fórmula sólo intervienen los tamaños de los grupos, las medias de los grupos y las cuasivarianzas de los grupos.

La prueba t anterior hace el supuesto de que la varianza de ambas muestras es igual en la población. Pero en ocasiones no se puede sostener la hipótesis de que la varianza poblacional de ambos grupos es la misma. Esto sucede por ejemplo si se ha aplicado previamente un *test de Levene de igualdad de varianzas* entre ambas muestras y este ha resultado significativo. En ese caso, *cuando no puede sostenerse el supuesto de igualdad de varianzas poblacionales con muestras independientes,* debe aplicarse la **t de Welch**, cuya fórmula es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

En la fórmula anterior sólo intervienen las medias muestrales, las cuasivarianzas muestrales y el tamaño de ambas muestras. La t empírica resultante distribuye como un t de Student con grados de libertad:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Las muestras a comparar no siempre son independientes entre sí. Decimos que tenemos dos *muestras dependientes* entre sí, cuando:

- comparamos dos muestras formadas por sujetos relacionados realmente en pares (cada varón con su pareja, cada madre con su hijo mayor, cada trabajador con su supervisor, una muestra de sujetos patológicos cada uno de ellos con un hermano no patológico...)

- comparamos las puntuaciones de una misma muestra de sujetos en el tiempo 1 con sus puntuaciones en el tiempo 2, (por ejemplo, antes y después de un determinado tratamiento o periodo evolutivo)

- comparamos las puntuaciones de una misma muestra de sujetos en dos tests distintos, o en dos partes de un mismo tests (por ejemplo, queremos saber si hay diferencias significativas entre las puntuaciones de una muestra de 200 escolares en fluidez verbal y las puntuaciones de esa misma muestra en razonamiento verbal).

En estos casos procede utilizar para el contraste de hipótesis la ***prueba t para muestras dependientes***, con la siguiente fórmula de cálculo:

$$t = \frac{\sum(X_1 - X_2)}{\sqrt{\frac{N\sum(X_1 - X_2)^2 - (\sum(X_1 - X_2))^2}{N-1}}}$$

aplicada con $N-1$ grados de libertad. (Referido N aquí al número de pares).

2. Cuando se desea comparar en el test a tres o más colectivos (p.e., tres confesiones religiosas; cuatro grupos expuestos a cuatro tratamientos distintos; cuatro muestras de sujetos con patologías distintas). En este caso los grupos se comparan mediante *análisis de varianza.*

Si los grupos se forman debido a un solo factor (p.e. cuatro grupos formados únicamente en función del partido político al que se ha votado en las últimas elecciones, considerando solo los 3 partidos más votados) se utiliza *análisis de varianza simple*.

Si los grupos se forman como resultado del cruce de dos o más factores se utiliza un *diseño factorial* de análisis de la varianza. (p.e., si se forman ocho grupos debido a la consideración simultánea del sexo (2 niveles) y el partido político al que se ha votado en las últimas elecciones (con 4 niveles) decimos que se trata de un diseño factorial "2 por 4".

Ejemplos resueltos de pruebas t

Prueba t para muestras independientes asumiendo igualdad de varianzas (contraste bidireccional)

Deseamos comprobar si existen diferencias significativas ($\alpha = 0'05$) en las puntuaciones en cierto test de dos grupos formados por 8 personas cada uno, después de

aplicar en cada uno de ellos dos técnicas distintas A y B. Los datos son los siguientes:

Grupo A: 12, 14, 12, 10, 9, 12, 15, 10

Grupo B: 14, 14, 12, 16, 17, 11, 14, 12

Hipótesis:

$H_0: \mu_A = \mu_B$

$H_1: \mu_A \neq \mu_B$

Estadístico de contraste:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{gl = n_1 + n_2 - 2}$$

Valor empírico:

$$t_e = \frac{11'75 - 13'75}{\sqrt{\frac{4'2143\ (8-1) + 4'2143\ (8-1)}{8+8-2}\left(\frac{1}{8} + \frac{1}{8}\right)}} = -1'9485$$

$\mu_1 - \mu_2$ desaparece de la fórmula porque la hipótesis nula supone que son iguales y por tanto su diferencia es nula.

Con gl=14, el nivel crítico bidireccional de esta t empírica es 0'0717.

Como esta probabilidad es mayor que $\alpha = 0'05$ concluimos que no se puede rechazar la hipótesis nula. Sin embargo, un nivel crítico como 0'07 está lo bastante cerca del nivel de significación 0'05 como para sugerir que deberíamos acumular mayor evidencia antes de tomar una decisión en términos sustantivos. Un resultado así puede sugerir que conviene repetir el experimento y quizás (si la naturaleza del problema lo permite y este ejemplo podría no ser el caso) aumentar el n en el nuevo experimento.

Prueba t para muestras independientes sin asumir la igualdad de varianzas (contraste unidireccional)

Supongamos que contrastamos la eficacia de un nuevo método de tratamiento B frente al método de tratamiento tradicional A para obtener una relajación profunda medida por el test X. Disponemos de datos de un grupo clínico de 10 personas a las que se ha aplicado el método tradicional y de otras 7 a las que se ha aplicado el nuevo método. Contrastar si el nuevo método es más eficaz que el antiguo.

Grupo A: 20, 21, 19, 20, 21, 23, 20, 20, 19, 19.

Grupo B: 20, 26, 24, 25, 22, 18, 17.

Hipótesis: $H_0: \mu_A \geq \mu_B;\ H_1: \mu_A < \mu_B$

Estadístico de contraste:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Valor empírico:

$$t_e = \frac{20'20 - 21'7143}{\sqrt{\frac{1'5111}{10} + \frac{12'2381}{7}}} = -1'0988$$

Grados de libertad:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{1'5111}{10} + \frac{12'2381}{7}\right)^2}{\frac{\left(\frac{1'5111}{10}\right)^2}{10 - 1} + \frac{\left(\frac{12'2381}{7}\right)^2}{7 - 1}} \approx 7$$

El nivel crítico de la t empírica unidireccional es 0'1541.

Como el nivel crítico es mayor que $\alpha = 0'05$ concluimos que no se puede rechazar la hipótesis nula. A pesar de que los resultados del método B pueden caracterizarse por actuar en el sentido de la hipótesis y también por producir una dispersión mucho mayor, la diferencia entre medias no es estadísticamente significativa.

Prueba t para muestras dependientes (contraste unidireccional)

Tenemos 5 clientes participando del tratamiento durante cierto periodo. La variable clave es el test X, y se espera que el tratamiento aumente la puntuación en este test. El contraste es obviamente unidireccional porque sólo se aceptará el nuevo método si aumenta la variable de interés y se desea plantear con un nivel $\alpha = 0'05$ Estos son los resultados obtenidos:

Caso:	Antes:	Después:	Diferencia d:
1	11	23	-12
2	17	34	-17
3	23	25	-2
4	11	19	-8
5	24	25	-1

Hipótesis:

$H_0: \mu_a \geq \mu_d$

$H_1: \mu_a < \mu_d$

Estadístico de contraste:

$$t = \frac{\sum d}{\sqrt{\frac{N\sum d^2 - (\sum d)^2}{n-1}}}$$

Valor empírico:

$$t_e = \frac{-40}{\sqrt{\frac{5(502) - 1600}{5-1}}} = -2'652$$

Para 4 gl y un contraste unidireccional el nivel crítico de este valor del estadístico es 0'0284. Cómo es menor que el nivel de significación al que se ha planteado el contraste se concluye rechazando la hipótesis nula. El nuevo método aumenta significativamente los valores de la variable de interés. Además de una observación de los valores puede advertirse que es más eficaz en aquellos casos en que la puntuación de partida es menor.

2. Comparaciones entre la puntuación de una persona y la de un grupo

Existen dos aproximaciones generales de interpretación (y por tanto de comparación) de las puntuaciones de los sujetos en el test: interpretación orientada a normas e interpretación orientada a criterios.

> La *interpretación orientada a normas* supone que la puntuación del sujeto adquiere su significado en relación con las puntuaciones de otros sujetos más o menos semejantes que forman una muestra con la que comparar denominada grupo normativo. En esta orientación la puntuación directa se interpretará en función de qué posición ocupa en la distribución de puntuaciones del grupo normativo.
>
> La *interpretación orientada a criterios* comparará la puntuación del sujeto con criterios (niveles) de puntuación fijados por razones distintas a la distribución de la muestra. Se comparan los resultados del sujeto con algún estándar (o estándares) fijado de antemano e independiente de la muestra.

En este apartado nos ocuparemos de diversos métodos de comparación de la puntuación de un sujeto con las puntuaciones de un grupo, una aproximación que se ubica en la interpretación orientada a normas.

2.1. Comparaciones con el grupo normativo para interpretar la puntuación del sujeto

Las puntuaciones de los tests clásicos construidos bajo la teoría clásica generalmente no pueden ser interpretadas sin compararlas con un grupo de referencia que se denomina *grupo normativo.* Es decir, las puntuaciones de los tests clásicos están generalmente orientadas a una interpretación mediante normas de grupo.

> Supongamos que tenemos un sujeto que ha obtenido 75 puntos en un test de inteligencia con 100 preguntas. ¿Es mucha o poca inteligencia? Depende. Si el sujeto tiene 6 años y la media de los sujetos de 6 años es 100, con una desviación típica de 16, el sujeto parece sustancialmente "retrasado" en inteligencia. Pero si la media de los sujetos de esa edad es 50 con una desviación típica de 10 el sujeto esta muy "adelantado", ocupa una posición avanzada en comparación con su grupo.

El grupo normativo es una muestra adecuada de la población de referencia con la que tiene sentido comparar al sujeto. Delimitar con qué grupo o grupos de referencia tiene sentido comparar al sujeto puede depender de los propósitos concretos de la comparación de modo que puede haber varios grupos normativos con los que tenga sentido efectuar la comparación.

> Por ejemplo, si estamos midiendo razonamiento abstracto en sexto grado lo razonable es comparar al sujeto con una muestra adecuada de sujetos de sexto grado. Es obvio que no deberíamos comparar a nuestro sujeto con sujetos de 4º grado, pero dependerá de los propósitos de la comparación utilizar una muestra adecuada de sujetos de 6º grado de ambos sexos o de su mismo sexo, de su centro educativo o de todos los centros educativos de la ciudad, o de la nación o del estado.

Las tablas que ponen en relación la puntuación directa obtenida en el test (la puntuación total del sujeto en el test) y algún valor que determina su posición en el grupo normativo se conocen como *baremos o normas.*

Los tests, en su manual, han de incluir los baremos adecuados para que podamos comparar la puntuación de cada sujeto con el grupo o grupos normativos que resulte conveniente.

La comparación de la puntuación directa con el grupo normativo a través de un baremo tiene por objeto generalmente convertir la puntuación directa en una *puntuación normativa.* Denominaremos puntuación

normativa a aquella que indica la posición del sujeto en un grupo normativo.

2.2. Principales tipos de puntuaciones normativas

Hay tres tipos principales de puntuaciones normativas:

1. Los cuantiles.
2. Las puntuaciones típicas.
3. Las puntuaciones típicas normalizadas.

Los cuantiles, y, en especial los percentiles son el tipo de puntuación normativa más utilizado. Existen otros tipos de puntuaciones normativas, como las normas cronológicas, que merecen consideración a parte dado que implican parcialmente un principio generador distinto.

2.3 Cuantiles

Los cuantiles relacionan los valores de la variable con su posición en una distribución de frecuencias. Esencialmente indican que valor de la variable deja por debajo de sí determinada proporción de casos.

Supongamos que deseamos dividir una muestra de 1000 casos en dos grupos de igual tamaño en función de una variable X. Necesitaríamos saber que valor de la variable X deja por debajo (y por tanto por encima) al 50% de la muestra, es decir, que valor deja 500 personas por debajo en X. Un modo de obtener ese valor consiste en ordenar a las mil personas por su puntuación en X de menor a mayor y contar hasta 500 desde un extremo, e identificar que valor de X que queda a mitad de camino entre el caso 500 y el 501. Una tabla de frecuencias acumuladas puede ayudar a hacer ese conteo más sencillo. El concepto de "valor de la variable que deja por debajo de sí al 50% de la muestra" es el concepto de un estadístico de posición que se denomina mediana y se representa Md. Este concepto puede extenderse fácilmente a otras particiones de la muestra ordenada por una variable X en grupos de igual tamaño.

Los 3 puntos de corte en X que dividen a la muestra en 4 partes iguales se denominan *cuartiles*, y se representan por Q_1, Q_2, y Q_3,

Los nueve puntos de corte en X que dividen a la muestra en 10 grupos de igual tamaño se denominan *deciles* y se representan por D_1, D_2, ...D_9,

Los 99 puntos de corte en X que descomponen la muestra en 100 partes iguales se denominan *percentiles* y se representan P_1, P_2, ...P_{99}.

En general los estadísticos de posición se denominan *cuantiles*. Un cuantil es un valor de la variable que deja por debajo de si una porción determinada de los datos. Así los percentiles son un tipo de cuantiles; también los cuartiles y los deciles son cuantiles. Los cuantiles están relacionados entre sí:

Los deciles son percentiles múltiplos de 10. Es decir, los deciles D_1, D_2, ...D_9, corresponden exactamente con los percentiles P_{10}, P_{20}, ...P_{90}

Los cuartiles equivalen a percentiles múltiplos de 25. Así, los cuartiles Q_1, Q_2, y Q_3, que corresponden con los percentiles P_{25}, P_{50}, y P_{75}.

Entre Q_1 y Q_3 se encuentra siempre el 50% central de los casos. Y por tanto, fuera del intervalo Q_1 - Q_3 se encuentra siempre el 50% periférico de los casos La mediana Md tiene especial interés al representar aquel valor de la variable que deja por debajo (y por encima de sí) el 50% de los casos. Se corresponde con Q_2 con D_5 y con P_{50}

2.4. Los Percentiles: Los Cuantiles más utilizados

Se denomina *cuantil k'* a aquel valor de la variable (en este caso a aquella puntuación total en el test) que es igual o superior a los obtenidos por una determinada proporción k de los sujetos.

El *percentil k'* es aquel valor de la variable (puntuación en el total del test) que es igual o superior a los obtenidos por el k por cien de los casos.

Por ejemplo, si en un test la puntuación 39 es el percentil k' = 55 esto significa que el 55% de la muestra ha obtenido una puntuación inferior o igual a 39. Dicho de otro modo, que un sujeto que obtiene una puntuación de 39 iguala o supera en el test al 55% del grupo.

Generalmente se define el **percentil k** como aquel valor de la variable que deja *por debajo de sí* al k por cien de las observaciones. En mi opinión es más cómodo para el trabajo posterior definir el *percentil k'* como valor de la variable *que iguala o deja por debajo* al k por cien. Esto facilitará una obtención inmediata de las tablas de baremos y en general clarifica más el significado del cuantil. Para no inducir a confusión vamos a llamar a un percentil así definido como percentil k' (k prima), mientras que a los percentiles tradicionales los llamaremos percentil k o simplemente percentil.

Muchos tests expresan sus baremos en percentiles. Un *baremo* en percentiles consiste en una tabla donde se refleja que porcentaje de la muestra normativa deja por debajo de sí cada puntuación total del test.

El *método clásico para establecer los percentiles* consiste en:

1) Obtener una muestra adecuada (grupo normativo) y administrarles el test. (El test se supone que ya ha cumplido con los criterios de bondad clásicos de fiabilidad y validez).

2) Agrupar las puntuaciones en intervalos. Tabular los resultados: Es decir, construir una tabla con frecuencias, porcentajes, frecuencias acumuladas y porcentajes acumulados.

3) Aplicar la **fórmula de percentilación** para cada percentil k en que estemos interesados:

$$P_k = L_{inf} + \left(\frac{\frac{k \cdot N}{100} - n_{bajo}}{n_{dentro}} \right) \cdot i$$

En la fórmula:

P_k es el valor de la variable que deja por debajo de si el k por cien de los casos.

L_{inf} es el valor de la variable que representa el límite exacto inferior del intervalo donde cae el $\frac{k \cdot N}{100}$ por cien de los casos en la columna de porcentajes acumulados. Para referirnos a ese intervalo cómodamente lo llamaremos "intervalo crítico" siguiendo una convención al uso.

k es el porcentaje de interés. Por ejemplo si k=50 estamos calculando la mediana.

N es, como resulta usual, el número de casos total de la muestra.

n_{bajo} representa el número de casos acumulados hasta el intervalo inmediatamente inferior al intervalo crítico.

n_{dentro} es el número de casos dentro del intervalo crítico.

i es la amplitud del intervalo crítico, es decir, su límite exacto superior menos su límite exacto inferior.

Esta fórmula, de apariencia complicada, es conceptualmente muy simple. Se limita a sumar al valor de la variable límite inferior del intervalo donde se acumulan el k por cien de los casos, el fragmento de intervalo, interpolado mediante una regla de tres, suficiente para encontrar el punto de la variable que hipotéticamente deja por debajo de si el k por cien de los casos.

En el modo de hacer clásico, generalmente, se aplica la fórmula para cada k múltiplo de 10 entre 0 y 100, o bien, si se quiere detallar más, para cada k múltiplo de 5. Rara vez, aunque es prácticamente posible, se calcula la fórmula para

cada k entre 1 y 100. Otras combinaciones en la elección de k son posibles obviamente.

Este procedimiento *no* es, a mi juicio, recomendable, por varias razones.

Primero, no tiene sentido hacer intervalos para después calcular los percentiles. Hacer intervalos (es decir, policotomizar una variable y peor aun sería dicotomizarla) aun cuando el número y amplitud de los intervalos sean razonables y sean razonablemente aplicables a los datos a lo largo de todos los valores de la variable (condiciones que generalmente se incumplen al menos parcialmente) es un procedimiento que distorsiona la información original.

En general: Cuanto más difiera la distribución de una distribución plana (rectangular) perfecta más distorsión introducen los intervalos. Cuantos menos intervalos más distorsión. Cuanto mayor la amplitud de intervalo más distorsión. Cuanto menos homogénea la distribución de la puntuaciones en el interior de los intervalos más distorsión. Cuanto menor sea el rango de la variable mayor distorsión.

Los intervalos, utilizados con extraordinaria prudencia, pueden ser un recurso gráfico para levantar histogramas y "resumir" la información de la variable. Pero no hay justificación *para efectuar el cálculo de ningún estadístico* (incluidos y quizás especialmente los percentiles) sobre una distribución agrupada en intervalos *cuando se dispone de la distribución natural sin agrupar*. El mejor intervalo es la puntuación natural de la variable (es decir, no hacer intervalos). Agrupar en intervalos, para después aplicar la fórmula anterior de percentilación, interpolando qué sucede dentro del intervalo, bajo el supuesto de que las puntuaciones se distribuyen homogéneamente en él, cuando se conoce de verdad como se distribuyen las puntuaciones, me parece poco aceptable. En este método primero se distorsiona la información con los intervalos; después se busca la precisión en la interpolación sobre información distorsionada. Se sabe como distribuye realmente la variable, pero se 'supone' que distribuye de otro modo.

Ciertos usos de los intervalos son una herencia de los tiempos en que los cálculos se hacían a mano o con muy limitados recursos mecánicos y era mejor una aproximación que un cálculo aritmético inacabable. Desde la aparición de los ordenadores, el uso de los intervalos puede restringirse a su aplicación como método de suavización y presentación de síntesis estadísticas y gráficas. En general, debe evitarse utilizarlos *como fuente para estimar estadísticos* sobre ellos. Particularmente, por lo que se refiere a la elaboración de normas percentiles, en mi opinión, son una práctica menos que recomendable.

Segundo, generalmente no tiene sentido hacer interpolaciones más allá de la unidad de medida del test. Las puntuaciones totales de los tests en teoría clásica son

generalmente discretas: una suma de puntos, de ítems acertados.

Por eso en la mayoría de los tests o se han acertado 14 items o se han acertado 15, pero no existe nada semejante a haber acertado 14'757 items (por ejemplo). Por esta razón no tiene sentido (se hagan o no intervalos) en términos prácticos interpolar que el percentil 80, por ejemplo, es la puntuación 27'56. Podemos estar seguros de que ningún sujeto al que administremos un test puntuado como suma de aciertos producirá una puntuación de 27'56. O puntuará 27 ó 28.

Si, por ejemplo, hemos encontrado que 27 puntos es el percentil 76 y que 28 puntos es el percentil 84, no podemos discriminar entre posiciones percentiles entre el percentil 77 y el 83. En ese caso, simplemente con ese baremo obtenido empíricamente en esa muestra no podemos encontrar ningún sujeto en el percentil 80 porque nadie puede materialmente puntuar entre 27 y 28. Se puede coger la fórmula de percentilación e interpolar que el percentil 80 es algún valor decimal entre 27 y 28 pero este es un ejercicio esteril en términos prácticos.

Si esta situación de que el baremo no discrimine entre el percentil 76 y el 84 nos parece insatisfactoria podemos proceder a incrementar el número de items y volver a baremar (es decir, administrar la nueva prueba a una nueva muestra normativa). En la mayoría de tests el número de items representa o afecta al número de grados de diferenciación que podemos establecer en la variable, de ese modo para diferenciar con mayor fineza entre sujetos es necesario aumentar el número de items.

Si hemos obtenido, por ejemplo, que 27 es el percentil 50 y 29 el percentil 59, y carecemos de casos en la puntuación 28, nuestra muestra normativa no permite conocer con exactitud realmente que posición ocuparía la puntuación 28 entre esos dos percentiles. *De hecho,* no ha aparecido en el grupo normativo. En esta última situación es en la que está más justificado interpolar, puesto que sí tiene sentido que aparezca en una medición posterior un sujeto con puntuación 28 aunque este valor no se haya dado en el grupo normativo. Sin embargo, si este caso nos sucede varias veces con un baremo ello significa probablemente que el baremo no es adecuado, posiblemente porque la muestra tomada como grupo normativo no ha sido lo bastante grande para presentar casos con todas las puntuaciones del test relevantes para ser interpretadas. En ese caso lo mejor que se puede hacer es volver a administrar el test a un grupo normativo con un N mayor para obtener unos baremos más detallados y aconsejables.

Si un test tiene n items valorados dicotómicamente solo puede producir n+1 grados de discriminación entre sujetos. Si el test sólo tuviera un ítem, podría discriminar sólo dos clases de sujetos, los que lo aciertan con total X=1 y los que lo fallan, con total X=0. Si hay dos items, el test introduce 3 grados de discriminación: 2 ciertos, 1 acierto y 0 aciertos. Si

tenemos, por ejemplo, un test con solo 37 cuestiones no podemos aspirar a que cada uno de los 99 percentiles comprendidos entre 1 y 99 estén representados por una puntuación directa. Para tener 99 percentiles representados individualmente hacen falta al menos 99 grados de discriminación en el test, en el mejor de los casos. Desde el punto de vista de un test con 37 items valorados dicotómicamente, en el mejor de los casos ese test permitirá determinar 37 percentiles pero es imposible que los otros 62 percentiles tengan asignada una puntuación directa distinta. Si deseamos mucho detalle en el baremo de percentiles (es decir poder determinar el percentil 1, el 2, el 3...) la primera condición es utilizar un test con muchos items. La segunda, disponer de una muestra suficientemente amplia para que todos los grados posibles, se manifiesten separadamente, es decir que de hecho aparezcan todas las puntuaciones que el test permite.

El número n de items requerido es como mínimo igual (en la práctica normalmente mayor) que el número n_k de percentiles que deseamos determinar:

$$n \geq n_k$$

El tamaño de una muestra de baremación debe ser substancialmente grande. Las indicaciones para la elección del tamaño de la muestra deberían seguirse de las especififcaciones asociadas al muestreo, de forma que la representatividad de la muestra respecto de la población quede garantizada, y el tamaño sea garantía de que todos los grados de diferenciación posibles en el test pueden aparecer.

2.5. Método para establecer baremos de percentiles k'

Para establecer percentiles k' el método es mucho más sencillo que para establecer percentules convencionales y los problemas asociados a los percentiles convencionales no están presentes. En mi opinión este es el método que debe usarse.

1) Obtener una muestra adecuada (grupo normativo) y administrarles el test. El test se supone que ya ha cumplido con los criterios de bondad clásicos de fiabilidad y validez, y presenta suficiente número de items para expresar todos los grados de diferenciación en que estamos interesados. La muestra debe tener el tamaño suficiente y debe obtenerse por un método de muestreo adecuado.

2) Tabular los resultados: Es decir, construir una tabla con frecuencias, porcentajes, frecuencias acumuladas y porcentajes acumulados. No agrupar las puntuaciones en intervalos.

Pues bien, en esta tabla, *el porcentaje acumulado* para cada valor de la variable indica ya, directamente, el porcentaje de casos que obtienen una puntuación *igual o inferior* a ese valor de la variable.

Si por ejemplo al valor (total del test) 34 le corresponde un porcentaje acumulado de 79'81 ello significa que 34 es el percentil 79'81. Dicho de otro modo, cualquier sujeto que obtenga un total de 34 en el test diremos que está en el percentil k' 79'81, ó, dicho de otro modo, que iguala o deja por de si al 79'81% de los casos del grupo normativo.

En el siguiente apartado veremos algunos ejemplos de obtención de pecentiles por ambos métodos.

2.6. Ejemplos de obtención de percentiles

Los percentiles son estadísticos de posición, indican que valor de la variable X deja por debajo de sí determinada proporción o porcentaje de datos. Por ejemplo, se denomina percentil 35 al valor de la variable que deja *por debajo de sí* al 35 de los casos. Como todos los demás cuantiles tienen su equivalencia en percentiles, basta con saber calcular percentiles para poder calcular cualquier cuantil.

Supongamos la variable X que representa la puntuación de una variable psicológica obtenida mediante un test para N=250 casos. El test presenta 14 items que el sujeto puede acertar o fallar y la variable puede valer entre 0 y 14 puntos. La tabla de frecuencias siguiente resume los datos obtenidos en la muestra.

X	f	fa	p	pa	PA
0	9	9	0,036	0,036	3,6
1	8	17	0,032	0,068	6,8
2	9	26	0,036	0,104	10,4
3	9	35	0,036	0,14	14
4	7	42	0,028	0,168	16,8
5	11	53	0,044	0,212	21,2
6	20	73	0,08	0,292	29,2
7	35	108	0,14	0,432	43,2
8	50	158	0,2	0,632	63,2
9	23	181	0,092	0,724	72,4
10	21	202	0,084	0,808	80,8
11	12	214	0,048	0,856	85,6
12	10	224	0,04	0,896	89,6
13	14	238	0,056	0,952	95,2
14	12	250	0,048	1	100

En las cabeceras de columnas, X es el valor de la variable (puntos en el test), f es la frecuencia o número de casos que han obtenido ese valor X, fa es la frecuencia cumulada, p es la proporción, pa la proporción acumulada y PA el porcentaje acumulado.

Proceso razonado de obtención de un percentil por interpolación

Supongamos que deseamos obtener P_{10}, es decir, el percentil 10. El percentil 10 es un valor de X tal que divide la muestra en dos partes desiguales. Deja por debajo de sí al 10% y, consecuentemente, deja por encima de sí al 90% restante de los casos.

Observemos la tabla. Con el valor 0 (X=0) hay 9 casos, que representan el 3'6% de la muestra. Con el valor 1 o inferior hay 17 casos (fa=17) que representa un 6'8% de la muestra (PA=6'8). Con 2 o inferior encontramos al 10'4% de los casos. El valor 1 no llega a dejar el 10% por debajo, y el valor 2 ya deja más del 10% por debajo. En los datos de esta muestra no hay ningún valor de X que deje exactamente el 10% de los casos por debajo.

Por tanto, si deseamos decir que valor de X dejaría exactamente el 10% por debajo es necesario que hagamos una *interpretación de la variable X como continua* para poder admitir que 2 es la marca de clase de un continuo de valores que van entre 1'5 y 2'5, e interpolar el valor exacto entre 1'5 y 2'5 que dejaría exactamente el 10% por debajo. Al realizar una interpretación de X como continua, consideramos los valores de la columna X como marcas de clase de intervalos unitarios. Generalmente, la *marca de clase de un itervalo* es su valor central, y un *límite exacto* entre dos intervalos es aquel valor del continuo equidistante entre dos marcas de clase sucesivas. El límite exacto superior de cierto intervalo es a la vez el límite exacto inferior del intervalo siguiente que ocupa la posición adyacente superior en la escala.

Buscamos el valor de X que deja por debajo el 10% de los casos. El 10% de 250 casos son 25 casos; por tanto, 25 casos es la frecuencia acumulada que nos interesa localizar. Hasta el límite exacto superior del primer intervalo (=0'5) se han acumulado 9 casos, hasta el límite 1'5 se han acumulado 17, y hasta el límite 2'5 se han acumulado 26, es decir, uno de más. Este intervalo, dentro del cual está el porcentaje acumulado que buscamos se llama *intervalo crítico.*

Suponiendo que los nueve casos que hay dentro del intervalo crítico se distribuyeran homogéneamente dentro del mismo, se trata de interpolar que punto de X deja los 8 primeros por debajo. Los 8 primeros porque basta estos 8 más los 17 acumulados hasta el límite exacto superior del intervalo inmediatamente inferior para obtener los primeros 25 casos con puntuaciones más bajas en X.

Se trata de una interpolación lineal que se resuelve por una regla de tres. 8 casos son a 9 (total de casos dentro del intervalo critico) como z (fragmento del intervalo critico que ocupan los 8 primeros casos) es a 1 (que es la amplitud del intervalo crítico). De donde z=8/9 =0'89. Por tanto el

fragmento del intervalo crítico ocupado por los 8 primeros casos vale 0'89 en términos de puntos de las variable X.

Ahora para saber el valor de X que deja por debajo 25 casos basta con sumar ese fragmento z al límite exacto inferior del intervalo crítico. Es decir, 0'89 + 1'5 = 2'39. Por tanto concluimos que P_{10}, = 2'39. Es decir, 2'39 es el valor de la variable que deja por debajo al 10% de la muestra.

Obtención de un percentil aplicando la fórmula de interpolación:

El proceso que hemos razonado paso a paso es el mismo que efectúa resumidamente la fórmula de interpolación percentil, útil para el cálculo de cualquier percentil:

$$P_k = L_{inf} + \left(\frac{\frac{k \cdot N}{100} - n_{bajo}}{n_{dentro}} \right) \cdot i$$

En la fórmula:

P_k es el valor de la variable que deja por debajo de si el k por cien de los casos. Es decir, el percentil k.

L_{inf} es el valor de la variable que representa el límite exacto inferior del intervalo donde cae el $\frac{k \cdot N}{100}$ por cien de los casos en la columna de porcentajes acumulados. Para referirnos a ese intervalo cómodamente lo llamaremos como es usual "intervalo crítico".

k es el porcentaje de interés. Por ejemplo, si k=25 estamos calculando el primer cuartil.

N es el número de casos total de la muestra.

n_{bajo} representa el número de casos acumulados hasta el intervalo inmediatamente inferior al intervalo crítico.

n_{dentro} es el número de casos dentro del intervalo crítico.

i es la amplitud del intervalo crítico, es decir, su límite exacto superior menos su límite exacto inferior.

Aplicando la fórmula al cálculo del P_{10} tendríamos:

$$P_k = L_{inf} + \left(\frac{\frac{k \cdot N}{100} - n_{bajo}}{n_{dentro}} \right) \cdot i = 1'5 + \left(\frac{\frac{10 \cdot 250}{100} - 17}{9} \right) \cdot 1 = 2'39$$

Los percentiles juegan un papel muy importante en el trabajo profesional de los psicólogos y otros profesionales

como médicos, educadores, etc. Ello es así debido a que los percentiles permiten interpretar una medición de un caso en términos de su posición en una muestra de referencia. Por ejemplo, permiten al psicólogo decir que la puntuación de 34 puntos en un test de fluidez verbal significa que la persona deja por debajo de sí al 80% de las personas de su edad y características. O al pediatra decir que el peso de un bebé es muy adecuado para su edad pues deja por debajo de sí al 60% de los niños de su sexo y edad.

Debido a su importancia práctica es conveniente ser consciente de los supuestos que implica la definición y el cálculo clásico por interpolación de los percentiles que hemos expuesto en el punto anterior, y examinar otras propuestas alternativas.

La definición clásica de percentil k es aquel valor de la variable que *deja por debajo de sí* el k% de los casos. En los percentiles convencionales el investigador o el profesional eligen que percentiles desean calcular (generalmente los múltiplos de 5) y los interpolan, *suponiendo que la variable es continua* y *suponiendo que los casos se distribuyen homogéneamente* dentro de los intervalos.

El supuesto de que la variable es continua es exacto en variables como peso o estatura y fruto de una conjetura teórica en el caso de variables como fluidez verbal, razonamiento matemático o satisfacción laboral. Podemos admitir ese supuesto, más que verificarlo, si asumimos por cierta la teoría psicológica que lo sustenta (que no es, por cierto, la única alternativa posible al respecto en la concepción de esas variables). Sin embargo, en términos aplicados, los instrumentos de medida siempre arrojan puntuaciones discretas determinadas por su límite de discriminación entre magnitudes contiguas. Y esto, que es cierto de una báscula, lo es de modo mucho más patente cuando hablamos de las puntuaciones de un test.

El método tradicional se esfuerza en interpolar entre unidades el valor de los percentiles. Pero, siendo el instrumento de medida discreto en sus puntuaciones, las personas pueden puntuar en él o 2 ó 3 puntos, pero por más que midamos casos jamás encontraremos uno que puntúe 2'39 y por tanto, en el ejemplo, jamás encontraríamos un caso exactamente en el percentil 10 por lo que la supuesta precisión del método tradicional es, en estas circunstancias, un artificio poco razonable. Si nuestro test solo produce puntuaciones *discretas* de modo que no puede obtener nunca una puntuación entre 2 y 3, interpolar que valor x, tal que $2 < x < 3$, deja exactamente el 10% por debajo es un ejercicio de poca utilidad práctica. Por ello definir los cuantiles en valores imposibles para el instrumento puede ser útil descriptivamente (como cuando calculamos una mediana como referencia de la posición central) pero carece de sentido en el uso más común de los percentiles como procedimiento para elaborar baremos.

La interpolación que efectúan los percentiles necesita suponer como se distribuyen los casos dentro de los intervalos. Y hace el supuesto más sencillo. Supone que los casos se distribuyen homogéneamente a lo largo de los valores del intervalo, lo que permite una interpolación lineal mediante la sencilla regla de tres que contiene la fórmula. Esto sería razonable si la distribución fuese rectangular, uniforme, pero es obviamente un supuesto menos razonable en cualquier caso en que el polígono de frecuencias no discurra paralelo al eje de abscisas. Puede argumentarse que la diferencia entre el modelo de interpolación adecuado, supuesto a partir de la forma de la distribución, y la forma lineal, puede ser despreciable en la mayoría de los casos dado que solo afecta a los decimales de los valores calculados, siempre que el tamaño de los intervalos sea suficientemente pequeño. Así es, pero ¿si no se puede interpolar con precisión por qué ese empeño en interpolar como si tuviera sentido obtener esa precisión?

Cómo obtener percentiles tradicionales con enfoque de cálculo discreto.

Se pueden obtener los percentiles definidos tradicionalmente por un procedimiento de cálculo con enfoque discreto que evita las críticas sobre la continuidad y la interpolación. Este método consiste en ordenar todos los casos en función de su puntuación. Determinar que número de caso corresponde con el porcentaje k buscado y atribuir a ese porcentaje el valor discreto correspondiente a ese caso, independientemente del número de casos que hayan con la misma puntuación.

Por ejemplo, vamos a determinar por este método que vale el percentil 10 en los datos del ejemplo anterior. Dado que N=250, el 10% de los casos es igual a 25. El enfoque discreto simplemente busca cual es la puntuación concreta que ha obtenido el caso que hace el número 25 estando los casos ordenados por sus puntuaciones de menos a más.

La columna de frecuencias acumuladas (fa) nos ayuda a encontrarlo rápidamente. Para X=0 se han acumulado 9 casos (fa=9), por tanto todavía no hemos llegado al número 25. Para X=1 se han acumulado 17 casos (fa=17), por tanto todavía no hemos llegado al número 25. Para X=2 la frecuencia es 9 (f=9), es decir hay 9 casos que han puntuado 2. Para X=2 se han acumulado ya 26 casos. Por tanto, necesariamente, el caso número 25 es uno de estos nueve que han puntuado X=2 y por tanto el percentil 10 desde la definición tradicional pero con enfoque de cálculo discreto, es 2. (Recuérdese que el percentil 10 desde la definición tradicional y el modo de cálculo más tradicional mediante interpolación en intervalos es la puntuación 2'39).

Crítica de los percentiles tradicionales obtenidos con enfoque de cálculo discreto.

La solución discreta va ganando adeptos y esto me parece razonable dado que en la mayoría de las situaciones los supuestos del enfoque de interpolación se sabe que no se sostienen. Sin embargo, desde el punto de vista aplicado esta solución no es, a mi parecer suficiente.

El punto crítico es el siguiente. Por este enfoque tradicional pero con cálculo discreto, el investigador o el profesional determina que percentiles está interesado en calcular. generalmente los múltiplos de 10, o con más detalle los de 5. Es decir, el percentil 5, el 10, 15, 20, 25 etc. Y después obtiene la puntuación discreta que corresponde a cada uno de ellos. *Pero*, frecuentemente, debido a la acumulación de casos en determinados valores del test, *varios percentiles consecutivos pueden ser iguales entre sí al corresponderles la misma puntuación discreta.*

Si en los datos del ejemplo anterior el lector calcula, por ejemplo, los percentiles 45, 50, 55 y 60 tradicionalmente definidos pero por el método de cálculo discreto encontrará que todos ellos corresponden a la puntuación en la prueba 8. El problema es que el psicólogo práctico que utiliza baremos no se encuentra con percentiles a los que busca asignar una puntuación, sino con puntuaciones que trata de decir que percentil son. Cuando el profesional encuentre un 8 ¿qué debe decir? ¿Qué ese 8 es el que ocupaba en el intervalo la posición correspondiente al percentil 45 ó es el 8 correspondiente al caso número 125 que en una muestra de 250 corresponde al percentil 50, ó bien que corresponde a la posición 60? La distribución empírica, concreta, de puntuaciones del test que hemos obtenido no permite contestar a esta pregunta con exactitud. No sabemos si ese caso es el primero (digamos el peor) o el último (el mejor) del intervalo. Y no podemos saberlo con esta información.

Para tratar de evitar estas ambigüedades es necesario que la escala de la prueba tenga un rango bien amplio y que el tamaño de la muestra sea suficiente. Sin embargo, aunque el rango de la prueba (generalmente, en un test, determinado por el número de items) fuera de 0 a 100, es difícil que estas dificultades desaparezcan completamente, dado que la mayoría de las variables distribuyen de tal modo que los casos tienden a concentrarse en torno a sus valores centrales, estén estos donde estén ubicados en términos de la escala del test.

Dados estos resultados, ¿cómo interpretar la puntuación 8? De una cosa estamos seguros, un sujeto que ha sacado una puntuación de 8 iguala o deja por debajo de sí al 63'2% de la muestra (=PA). Está respuesta abre el camino al concepto de percentil k**'**.

*Una propuesta alternativa más sencilla: los percentiles k**'***

Para evitar estas dificultades y simplificar las cosas prefiero introducir un concepto, más sencillo, que llamaré percentil k**'** (percentil k prima).

A efectos prácticos lo que nos interesa en un baremo es saber qué porcentaje de la muestra ha obtenido una puntuación inferior *o igual* a cada una de las 20, 30 ó 100 puntuaciones concretas discretas que puede arrojar el test. Los percentiles k' responden a esta cuestión.

Denominaremos **percentil k'** a aquel valor de una variable que *iguala o deja por debajo de sí* el k% de los casos. Los percentiles k' se representan $P_{k'}$

El percentil k' ($P_{k'}$) es aquel valor de X ($P_{k'}$=X) que presenta un PA igual a k' (PA=k').

Una ventaja adicional de este enfoque es su sencillez de cálculo dado que la información que determina que percentil k' es cada valor de X se encuentra directamente en la tabla de frecuencias en la columna de porcentajes acumulados (PA). Esta columna indica precisamente que porcentaje de casos presenta una puntuación X *igual o menor* a una dada.

En la tabla del ejemplo anterior puede interpretarse que el 14% de los casos tienen una puntuación igual o inferior a 3 (PA=14 para X=3). El percentil k'=14 es 3.

O, por ejemplo, que el 80'8% puntúan 10 o menos, lo que implica, claro, que solo el 19'2% restante puntúa más de 10. El valor X=10 es el percentil k'=80'8.

En los datos no se ha determinado empíricamente el percentil k'=10, pero si el k'=10'4, de modo que de una persona que puntúa X=2 decimos que está en el percentil k'=10'4.

Para la puntuación X=8 la respuesta desde este enfoque es que 8 es el percentil k'=63'2. Es decir, que a una persona que ha puntuado 8 le atribuimos el igualar o dejar por debajo al 63'2% de la muestra.

La nueva definición de percentil k' como aquel valor de la variable que *iguala o deja por debajo de sí* el k% de los casos resulta práctica cuando se dispone de tablas de tabulación de las variables, evita efectuar interpolaciones, frecuentemente injustificadas, fuera de los valores que empíricamente están disponibles en una distribución, y ofrece una respuesta más honesta con la información disponible para ubicar a una persona que ha obtenido una puntuación determinada.

En los percentiles k' el investigador o el profesional no eligen los percentiles k' que desean calcular; por el contrario, obtienen para cada valor de X *que de hecho ha aparecido empíricamente* en la muestra su significado percentil en términos de que porcentaje de casos han obtenido ese valor u otro inferior. Este enfoque evita los supuestos de los percentiles convencionales por interpolación, que son difíciles de sostener y que no se compadecen con la naturaleza discreta de los instrumentos, ni con la forma usual de las distribuciones. También evita las ambigüedades de interpretación a que dan lugar los percentiles tradicionales obtenidos con enfoque discreto.

Por estas razones creo que los percentiles k' son un enfoque más realista y práctico para la elaboración de baremos.

Baremos y transformación a percentiles

Una tabla que relaciona los valores de la variable con los percentiles correspondientes se denomina un **baremo** de la variable. Aunque existen otras clases de baremos los de percentiles son los más utilizados en muchos contextos.

Un "baremo" permite transformar los valores de la variable en una nueva puntuación transformada que expresa la posición relativa de la puntuación en una muestra y facilita la interpretación de la misma.

Por ejemplo, un baremo de percentiles k' permite reexpresar los valores de la variable en una escala de 0 a 100 que expresa que proporción de la distribución deja por debajo o iguala ese valor de la variable, lo que facilita entender el significado de un valor de la variable en relación a los valores obtenidos por los casos de esa muestra.

Convertir los valores de la variable X en sus valores k' respectivos es un caso particular de transformación. Convertir cada puntuación en su valor k' asociado es una transformación no lineal de la variable. Los psicólogos y otros profesionales efectúan frecuentemente estas transformaciones al interpretar puntuaciones, elaborar informes, y explicar los resultados en pruebas y mediciones.

Por ejemplo, si de una persona determinada nos dicen que puntúa 5 en nuestro test ¿cómo interpretar esta cifra? ¿es mucho o es poco?. Evidentemente, depende. Depende de con qué se compare.

Una tabla de baremos percentiles implica el supuesto de que tiene sentido comparar un caso con la muestra a la que pertenece e interpretar su puntuación en relación a su posición en esa muestra. Así, si miramos en la tabla anterior veremos que 5 deja por debajo o iguala al 21'2% de los casos, lo que significa que esta persona está en el cuarto inferior de los casos.

Una tabla que presenta junto a los valores de la variable los porcentajes acumulados puede considerase una tabla de baremos de percentiles k' (bastaría pues la primera y última columna de la tabla del ejemplo anterior para constituir la tabla de baremos).

Un ejemplo de elaboración de baremos percentiles de las puntuaciones de un test

Hemos aplicado un test a una muestra de N=1087 personas. En el test puede obtenerse una puntuación entre 0 y 30 puntos. Para simplificar la presentación del problema, vamos a partir de los datos ya tabulados. Nuestro objetivo es obtener una tabla de baremos percentiles. Esta

tabla de baremos percentiles nos será muy útil para ubicar a cada persona en la muestra e "informar sus puntuaciones", es decir, generar informes sobre esas puntuaciones ofreciendo afirmaciones sobre su posición relativa en su grupo de comparación.

X	f	fa	p	pa	PA
0	3	3	0,0028	0,0028	0,276
1	7	10	0,0064	0,0092	0,92
2	16	26	0,0147	0,0239	2,3919
3	22	48	0,0202	0,0442	4,4158
4	24	72	0,0221	0,0662	6,6237
5	29	101	0,0267	0,0929	9,2916
6	34	135	0,0313	0,1242	12,42
7	36	171	0,0331	0,1573	15,731
8	37	208	0,034	0,1914	19,135
9	43	251	0,0396	0,2309	23,091
10	46	297	0,0423	0,2732	27,323
11	49	346	0,0451	0,3183	31,831
12	52	398	0,0478	0,3661	36,615
13	54	452	0,0497	0,4158	41,582
14	58	510	0,0534	0,4692	46,918
15	62	572	0,057	0,5262	52,622
16	67	639	0,0616	0,5879	58,786
17	54	693	0,0497	0,6375	63,753
18	51	744	0,0469	0,6845	68,445
19	43	787	0,0396	0,724	72,401
20	41	828	0,0377	0,7617	76,173
21	40	868	0,0368	0,7985	79,853
22	38	906	0,035	0,8335	83,349
23	36	942	0,0331	0,8666	86,661
24	32	974	0,0294	0,896	89,604
25	28	1002	0,0258	0,9218	92,18
26	25	1027	0,023	0,9448	94,48
27	21	1048	0,0193	0,9641	96,412
28	18	1066	0,0166	0,9807	98,068
29	16	1082	0,0147	0,9954	99,54
30	5	1087	0,0046	1	100
Sumas	1087		1		

Desde una definición tradicional de percentil vamos a determinar que valores dejan por debajo el 1, 5, 10, 15, ... 90, 95, y 99% de los casos la distribución de la variable. Hallaremos los percentiles en su concepción clásica primero desde el enfoque más tradicional de interpolación, y segundo desde el enfoque discreto. Por último, desde un enfoque de percentiles k' determinaremos que percentil k' es cada puntuación empírica de las 31 puntuaciones (de 0 a 30) que puede arrojar el test.

Concepción clásica de los percentiles, enfoque de cálculo por interpolación.

Vamos a ilustrar el procedimiento tradicional que supone que (1) la variable es continua con intervalos unitarios, y (2)

que las puntuaciones presentan una distribución rectangular dentro de cada intervalo, de modo que se busca encontrar que valor exacto de la variable deja por debajo de si el k % de los casos. Por ejemplo, para el percentil 1, qué valor exacto de la variable en el intervalo 1'5 a 2'5 deja por debajo de sí al 1% de los casos.

Necesariamente hay que partir de la tabla de frecuencias, incluyendo las columnas de acumulación imprescindibles para localizar el intervalo crítico en que interpolar cada percentil. Por simplicidad en la tabla siguiente están las frecuencias que reflejan los datos y el resto de resultados de tabulación.

Fórmula general:

$$P_k = L_{inf} + \left(\frac{\frac{k \cdot N}{100} - n_{bajo}}{n_{dentro}} \right) \cdot i$$

Para el percentil 1 tendremos:

$$P_1 = 1'5 + \left(\frac{\frac{1 \cdot 1087}{100} - 10}{16} \right) \cdot 1 = 1,5544$$

Para el percentil 5:

$$P_5 = 3'5 + \left(\frac{5 \cdot \frac{1087}{100} - 48}{24} \right) = 3,7646$$

Como N e i son constantes en la fórmula para unos mismos datos, simplificamos elementos constantes y seguimos calculando cada uno de los percentiles:

$$P_{10} = 5'5 + \frac{(10 \cdot 10,87) - 101}{34} = 5'7265$$

$$P_{15} = 6'5 + \frac{(15 \cdot 10,87) - 135}{36} = 7'2792$$

etc....

Por este procedimiento calculamos cada uno de los percentiles P_k solicitados. En la tabla siguiente se resume el valor de los términos en la fórmula de cálculo de cada uno de ellos, y, en la última columna el valor de X que es el percentil k correspondiente:

k	li	nb	nd	pk
1	1,5	10	16	1,5544
5	3,5	48	24	3,7646
10	5,5	101	34	5,7265
15	6,5	135	36	7,2792
20	8,5	208	43	8,7186
25	9,5	251	46	9,9511

30	10,5	297	49	11,094
35	11,5	346	52	12,163
40	12,5	398	54	13,181
45	13,5	452	58	14,141
50	14,5	510	62	15,04
55	15,5	572	67	15,886
60	16,5	639	54	16,744
65	17,5	693	51	17,766
70	18,5	744	43	18,893
75	19,5	787	41	20,189
80	21,5	868	38	21,542
85	22,5	906	36	22,999
90	24,5	974	28	24,654
95	26,5	1027	21	26,769
99	28,5	1066	16	29,133

La tabla de baremos en percentiles tradicionales por interpolación quedaría, pues, prescindiendo de los elementos de cálculo innecesarios en el baremo, del siguiente modo:

k	pk
1	1,5544
5	3,7646
10	5,7265
15	7,2792
20	8,7186
25	9,9511
30	11,094
35	12,163
40	13,181
45	14,141
50	15,04
55	15,886
60	16,744
65	17,766
70	18,893
75	20,189
80	21,542
85	22,999
90	24,654
95	26,769
99	29,133

En la tabla anterior leemos. por ejemplo, que una persona cuya puntuación en el test sea 26'769 está en el percentil 95, es decir, deja por debajo de sí al 95% de la muestra.

Concepción clásica de los percentiles, enfoque de cálculo discreto

En este método se supone que los casos están ordenados por puntuaciones de menor a mayor, y se atribuye al percentil k la puntuación concreta que ha obtenido el caso que ocupa la posición (fa) correspondiente.

k	fa	pk
1	10'87	2

5	54'35	4
10	108'7	6
15	163'05	7
20	217'4	9
25	271'75	10
30	326	11
35	380'45	12
40	438'8	13
45	489'15	14
50	543'5	15
55	597'85	16
60	652'2	17
65	706'55	18
70	760'9	19
75	815'25	20
80	869'6	22
85	923'95	23
90	978'3	25
95	1032'6	27
99	1076'1	29

La confección de la tabla anterior permite ilustrar claramente tres dificultades de esta aproximación:

1) Cuando el porcentaje de casos cae entre dos valores de X (como sucede para k=1, fa=10'87 entre X=1 con fa=10 y X=2 con fa=26) hay que hacer un supuesto sobre la ubicación de la parte decimal 0'87 y tomar una decisión de asignación. La solución adoptada en el ejemplo es coherente con el supuesto de la parte decimal correspondería al intervalo 2.

2) ¿Por qué decir que una persona que puntúa 14 deja por debajo de sí al 45%? Decimos esto porque hemos hecho la elección arbitraria de escoger interesarnos por el percentil 45. Pero con el mismo rigor y método podríamos decir que una persona con 14 puntos deja por debajo de sí el 42, el 43, el 44, el 45 y el 46%. Porque todos esos percentiles son la puntuación 14 con este método. (¿No será más exacto decir que una persona con puntuación de 14 iguala o deja por debajo de si al 46'91% de la muestra?)

3) ¿Qué pasa con las puntuaciones que no están en la tabla? Este es un problema práctico importante de los percentiles clásicos calculados por el método de interpolación o por el de aproximación discreta. ¿Qué pasa con las personas que al aplicarles el test puntúen 0, 1, 3, 5, 8, 21, 24, 26, 28 y 30? ¿Cómo interpretamos esas puntuaciones? ¿A qué percentil corresponden? Con estos baremos clásicos el profesional tendrá que hacer interpolaciones, en el mejor de los casos lineales, si no tiene acceso a la distribución completa (como suele pasar cuando se dispone de un test publicado y su manual).

Percentiles k'

Considerando que el test arroja puntuaciones discretas y utilizando percentiles k' el problema de obtener una tabla de baremos ya está resuelto al construir en la tabla de tabulación la columna PA de porcentajes acumulados. En efecto, una persona que obtenga en el test una puntuación de 3 puntos, por ejemplo, iguala o deja por debajo de sí exactamente al 4'42% de la muestra (y es superado por el 95'58% restante).

Utilizando percentiles k' el baremo sería el siguiente:

X	PA
0	0,28
1	0,92
2	2,39
3	4,42
4	6,62
5	9,29
6	12,42
7	15,73
8	19,14
9	23,09
10	27,32
11	31,83
12	36,61
13	41,58
14	46,92
15	52,62
16	58,79
17	63,75
18	68,45
19	72,40
20	76,17
21	79,85
22	83,35
23	86,66
24	89,60

25	92,18
26	94,48
27	96,41
28	98,07
29	99,54
30	100,00

En esta tabla puede leerse que una persona que puntúa en el test 26 deja por debajo o al menos iguala en el test al 94'48% de la muestra, mientras que una persona que puntúe 27 deja por debajo o al menos iguala al 96'41. La tabla no dice que porcentaje de casos deja por debajo de si una persona con 26'769 puntos en el test. (Pero tampoco hace falta si las respuestas ante el test son de naturaleza discreta y, dado que o bien se puntúa 26 ó bien se puntúa 27).

Con este método *todas las puntuaciones en el test tienen una y solo una interpretación percentil*, y esa interpretación no es arbitraria, responde a la mejor interpretación posible de la puntuación que permiten los datos. ("Mejor" en el sentido de puntuación más alta y suponiendo que esta tenga un sentido sustantivo positivo).

Los baremos en percentiles k' no hacen supuestos de continuidad ni de distribución homogénea, ofrecen información exacta sobre el porcentaje que cada valor X iguala o deja bajo sí, y aprovechan al máximo la información disponible de modo que no dejan puntuaciones X sin referirlas a un valor k'.

2.7. Las muestras para baremación

Las muestras para baremación deben presentar las cualidades propias de una buena muestra y algunas específicas por tratarse de una muestra de comparación que definirá la interpretación de las puntuaciones de los sujetos que midamos con el test.

Las muestras de baremación han de ser *representativas* del tipo de sujetos con el que se va a comparar para interpretar sus puntuaciones.

Han de tener un *tamaño suficiente*. Generalmente se admite que hacen falta aproximadamente unos 1.000 casos para disponer de un baremo razonable.

Los baremos han de estar actualizados, es decir, obtenidos a partir de *sujetos medidos en fechas recientes.* Los sujetos cambian, la cultura cambia y los baremos han de ser actuales. En muchos contextos no se pueden usar razonablemente baremos que tengan más de cinco o seis años de antigüedad.

Las mediciones de todos esos sujetos se tienen que haber hecho bien. En *buenas condiciones de administración* de los tests y con una administración escrupulosamente correcta, pero en el mismo tipo de ambiente en que después serán medidos los sujetos.

Las condiciones de obtención de las mediciones de la muestra han de ser lo más semejantes posibles a las de la medición profesional posterior.

Las muestras de baremación, denominadas *grupos normativos*, han de ser representativas y razonablemente comparables en variables demográficas (sexo, edad, zona geográfica...) y descriptivas generales (nivel educativo, nivel y tipo de actividad profesional, otras características específicas que sean relevantes en ese caso...) en función de los sujetos a los que luego se va a medir.

Por ejemplo, para sujetos en edad escolar lo razonable suele ser comparar con baremos obtenidos a partir de sujetos de su mismo grado educativo en centros de su entorno cultural próximo.

Es un disparate comparar la puntuación de un adolescente de Gandía, de Elche, de Valencia o de Vinaroz (por ejemplo) con una muestra de sujetos de Illinois o de Los Angeles (por ejemplo). Ni siquiera sería suficientemente adecuado compararlo con una muestra de centros de la periferia de Madrid o de la ciudad de Sevilla, por poner un ejemplo.

Debe tenerse en cuenta que no esta garantizado a priori que el comportamiento de dos muestras de baremación sea el mismo o semejante porque compartan *una* característica, por ejemplo la edad o el grado educativo. No está garantizado a priori que no pueda haber funcionamiento diferencial de los items y del test y no está garantizado a priori que el uso de la referencia de comparación inadecuada no pueda resultar perjudicial para la persona evaluada.

Cuanto más semejantes en condiciones culturales, educativas y personales sean los sujetos del baremo con aquellos otros que después mediremos con el test, más adecuado resultará el baremo y más “justa” la comparación.

Posiblemente en la actualidad un nivel geográfico adecuado para elaborar baremos sea el de nacionalidad o comunidad autónoma, aunque un nivel de estado, de comarca o de ciudad pueden considerarse también adecuados siempre que la muestra sea adecuada y aquel sujeto con el que se vaya a comparar pertenezca a esa población muestreada.

La cuestión de si deben efectuarse baremos separados por sexos depende esencialmente de si ambos grupos difieren o no significativamente en los principales estadísticos que describen la distribución de la variable medida. La distribución misma debe observarse cuidadosamente para evitar que diferencias no manifiestas en los estadísticos produzcan efectos indeseados (p.e. si una distribución tiene un comportamiento anómalo en una cola, o en una determinada zona de la escala).

En general, si *no* existen diferencias significativas entre los sexos en la variable medida, puede ser razonable utilizar un mismo baremo. Ahora bien, puede tener sentido por razones prácticas efectuar una comparación con el grupo total aun cuando existan diferencias significativas entre los grupos. Habrá que tener en cuenta que, en ese caso, la interpretación tenderá a beneficiar a un sujeto del grupo con distribución más desplazada hacia la derecha y tenderá a perjudicar a un sujeto del grupo cuya distribución esté más desplazada hacia la izquierda, supuesto que el polo deseable de la variable esté a la derecha. Sin embargo, en ese caso, es inexcusable, además, que se efectúe la comparación con el grupo del propio sexo como elemento de interpretación principal.

Debe tenerse en cuenta que la media de los grupos no es el único estadístico relevante para tomar la decisión de utilizar baremos únicos o separados. Para dos grupos con la misma media pero distintas dispersiones, o distintos sesgos, puede seguir siendo más correcto efectuar baremos separados.

El problema de los baremos separados es que, generalmente, implicarán muestras más reducidas, lo que, en general, perjudica la calidad de la comparación sujeto-baremo. Si la variable medida se trata de una característica en la que se aprecian diferencias entre ambos sexos y se puede obtener una buena muestra -a ser posible en torno a 1000 sujetos- de cada sexo, los baremos separados serán una solución adecuada.

El lector puede contestar la pregunta de baremos únicos o separados para otras características diferentes del sexo que pueden permitir separar muestras por analogía con los argumentos anteriores.

La calidad de las muestras es esencial para determinar el valor de un baremo. Y la calidad del baremo a su vez es esencial para que el test pueda utilizarse ofreciendo una buena interpretación de las puntuaciones de los sujetos. Baremar los tests implica estudios específicos que pueden ser muy costosos en tiempo y esfuerzo.

Debe recordarse que una muestra sólo puede considerarse representativa de la población de la que realmente se ha extraído por muestreo. Sin muestreo las muestras no son muestras propiamente y entonces el extraordinario instrumento de la inferencia no está justificado. En nuestro país no existe una política pública que garantice y financie sistemáticamente este tipo de estudios de medición psicológica y educativa. Sí la hay en otras áreas técnicamente comparables como encuestas políticas, de opinión y por supuesto de mercado, y sí la hay en otros países occidentales donde existen organismos técnicos dedicados a la evaluación y medición psicológica y educativa. En estas condiciones en nuestro país los investigadores se enfrentan a enormes dificultades prácticas para obtener muestras adecuadas. Muchos tests publicados tienen dificultades en el tamaño, en la localización y representación geográfica, en las características o en el tamaño de las muestras.

Otro problema considerable es la antigüedad de los baremos. Los comportamientos humanos están irremisiblemente anclados históricamente y la clase de problemas que resuelven bien los niños de una cohorte difieren de los que son capaces de resolver los de otra cohorte 10 años después. Esto implica la necesidad de renovar baremos periódicamente o de un modo cíclico, incorporando información actualizada. En muchos casos los mismos tests deben sufrir pequeñas o grandes modificaciones en su contenido en ese proceso de actualización. De nuevo para ello es evidente la necesidad de disponer de recursos e instituciones para realizar estudios sistemáticamente, y no de modo aislado.

Los tests deben dar en sus manuales cumplida cuenta de las características de la muestra o muestras de baremación y de validación. Debe informarse sobre sus características personales, demográficas, geográficas, culturales y sociales que sean relevantes a la naturaleza de la prueba. Entre ellas, pero no exclusivamente, debería informarse claramente de las distribuciones de las características de edad, sexo, grado escolar y cultural, colectivo y nivel profesional, si procede, así como del contexto y condiciones de administración de las pruebas, año o años en que fueron administradas y cuantas características permitan valorar el método de muestreo utilizado y las limitaciones a que está expuesta la muestra obtenida.

Muchos de los tests de los que disponemos han sido creados originalmente en otra lengua y contexto cultural y después adaptados. Este proceso de ***adaptación del test*** es en realidad un nuevo proceso de validación y baremación en la nueva población a la que pretende aplicarse el test, aunque se beneficie de la experiencia obtenida en la validación y baremación en la población original.

Un test no se puede simplemente “traducir”. Hay que adaptarlo, que es mucho más que eso. *Adaptar un test* significa que hay que volver a efectuar en la nueva

población de interés prácticamente todos los estudios que llevaron a su construcción, validación y baremación, quizás con la excepción del paso de formulación de items que aquí se convierte en traducción (aunque también en adaptación de expresiones, ideas etc.).

Para mejorar la fidelidad con el original se utiliza un método de *doble traducción*. La doble traducción consiste básicamente en que el test en la lengua origen es traducido por algunos expertos a la lengua destino y después otros expertos, que no conocen el original ni el proceso de traducción anterior, vuelven a traducirlo a la lengua origen. La versión real en la lengua original y la versión fruto de la doble traducción en la lengua original son comparadas a la búsqueda de dificultades semánticas, cambios indeseados, etc. Para mejorar la calidad si es posible se utiliza más de un traductor o más de un equipo de traducción en cada proceso de traducción, y finalmente en el proceso de comparación todos los expertos pueden participar en discutir las diferencias y las dificultades halladas para tratar de ofrecer las mejores soluciones.

Es importante desde luego obtener una traducción adecuada. Sin embargo, en mi opinión la fidelidad a lo que decía el test original es menos importante que la fidelidad a lo que trataba de medir. Una buena fidelidad filológica no tiene porque garantizar los mismos procesos de modo que garantizar una extremada fidelidad lingüística podría llegar a dificultar medir lo que se pretendía medir. Que la fidelidad lingüística es sólo una parte –pero no la esencial del problema- es patente cuando se observan ciertas dificultades con tests que sólo utilizan material gráfico. Allí no hay problemas de traducción, pero *un mismo problema* puede interpretarse de modo muy diferente por dos personas de culturas distintas. Hasta la dirección en que las personas leemos "un problema" gráfico y esperamos encontrar las soluciones tiene raíces culturales. Para nosotros, por ejemplo, los problemas aunque sean gráficos tienden a ser leídos, en general, de izquierda a derecha y de arriba abajo, probablemente porque este es el modo en que nos han entrenado a leer el texto escrito, pero esto no es así en todas las culturas. Al adaptar tests a otras culturas más que la literalidad habría que conservar la capacidad de medir lo mismo. No es pues una cuestión de "traducir" bien sino de medir lo mismo.

Un problema esencial es que ese "lo mismo" debe poder definirse transculturalmente para poder adaptar un test de una cultura a otra. Sin embargo, en mi opinión las mismas aptitudes, actitudes, factores de personalidad, etc. pueden no tener ninguna existencia ni significado fuera de su ámbito cultural. Nuestras teorías psicológicas también son un producto social e histórico y cual es el sustrato general del comportamiento humano a través de épocas y culturas es una cuestión empírica difícil de dilucidar. ¿A alguien se le ocurría medir personalidad tipo A o necesidad de logro en una comunidad campesina en Vietnam o sin ir tan lejos simplemente en un pueblo tradicional de una zona rural?

Podría argumentarse que, simplemente, estas personas tienen puntuaciones bajas en esta dimensión. En mi opinión está opinión esta sesgada, presupone la universalidad del rasgo y la certeza de la teoría, y es semejante a la de los colonialistas que encontraban ignorantes a los pigmeos porque no sabían tomar el té o geografía británica. No veo la evidencia que sostenga la aplicabilidad ni siquiera la existencia de tales dimensiones en estas circunstancias.

Aunque el constructo a medir tenga sentido en las dos culturas, aunque se consiga traducir de modo que ambas versiones "midan lo mismo" el proceso de adaptación sólo acaba de empezar. Es necesario volver a afirmar la validez de la medida en la población de destino.

Por supuesto es absurdo psicométricamente "traducir" baremos. En la mayoría de las variables de interés no puede sostenerse en absoluto que los sujetos de países distintos formen una misma población. Por tanto no tiene ningún sentido utilizar los baremos de otros. Es como limpiarse los dientes con el cepillo del vecino del sexto o ponerse su ropa interior.

Los psicólogos han de ser cuidadosos con la elección del material psicométrico que adquieren para aplicar a sus clientes (tanto en clínica, como en educación, como en organizaciones, como en servicios sociales etc.). Aquí hemos comentando los problemas relativos a la baremación, pero estos no son los únicos ni quizás los más importantes que pueden presentar los tests. Si un test no tiene una validación aceptable entonces carece de valor como instrumento de medición psicológica. Esto sucede a veces con tests "adaptados", cuyos estudios originales son valiosos pero cuyos estudios de adaptación sobre la población de destino adolecen de defectos técnicos.

Cuando un psicólogo o un gabinete de psicólogos trabajan de modo profesional en un campo especializado durante un tiempo suficiente pueden contribuir a desarrollar sus propios estudios de validación y sus propios baremos. Algunos gabinetes importantes desarrollan así sus propios baremos. Esto es particularmente difícil para tests especializados en el trabajo psicológico en clínica y patología -salvo que se esté conectado a un gran hospital o institución- pero relativamente fácil en aquellos contextos donde se trabaja con colectivos numerosos, como en educación o en empresas. Un términos generales una psicóloga o psicólogo siempre debe estar aprendiendo sobre los tests que utiliza, porque cada caso o cada grupo de casos es una oportunidad para entender mejor qué significan las respuestas de los sujetos, y, en términos de baremos, para acumular información que puede ser útil para rebaremar la prueba.

Un recurso adecuado para obtener muestras de tamaño adecuado -en torno a unos mil sujetos- consiste en acumular los resultados durante años sucesivos. Los baremos han de actualizarse, y un recurso técnico correcto y práctico consiste en renovar cada año los sujetos del año

más antiguo que incluye el baremo. Por ejemplo, si hemos necesitado n años (pongamos 5 años) para reunir una muestra de N sujetos (pongamos 978 sujetos) en un baremo de alta calidad, a partir de ahora, cada año retiraremos del baremo los aproximadamente N/n sujetos más antiguos (es decir, por ejemplo, los 207 sujetos que conseguimos el primer año) y los sustituiremos por los nuevos datos obtenidos este mismo año. Y volvemos a baremar. De este modo el baremo siempre se refiere a los últimos n años y mantiene, si el trabajo es regular, una muestra razonable y actualizada de sujetos. Esta técnica es utilizada por algún prestigioso gabinete de psicología educativa que utiliza sistemáticamente baterías completas de tests en un seguimiento psicológico y educativo de los sujetos cada dos cursos aproximadamente.

De todos modos los mismos tests no pueden permanecer inalterados durante muchos años y han de ser actualizados y vueltos a validar periódicamente. Por supuesto, cuando un test es actualizado y se cambian, adaptan, eliminan, o incluyen items, los baremos anteriores hay que desecharlos y volver a comenzar.

Un ciclo indicativo para volver a validar y reformular un test puede estar en torno a unos diez años para muchas variables psicológicas, pero esto depende esencialmente de la naturaleza de la variable medida y de los cambios sociales, educativos y culturales que le pueden afectar.

Un test no puede ser un anticuario de items y debe estar adaptado a las circunstancias educativas, históricas y sociales de los sujetos a los que se aplica. Los casos de incumplimiento de este principio pueden ser notorios o no, pero en ambos casos la calidad del test y de las mediciones con él efectuadas se verá afectada.

Por supuesto no deben utilizarse baremos de sujetos universitarios (por lo general y casualmente estudiantes de psicología) como baremos para la población general. Es legítimo obtener baremos de estudiantes de psicología pero sirven justamente para comparar con ellos a estudiantes de psicología. Incluso si uno desea una muestra general de estudiantes universitarios debe muestrear toda la universidad, no solo el patio de su casa.

Un caso particular muy relevante para muchos profesionales de la psicología lo constituyen los baremos de niños. Aquí hay que tener cuidado con aquellas dimensiones psicológicas expuestas a rápido cambio evolutivo y muy especialmente en las épocas de la vida de más rápido cambio evolutivo. En estos casos comparar a un niño con una muestra de niños tan solo seis meses mayor que él puede, en algunos casos, no ser completamente adecuado. Procede disponer de baremos distintos para tramos cortos de edad, tan cortos como lo requiera el cambio evolutivo de la variable en cuestión, lo cual puede resultar difícil en términos prácticos.

Esto sucede, por ejemplo, con aspectos de preescritura y prelectura en el periodo crítico de adquisición de esas habilidades, tanto por cambios madurativos como por cambios inducidos por el sistema educativo, padres, etc. No podemos tener unos baremos de estas variables, por ejemplo, para niños entre 4 y 6 años, porque el cambio ahí es tan rápido que el baremo no serviría bien para niños ni de 4, ni de 5, ni de 6. Procede baremar por periodos cortos que pueden ser, en esa etapa y para este tipo de temas, de unos seis meses. Baremar mes a mes puede parecer de una precisión compulsiva aunque quizás mejorará algo la comparación, si es que podemos disponer de muestras de tamaño adecuado (cosa que, a este nivel de detalle, es extraordinariamente infrecuente, salvo que estemos conectados profesionalmente con alguna gran institución educativa pública o privada o conjuntos de ellas).

Debe advertirse que esta clase de necesidades de muestro adecuado, descripción rigurosa de la muestra de baremación y actualización de la información no es exclusiva del modelo clásico. Estas dificultades son inherentes a la naturaleza histórica, cultural y social de las variables psicológicas que no deberíamos olvidar nunca, y, por tanto, no pueden resolverse utilizando un modelo de medida más sofisticado –aunque quizá puedan paliarse algunas de las dificultades prácticas–.

2.8. Interpretación de los percentiles

Los percentiles son las puntuaciones normativas más utilizadas. Casi inevitablemente un psicólogo que utilice mediciones tendrá que ser capaz de obtenerlos y, sobre todo, de interpretarlos adecuadamente como parte de su trabajo. Y es muy posible que tenga que ayudar a otros, clientes, familiares y otros profesionales, a interpretarlos adecuadamente.

Una vez que disponemos de una tabla, denominada **baremo**, que relaciona valores de la variable (puntuaciones totales en el test) y percentiles o porcentajes acumulados hasta ese punto, el uso e interpretación es bien sencillo.

Se administra el test a un sujeto, se obtiene la puntuación total. Se busca la puntuación total en la tabla y se determina que porcentaje del grupo normativo iguala o deja por debajo. Un sujeto con 21 puntos que, según la tabla, iguala o deja por debajo al 58% decimos que está en el percentil 58.

Los percentiles son muy fáciles de obtener y de interpretar, y las personas no especializadas pueden comprender su significado de modo sencillo con pocas explicaciones, lo que es una ventaja en el trabajo práctico para comunicar los resultados de las mediciones psicológicas a los interesados, sus familiares u otros profesionales.

Los percentiles tienen significado constante en términos de porcentajes de una muestra. Por tanto, si un sujeto está en

el percentil 80 en razonamiento verbal, y en el 50 en el razonamiento matemático, en términos de posición en la muestra podemos decir que está mejor en el razonamiento verbal. Esto permite comparar a los sujetos entre sí en el mismo test o a las puntuaciones de un mismo sujeto en el mismo o en diversos tests en términos de diferencias en porcentaje acumulado sobre la o las respectivas muestras normativas.

Sin embargo, *los percentiles no tienen un significado constante en términos de puntuación en el test.* La distancia en puntos de test entre el percentil 5 y el 10 puede ser muy diferente de la distancia en puntos entre el percentil 45 y 50. En muchos tests donde los sujetos tienden a estar más acumulados en las puntuaciones en torno a la mediana, dos percentiles consecutivos en los percentiles muy altos o muy bajos pueden suponer muchos puntos de diferencia en el test. Sin embargo, es característico que, en la zona en torno a la mediana, donde se acumulan más casos, dos percentiles consecutivos ocupen la misma puntuación del test o puntuaciones muy próximas.

Por supuesto tests distintos tienen distintos valores en un mismo percentil. Pero, además, un mismo test tiene valores distintos en un mismo percentil dependiendo del grupo normativo de que se trate. Esto último produce que una misma puntuación de un sujeto en un test corresponde a percentiles distintos en función del baremo (grupo normativo) que se consulte. Este fenómeno lo analizaremos en un apartado siguiente con más detalle.

Los percentiles reflejan una unidad constante en términos de porcentaje de muestra. Un 1% en una misma muestra siempre es un 1%, sea el que va entre 0 y 1 ó el que va entre 77 y 78. Pero los percentiles no reflejan una unidad constante en términos de puntuaciones directas del test. Por ejemplo, el percentil 6 puede ser 44 y el 7 ser 50 (una diferencia de 6 puntos), pero el percentil 51 ser 101 y el 52 ser 102 (una diferencia de 1 punto). Por supuesto utilizamos percentiles porque estamos interesados en una interpretación normativa, relativa al grupo normativo de comparación. Si estamos interesados en una interpretación y comparación en términos de puntos en el test no necesitamos para nada percentiles.

Un informe psicológico expresado en percentiles, especialmente si estos se reflejan gráficamente (por ejemplo a modo de “termómetros” de 0 a 100 donde una raya roja sube hasta el punto que refleja el percentil en que está el sujeto), es fácilmente inteligible y permite a las personas evaluar comparativamente su posición en diversas áreas, capacidades o factores.

Este tipo de informes son de uso extendido, y bien aceptados por las personas a las que van dirigidos, clientes, educadores, padres, estudiantes, trabajadores o empresarios. No obstante siempre deben acompañarse de las explicaciones verbales pertinentes que aclaren qué

significan los percentiles, que significado tiene cada nivel de puntuaciones, qué significan las variables y qué importancia tienen para el sujeto.

Los informes deberían poder completarse siempre, además, con informaciones verbales, al menos colectivas, que aclaren bien las dudas y malinterpretaciones que las personas no formadas en psicología pueden hacer y hacen de los tests y de sus resultados. La mejor medición con una explicación insuficiente o inadecuada no alcanzará su propósito de ser útil a las personas para tomar sus propias decisiones sobre sus vidas. Siempre debe haber una oportunidad de consultar personalmente y debería disponerse la entrega de la información de modo que la psicóloga o psicólogo pueda estar presente y explicar las cosas.

2.9. Las puntuaciones típicas z

Cualquier variable -y por tanto también la puntuación total de un test- puede expresarse de tres formas distintas:

1. En **puntuaciones directas**: En el caso del total de un test clásico con respuesta verdadera esto suele significar el número de items acertados. La unidad de medida es la propia de la variable: por ejemplo si estamos midiendo metros las directas están, claro, en metros, y si son aciertos, entonces 1 punto es 1 acierto.

2. En **puntuaciones diferenciales**: La puntuación directa menos la media. La media de las puntuaciones diferenciales es 0 y la desviación típica la misma de las directas. La unidad de medida es la propia de la variable, igual que en las directas.

3. En **puntuaciones típicas**: La puntuación diferencial dividida por la desviación típica. La media de las típicas es cero y su desviación típica 1. La unidad de medida ya no es la de la variable: la unidad de medida es la desviación típica. Una puntuación típica es el número de desviaciones típicas que un sujeto dista de la media.

Por ejemplo si la desviación típica de un test (en directas o diferenciales) es 16 y un sujeto tiene una puntuación típica de -1 ello significa que está 16 puntos por debajo de la media. Si la media en directas fuese 100 entonces este sujeto tiene una puntuación directa de 84 y una diferencial de -16.

Puede decirse que una puntuación típica mide la distancia a la media en desviaciones típicas. Dicho de otro modo la unidad de las puntuaciones típicas es la desviación típica de la variable. Cualquier variable cuando se transforma a típicas queda expresada en esa unidad abstracta. Esta

cualidad es muy interesante porque permite comparar las puntuaciones típicas de diferentes distribuciones de diferentes variables: En todas ellas una típica es la distancia a la media en unidades de desviación.

> En este sentido una típica de +3 siempre significa una desviación típica más lejos de la media por la derecha de la distribución que una típica de +2, independientemente de la naturaleza de las variables de que hablemos. La puntuación +3 estará siempre una desviación típica más a la derecha de la media (por encima) que una típica de +2.

Sin embargo, esa unidad "una desviación típica", que es igual en abstracto, no tiene porqué serlo en términos de puntuaciones directas. Ni siquiera cuando tratamos de un mismo test aplicado a dos grupos normativos distintos.

> Podemos encontrar que un mismo test presenta en un grupo más heterogéneo una desviación típica de 7'89, por ejemplo, mientras que en otro grupo normativo más homogéneo en la variable presenta una desviación típica de 4'12, por ejemplo. Por supuesto, entonces una misma típica de +2 significa cosas distintas en términos de puntuaciones directas en ambas distribuciones, aunque significa lo mismo en términos relativos de unidades de desviación de la media. Téngase en cuenta que un mismo test administrado en dos muestras distintas presentará, aunque sólo sea como variación muestral, una media distinta y una desviación típica distinta. Por tanto distintos grupos normativos darán lugar a distintas puntuaciones típicas para una misma puntuación directa de un mismo test.

Las transformaciones de directas a diferenciales o a típicas suponen un cambio lineal de escala, dejando inalterada la forma de la distribución y la correlación de la variable con cualquiera otra. Es simplemente un modo de reexpresar la puntuación en otra escala, sin que ello produzca ninguna distorsión sobre la información.

Si se está familiarizado con las puntuaciones típicas y sus características, disponer de una tabla de baremos que exprese para un grupo normativo que típica corresponde a cada directa, o, simplemente, transformar cada directa en típica sustrayéndole la media y dividiendo por la desviación típica, puede ayudar a comprender la posición relativa de un sujeto en la distribución de puntuaciones del grupo normativo que se utilice. Pero esto implica, efectivamente, familiaridad con estas puntuaciones para ese test, lo que no las hace demasiado idóneas para comunicar resultados a las personas medidas, y, a veces, ni siquiera a otros profesionales.

2.10. Puntuaciones típicas derivadas

Un inconveniente adicional de las puntuaciones típicas es que, al tener media 0, la mitad de las típicas aproximadamente que encontremos serán negativas, lo que las hace de manejo incómodo. Además, en la mayoría de las distribuciones es excepcional encontrar una típica por debajo de -3 o por encima de +3, lo que significa que los decimales son muy importantes en la escala.

Para solucionar esto se pueden efectuar transformaciones lineales de las puntuaciones típicas que cambien su escala, (muevan hacia arriba su media y utilicen una unidad de medida más pequeña que la desviación típica, pero proporcional a ella, que evite que los decimales sean tan relevantes). Estas transformaciones se conocen como escalas típicas derivadas.

Todas estas escalas se basan en multiplicar la típica por una constante b y sumar otra constante a, para lograr una nueva puntuación transformada P. La constante b determina la unidad de medida y la constante a la posición de la media en la escala resultante:

$$P = a + bz$$

Dos de los esquemas más utilizados de transformación lineal de puntuaciones típicas en puntuaciones típicas derivadas son:

$$T = 50 + 10\,z$$

$$D = 50 + 20\,z$$

Una vez que el psicólogo se ha familiarizado con una escala típica derivada determinada, por ejemplo, con las puntuaciones T, ésta resulta un modo muy intuitivo de reflejar las puntuaciones de los sujetos, actuando como una escala común que se interpreta razonablemente de un modo semejante de test a test o de sujeto a sujeto como sucede con las mismas puntuaciones típicas. Que las personas no especialistas en psicología se familiaricen lo suficiente con estas escalas puede ser un poco más difícil, especialmente si su contacto con una medición psicológica y sus resultados es esporádico. Si se utilizan hay que esforzarse en dar las explicaciones oportunas para que la interpretación sea adecuada.

En realidad hay una infinidad de escalas típicas derivadas. Muchos de los tests más utilizados han acuñado su propia escala de transformación. La escala en que se expresen las

típicas es arbitraria, pero, como se ve, no se ha alcanzado un alto grado de convención, que es la característica requerida inmediatamente a todo lenguaje arbitrario para ser útil. La cuestión reside en que, para cada test habrá que familiarizarse con el modo en que se reexpresan sus típicas.

Por cierto que no hay inconveniente en que el lector elabore escalas típicas derivadas propias, si le place, con números para la media "a" y para la unidad de medida "b" más de su gusto. Por ejemplo utilizando el día y mes de su cumpleaños, o, por ejemplo, creando una escala para que coincida básicamente con el esquema tradicional de notas de 0 a 10 ¿qué transformación debería hacerse para conseguir esto último?. Obviamente la arbitrariedad de estas escalas, como la de las escalas típicas derivadas de algunos tests, es la principal objeción para esta multiplicidad.

2.11. Puntuaciones típicas normalizadas

Una puntuación típica normalizada es la típica que le hubiera correspondido a una puntuación directa si, en lugar de formar parte de su distribución empírica real, hubiera formado parte de una distribución normal perfecta.

Se obtiene en dos pasos: Primero hay que conocer que porcentaje acumulado corresponde en la distribución a cada puntuación directa. Segundo mediante una tabla de la curva normal (ó una calculadora) se establece que puntuación típica z' de la curva normal corresponde a ese porcentaje acumulado.

Una vez que se dispone de una tabla de baremos indicando para cada puntuación directa que z' le corresponde en la curva normal, podemos convertir, según esa tabla, el resultado de cualquier persona en el test en la z' correspondiente para ese grupo normativo. La comprensión de la puntuación resultante supone estar familiarizado con la curva normal.

Este procedimiento produce una distorsión sistemática de la información original que es forzada a adoptar una forma de distribución normal a partir de los porcentajes acumulados independientemente de cual sea la forma real de la distribución. Si la distribución empírica, real, observada, del grupo normativo ya es normal o casi normal no tiene sentido convertirla en lo que ya es, y si no lo es todavía tiene menos sentido elaborar baremos en base a distorsionar la distribución real. Mi opinión personal es que este método de trabajo debe ser abandonado porque distorsiona deliberadamente la información. Una discusión crítica más detallada de esta cuestión y un ejemplo de aplicación puede verse en Meliá (1991); más información y ejemplos acerca de como aplicar y hacer operar el método puede encontrarse, por ejemplo, en Yela (1984) que recomienda utilizarlo cuando la distribución empírica se aproxima a la normal.

2.12. Crítica de la aproximación normativa

Esta necesidad de comparar la puntuación de un sujeto para poder interpretarla significa que la interpretación de la puntuación del sujeto es *relativa a las del grupo normativo* con el que se le compare. Si se dispone de normas relativas a diferentes grupos normativos con los que resulta razonable comparar a un mismo sujeto es posible obtener interpretaciones distintas en función del grupo de comparación.

Por una parte esto puede interpretarse como una **indefinición** de la teoría métrica y psicológica (p.e. ¿cómo describir bien la inteligencia de un sujeto si esto depende del grupo con el que se le compare? ¿cuál es la "verdadera" inteligencia del sujeto?) Paradójicamente una teoría que comienza explicando que cada puntuación empírica se debe a una variación aleatoria sobre una puntuación verdadera termina sosteniendo que no hay tal cosa como la puntuación "verdadera" y que, al final, sólo se trata de saber que posición ocupa el sujeto en el grupo. Para llegar ahí sobra toda la teoría sobre puntuaciones empíricas y verdaderas, y toda la teoría de la fiabilidad montada sobre esta concepción.

De hecho, la mayoría de los tests clásicos construídos y contrastados bajo la teoría clásica de tests:

1.Valoran cada ítem acertado con un punto y construyen el total como la suma de puntos.

Esto significa ignorar el índice de dificultad de los items y cualquier otro estadístico referido a los items. En la práctica se tratan los items como "todos iguales", el muy fácil y el muy difícil.

2. No pueden interpretar el total directamente en términos psicológicos.

Esto significa una ausencia de elaboración de teoría psicológica para cada total.

Desde una nueva concepción más profunda de validez, en el marco de lo que denomino teoría restrictiva de la validez, de la que nos ocuparemos después, podríamos considerar que un **test válido** *en la medida en que pudiéramos atribuir un significado psicológico determinado e inequívoco a cada posible puntuación total*.

Esto no implica necesariamente un modelo determinista que resultaría irreal, aunque si un trabajo de validación directamente enfocado a contrastar el significado de las puntuaciones, del conjunto de puntuaciones y de cada valor en la escala. Bajo el modelo tradicional, este estudio se efectúa sólo superficialmente estableciendo alguna función lineal

que relacione el total del test con algún criterio o criterios.

Con esta definición de validez el enfoque normativo de interpretación de las puntuaciones podría considerarse muy deficiente.

3. Interpreta la puntuación total como la posición relativa del sujeto en un grupo.

Toda la teoría métrica que hace falta para esto es suponer que t+1 puntos es más (no sabemos cuanto más) que t puntos y que es legítimo comparar a los sujetos con determinados grupos en función de *alguna* afinidad.

Dicho de otro modo la interpretación normativa de las puntuaciones sólo dice a cuantos sujetos ha "ganado" o "superado" un sujeto dado. Una información por sí lamentablemente pobre.

O bien se sabe que significa esa posición en el grupo en términos psicológicos, o bien no se sabe. Si se sabe, entonces la puntuación normativa es innecesaria: ¿para que convertir un total en una puntuación normativa para traducir esta a un significado psicológico? Bastaría con traducir la puntuación total a su significado psicológico (una parte del cual quizás relativamente poco importante es saber que posición supone en un grupo). Si no se sabe el significado psicológico de la puntuación normativa entonces prácticamente no se sabe nada. Esto equivale a afirmar que con las puntuaciones de muchos tests construidos bajo teoría clásica prácticamente no se sabe nada.

Estos razonamientos nos aclaran la perplejidad legítima de cualquier estudiante cuando después de estudiar muchas páginas de teoría clásica orientadas -se suponía- a poder contrastar la calidad de las puntuaciones de los tests y a poder explicarlas (la misma descomposición de X en V y E no es más que un intento de explicación) encuentra que, al final, se toma el total del test y se dice, por ejemplo, como interpretación final "este sujeto deja por debajo de sí al 76% del grupo normativo". ¿Y...? ¿Qué significa eso en términos psicológicos ...? ¿Qué le pasa al sujeto? ¿Qué clase de problemas es capaz de resolver y qué clase no? ¿Qué síntomas presenta y cuales no? ¿Cuál es su conducta típica, mínima o máxima en la variable bajo estudio? Y son preguntas perfectamente legítimas porque estas son las preguntas importantes.

Después de llegar a la puntuación normativa, los estudiantes preguntan aquí "¿y cómo se interpreta la puntuación normativa?" Es una pregunta paradójica que la aproximación usual de la teoría clásica de los tests no está preparada para contestar adecuadamente. Esa aproximación supone que la puntuación normativa *es* la interpretación. Pero ¡si esta puntuación normativa no es mas que saber si el sujeto es el primero el segundo o el vigésimo del grupo! La puntuación normativa en sí es una aproximación extraordinariamente pobre para la que,

además, es indiferente toda la teoría clásica, desde sus fundamentos hasta la más sofisticada fórmula de fiabilidad.

Puede argumentarse que si el grupo normativo describe bien la población general (es decir, es una buena muestra de la misma) entonces las normas son un buen punto de comparación. Por ejemplo, si tenemos una muestra excelente de la población general en inteligencia (como sucede con algunos tests clásicos, especialmente en poblaciones USA) puede argumentarse que ésa es una buena descripción de los grados o niveles de inteligencia que *de hecho* existen en la realidad y que es legítimo ubicar al sujeto en un punto preciso de esos niveles a partir de su ubicación relativa en la muestra. No se cuestiona esto, pero, la cuestión sigue siendo que esto, sin más no dice apenas nada en términos psicológicos.

El problema de la población de referencia. Un problema adicional esencial es *la elección del grupo normativo con el que comparar.* Un test clásico bien estudiado vendrá acompañado de diferentes y variados baremos y hay que decidir con cual o cuales comparar al sujeto. Si el sujeto sólo puede compararse razonablemente con uno de esos grupos la psicóloga o psicólogo no tiene alternativa. Pero si es razonable comparar al sujeto con más de un grupo, entonces la psicóloga o psicólogo se enfrentan ante una decisión que no siempre es fácil de fundamentar. La cuestión es que, al menos hasta donde yo se, carecemos de una teoría sobre la elección de población de referencia bien fundamentada cuando hay más de una posible población de referencia razonable.

El problema aquí es, de hecho, el inverso del muestreo que sostiene la inferencia. En el problema clásico del muestreo se dispone de la población y se arbitran métodos para elegir una muestra adecuada, representativa y suficiente. La cuestión aquí es que 1) se dispone de varias poblaciones o subpoblaciones de referencia que fueron en su día muestreadas (en el mejor de los casos) para obtener grupos normativos, 2) se dispone de un sujeto para el que la elección de grupo normativo puede no ser indiferente y 3) el sujeto comparte parcialmente características que describen los grupos normativos y puede tener sentido compararlo con más de uno de ellos.

Por ejemplo, un test puede aportar baremos por edades (uno para cada edad o grupo de edad), por sexos, por grupos profesionales. Aunque la cuestión principal no es si el test los lleva o no, pues aunque no los lleve estos baremos variados *deberían* elaborarse si hay una razón para ello, y no deberían elaborarse si hay una razón para ello. Y una vez elaborados y disponibles ¿qué procedimiento de elección debe seguirse? En términos prácticos la respuesta puede tener una solución fácil, pero en términos teóricos, si se pretende que se está midiendo una variable psicológica, la ausencia de una teoría de elección del grupo normativo a posteriori complica la interpretación de la puntuación del sujeto.

Si la posición del sujeto en el grupo normativo varía de grupo normativo a grupo normativo y escoger con qué grupo comparar depende de la afinidad entre sujeto y grupo, entonces deberían haber condiciones claras para establecer cuando hay suficiente afinidad y de qué clase o clases de un sujeto con un grupo normativo para poder aplicarle sus baremos. Sorprendentemente en teoría clásica hay poca o ninguna información más allá del sentido común en este respecto.

Por ejemplo, hemos aplicado un test de inteligencia a una mujer de 29 años maestra de la escuela pública (mañanas), estudiante de Psicología (tardes), casada y con un hijo (noches), que (sobre)vive y estudia y trabaja en la ciudad de València. Corregimos el test y obtenemos un total de puntos, por ejemplo 27. Vamos al manual a buscar la tabla de baremos y encontramos lo siguiente:

a. Tabla de baremos para la población general española (Muestra N=1.875). El manual explica que esta tabla se obtuvo en un estudio de hace una década.

b. Tabla de baremos para la población general española de mujeres (Muestra N=967) Se obtuvo en el mismo estudio que los baremos del punto a.

c. Tabla de baremos para mujeres profesionales. (N=754). El manual explica que se obtuvo en un estudio posterior sobre una muestra formada por mujeres que trabajan en su mayoría en el sector servicios, con una representación sustancial de mujeres que trabajan en educación (39%) y en sanidad (43%); el resto son otras profesiones.

d. Tabla de baremos para mujeres amas de casa (N=1321). Perfil de mujeres con uno o dos hijos y con “obligaciones” domésticas sustanciales impuestas por nuestro sistema social. Proviene de un estudio del año pasado.

e. Tabla de baremos de estudiantes de psicología. (N=645) obtenida de un estudio efectuado en la facultad hace tres años (El 89’7% de la muestra son mujeres. El 7’3% de la muestra son mujeres que trabajan. El 2’8% son mujeres que trabajan en enseñanza. El 1’7% son mujeres que trabajan en enseñanza y están casadas. El manual no especifica baremos para estos subgrupos pero si la composición detallada de la muestra.

f. Tabla de baremos por edades de la Comunidad Valenciana para el grupo entre 26 y 30 años. (N=545) Proviene de otro estudio del que no se menciona fecha.

Si comparamos la puntuación total de 27 con cada uno de esos baremos encontramos un resultado distinto para cada baremo. El resultado puede ser a menudo incluso marcadamente distinto. Por ejemplo, podemos encontrar que esa misma puntuación 27 significa un percentil 77 para los baremos d mientras que significa un percentil 43 para los baremos e o un percentil 54 para los f ¿Con qué baremos comparamos? La persona examinada comparte en algún grado características con todos estos grupos. Es decir, mantiene algún grado de afinidad con todos ellos,

pero ¿qué clase de afinidad es más importante: la edad, el sexo, la profesión, el rol social de esposa y madre, el rol social de estudiante...? La teoría clásica no ha resuelto en general está cuestión y, paradójicamente, el uso de baremos vuelve todavía más relativa la puntuación 27 que puede convertirse aquí en seis posiciones centiles distintas. La ausencia de criterio para elegir baremo produce una fuerte indefinición del significado de la puntuación incluso en el reducido (y muchas veces poco relevante) enfoque de determinar la posición relativa respecto a otras personas.

Si se apuesta por comparar a la persona con un grupo muy semejante en muchas características a ella misma podemos acabar comparándola con el 1'7% de la muestra e que tanto se le parece. Está formado por una muestra tan pequeña y accidental como un pequeño grupo de compañeras de carrera, profesión y estado. Se convendrá en que es bastante ridículo realizar todo este trabajo con un test para acabar informando si la persona en cuestión es la "primera de la clase". En el otro extremo, si se opta por la población más general, (por ejemplo, la población general española, suponiendo que nuestra persona sea española) la muestra puede ser tan distinta que la comparación esté poco justificada. Piénsese que esa muestra general puede contener, por ejemplo, sólo uno o dos casos más o menos parecidos a nuestra persona en *todas* las características, junto a mineros asturianos, bailadores andaluces, y campesinos castellanos, por mencionar solo algunos estereotipos más o menos infundados que ponen de relieve la impresionante heterogeneidad de la muestra.

La cuestión es ¿con qué propósito hemos "medido" la inteligencia de nuestra profesora estudiante de psicología? La mayoría de las veces aunque tengamos razonablemente claro el propósito, la indefinición y falta de significado psicológico que trasluce el ejemplo anterior para el método normativo no se desvanecerá suficientemente. Esto por supuesto produce una sensación de disgusto con este modo de trabajar en el que se comienza por renunciar a establecer un significado psicológico preciso para cada puntuación y se acaba por no saber tampoco el significado normativo de las puntuaciones.

Dicho sea de paso la Teoría de la Respuesta al Item es un ensayo para resolver *una parte* de este problema, aunque el problema de la elección de población de referencia y la cuestión esencial del significado psicológico pueden quedar esencialmente no resueltos (aunque quizá de modo menos patente).

3. Comparación entre puntuaciones particulares

3.1. Introducción: Comparación de puntuaciones como contraste de hipótesis

El modo más usual y sencillo de comparar las puntuaciones de un sujeto en el mismo o en diversos tests, o de dos sujetos en el mismo o en diversos tests consiste en reducirlas a una escala común, razonablemente comparable, como los percentiles o las puntuaciones típicas o las puntuaciones típicas derivadas. En esas escalas puede apreciarse comparativamente si una puntuación es mayor o no que otra y entender qué significa en términos de posición respecto a un grupo normativo de referencia o respecto a una unidad abstracta bien conocida denominada desviación típica. Para la mayoría de los propósitos prácticos este tipo de comparación es suficiente.

Este tipo de comparación permite observar si las puntuaciones son iguales o diferentes y si son muy diferentes o poco diferentes, esto es, si la diferencia entre las puntuaciones es grande o pequeña. Sin embargo, esta comparación simple por reducción a una escala razonablemente común no produce un juicio probabilístico acerca de la significación de las diferencias. Es decir, dada una diferencia determinada ¿en qué medida puede ser atribuida al error de medida que, como sabemos, introduce un componente de azar en cada puntuación empírica?. Esta pregunta puede hacerse de un modo más sencillo ¿En qué medida podemos creer que una determinada diferencia se debe al azar?

Si deseamos evaluar las diferencias entre dos puntuaciones en términos de contraste estadístico de hipótesis es necesario enfocar el asunto desde otro punto de vista, distinto al de obtener percentiles o puntuaciones típicas. Se tratará de evaluar la diferencia entre puntuaciones hallada en función de cuan probable sería hallar por azar una diferencia como esa en sujetos que realmente (puntuación verdadera) no difiriesen nada en la variable medida.

Las comparaciones entre puntuaciones de sujetos de las que nos ocuparemos en este apartado tienen precisamente un enfoque de contraste estadístico de hipótesis. Aunque realmente este tipo de contrastes son usados con poca frecuencia, tienen la virtud de asignar una probabilidad a las diferencias observadas entre puntuaciones, lo que permite tomar decisiones acerca de si tales diferencias parecen o no debidas al azar.

La cuestión es que, dado que los instrumentos llevan asociado un error de medida, dos sujetos que realmente no difiriesen (puntuación verdadera igual) probablemente presentarían alguna diferencia entre sus puntuaciones empíricas ¿cuan grande tiene que

ser una diferencia empírica entre puntuaciones para que creamos que está más allá de la variación que puede producir el error de medida? O, preguntado de otro modo, dada una diferencia concreta entre puntuaciones empíricas, ¿cuan probable es que diferencias de ese tamaño o mayores aparezcan por azar?

La hipótesis nula establece que no hay diferencias significativas entre las dos puntuaciones; o lo que es lo mismo que las diferencias entre ambas se deben al azar inducido por el error de medida en ambas puntuaciones; o lo que es lo mismo que la diferencia entre ambas puntuaciones no es significativamente distinta de 0.

Se sabe que la distribución de las diferencias entre dos puntuaciones empíricas cuando la puntuación verdadera es igual puede describirse razonablemente con una curva normal. Dicho de otro modo que la distribución muestral de las diferencias entre puntuaciones bajo hipótesis nula (cuando realmente no hay diferencias) es normal. La desviación típica de esa distribución muestral se conoce como error típico de la diferencia.

-La distribución muestral de las diferencias es una distribución hipotética formada por N (en el extremo igual a infinito) diferencias entre dos mediciones debidas al azar; es decir, a sabiendas de que los sujetos comparados realmente no difieren. Esta distribución ayuda a comprender cuan grandes y cuan frecuentes son las diferencias entre puntuaciones debidas únicamente al azar. Si además se averigua que esa distribución muestral es normal, entonces, como conocemos muy bien la distribución normal y sabemos cada punto de ella cuan probable es, podremos saber cuan probable es que una diferencia dada empírica pueda ser fruto del azar.-

Si se conoce la distribución muestral de las diferencias, que en ese caso se asume normal, para una diferencia dada, una vez estandarizada dividiendo por el error típico de la diferencia, podrá establecerse cuan probable es que una diferencia de ese tamaño (o mayor) haya aparecido por azar. Esta probabilidad nos ayudará a decidir si creemos que esa diferencia sea un efecto del azar (no rechazar la hipótesis nula) o no (rechazar la hipótesis nula, cuando el que esa diferencia sea fruto del azar sea tan improbable que una diferencia de ese tamaño o mayor aparezca sólo el 5% de las veces o menos).

3.2. Situaciones en que puede aplicarse a puntuaciones particulares el enfoque de contraste de hipótesis

Hay varias *situaciones* en las que podemos estar interesados en averiguar en que grado una diferencia entre dos puntuaciones puede deberse al azar introducido por el error de medida.

Las dos puntuaciones a comparar pueden pertenecer al mismo sujeto, en cuyo caso se dice que estamos evaluando diferencias intraindividuales; o pueden pertenecer a dos sujetos diferentes, en cuyo caso estamos evaluando diferencias interindividuales. Los casos más representativos han recibido una denominación específica en las exposiciones de orientación clásica sobre el tema.

Diferencias intraindividuales

-Evaluar la puntuación de un sujeto en un test en el tiempo 2 frente a la puntuación del mismo sujeto en ese test en el tiempo 1, para determinar si ha habido cambios en ese periodo.

Puede ser de interés para estudiar cambios evolutivos o para averiguar si ha habido cambios individuales, más allá de variaciones debidas al azar, después de algún tratamiento o intervención.

-Evaluar la diferencia entre las puntuaciones de un mismo sujeto en dos tests paralelos.

Teóricamente un mismo sujeto ha de presentar la misma puntuación verdadera en dos tests o formas paralelas, de manera que no debería haber diferencias más allá de las introducidas por la variación aleatoria que supone el error de medida. No obstante, si ambas pruebas se administran con un periodo temporal de diferencia estamos prácticamente frente a la circunstancia del caso anterior. En la práctica es frecuente que un tiempo 2 se administre una forma paralela en vez del mismo test para evitar efectos de memoria de las respuestas y preguntas.

Este caso se puede denominar “diferencias intraindividuales en tests paralelos” (Yela, 1984)

-Evaluar la diferencia de las puntuaciones de un mismo sujeto en dos tests diferentes.

Por ejemplo, estamos interesados en dos capacidades o dos factores de la inteligencia, averiguando si existe una verdadera diferencia entre las puntuaciones del sujeto en ambas.

Este caso se puede denominar "diferencias intraindividuales en tests distintos" (Yela, 1984)

Diferencias interindividuales

-Evaluar la diferencia de las puntuaciones de dos sujetos distintos en el mismo test, para determinar con que grado de certeza esas diferencias son reales (y no un efecto del azar).

Yela (1984) denomina este caso "diferencias interindividuales"

-Evaluar la diferencia entre las puntuaciones de dos sujetos en dos tests paralelos (a cada sujeto se le ha administrado una forma paralela distintas);

3.3. El enfoque clásico de las diferencias entre puntuaciones

El enfoque clásico se caracteriza por:

1. Analizar particularmente a qué es igual el error típico de la diferencia en cada caso, elaborando fórmulas particulares del error típico de la diferencia. Para cada caso se calcula un error típico de la diferencia 'particular', adecuado al caso.

El enfoque tradicional distingue si se trata de dos puntuaciones de un mismo sujeto en dos test paralelos, de dos puntuaciones de un mismo sujeto en tests distintos, o de las puntuaciones de distintos sujetos en el mismo o en distintos tests. Cada uno de los casos anteriores, en un enfoque tradicional, hay que explicarlos, y así se explican, por separado, como cosas distintas aunque relacionadas.

2. Dividir la diferencia entre puntuaciones (en directas) entre el error típico de la diferencia pertinente y,

3. Comparar esa diferencia estandarizada con una puntuación zeta crítica (1'96 para el nivel de confianza del 95%; ó 2'58, para un nivel de confianza del 99%). Si la diferencia es mayor que la zeta crítica se rechaza la hipótesis nula y se admite que las diferencias son significativas.

El enfoque tradicional propone evaluar la hipótesis de ausencia de diferencias en términos de aceptación o rechazo para un nivel convencional dado, por ejemplo, para un nivel alfa 0'05 o nivel alfa 0'01. Es decir, para un nivel alfa concreto, pongamos $\alpha = 0'05$, se determina que la zeta crítica que tiene que superar una diferencia, conforme a la curva normal, para ser significativa es $Z_C = 1'96$. Si la puntuación zeta empírica Z_e la supera o iguala se dice que hay diferencias significativas.

Que a un nivel $\alpha = 0'05$ corresponda una $Z_C = 1'96$ significa que en la curva normal las puntuaciones iguales o mayores a 1'96 solo aparecen el 5% de las veces o menos. (Es decir, si estamos hablando de diferencias entre puntuaciones, una diferencia de 1'96 o mayor tan solo aparecerá por azar un 5% de las veces).

En síntesis, el modo tradicional de resolver esta cuestión es casuístico (con un error típico de la diferencia distinto para cada caso que lo requiere) y orientado al contraste.

El lector puede consultar los manuales de Yela (1984; Cap. 5), o Santisteban (1990; Cap. 7), donde hay una exposición muy bien ilustrada acerca de estas cuestiones, analizada con el enfoque casuístico clásico.

4. Un método general de contraste de puntuaciones

4.1. Características del método general

En mi opinión puede proponerse otro modo de resolver estos casos que sea general y orientado al nivel de significación exacto de la diferencia entre puntuaciones.

Este método *general* permitirá al lector efectuar cualquier contraste entre puntuaciones particulares de sujetos, independientemente del caso concreto. Es decir, es útil tanto si las puntuaciones que se comparan provienen del mismo sujeto como si no, y tanto si provienen del mismo test como si no, y tanto si los tests son paralelos como si no. Con la única salvedad de que cuando estén en escalas distintas (p.e. diferentes tests) hay que pasarlas primero a una escala común (puntuaciones típicas).

El enfoque de contraste general que propongo es razonable siempre que tenga sentido la cuestión *¿cuán probable es que se deba al azar esta diferencia entre puntuaciones?*

Al contrario de lo que es usual el método que propongo está *orientado a determinar el nivel de significación*, antes que al contraste. En el método general que propongo aquí hay un matiz de enfoque, buscando establecer primero la probabilidad asociada a la diferencia (que puede ser mayor o menor que esos niveles convencionales) permitiendo una comprensión más gradual de la magnitud probabilística de la diferencia.

Por lo demás, el método general producirá los mismos resultados que el método particular clásico *adecuado al caso,* en términos de contraste.

Además, introduciremos un nuevo concepto, DSM, la Diferencia Significativa Mínima, que nos permitirá, para un error típico de la diferencia determinado y un nivel de confianza elegido saber inmediatamente que diferencias entre puntuaciones son significativas y cuales no.

4.2. Una fórmula general de error típico de la diferencia

El *método general de contraste de diferencias de puntuaciones* necesita una fórmula general del error típico de la diferencia, que valga para todos los casos.

Para simplificar llamaremos al error típico de la diferencia s_d, como es tradicional.

La varianza de la diferencia s_d^2 es igual a la suma de las varianzas de error asociadas a cada puntuación:

$$s_d^2 = s_{e_1}^2 + s_{e_2}^2$$

de ahí se desprende una fórmula general del error típico de la diferencia como raíz cuadrada de la suma de las varianzas de error (errores típicos de medida al cuadrado) de los tests o mediciones implicadas.

$$s_d = \sqrt{s_{e_1}^2 + s_{e_2}^2}$$

El error típico de la diferencia es la desviación típica de la distribución formada por N errores típicos de medida -obtenidos entre pares en los que se sabía que no había diferencias reales-.

En diversos casos esta fórmula general adopta diversas apariencias, lo que permite al enfoque tradicional tratar separadamente diversos casos, introduciendo, a mi juicio, una complicación innecesaria.

4.3. Pasos del método

El *método general de contraste de diferencias de puntuaciones* en todos los casos es siempre el mismo y solo presenta dos pasos:

Primero hay que tipificar la diferencia entre las dos puntuaciones a comparar dividiéndola por el error típico de la diferencia

$$z_d = \frac{X_1 - X_2}{\sqrt{s^2_{e_1} + s^2_{e_2}}}$$

Hay que hacer la salvedad de que si las dos puntuaciones están en escalas distintas (como suele ser si pertenecen a tests distintos) hay que pasarlas primero a típicas para que la diferencia entre ellas pueda hacerse con sentido.

Segundo, se determina el nivel de significación. Es decir, se determina -con la ayuda de una tabla de curva normal- qué probabilidad hay de que una diferencia como esa o mayor se deba al azar.

Explicación de los pasos

1. Tipificar una diferencia entre puntuaciones. Estamos interesados en tipificar la diferencia entre dos puntuaciones

$$X_1 - X_2$$

para poder comprobar cuan probable es que esa diferencia se deba al azar.

Para tipificar una diferencia hay que dividirla por el error típico de la diferencia. Esto equivale a ponerla en "puntuaciones típicas" comparables con las puntuaciones típicas de una tabla de curva normal.

Por tanto para tipificar la diferencia $X_1 - X_2$ dividiremos por s_d lo que dará como resultado la diferencia expresada en puntuaciones típicas, que designaremos por z_d:

$$z_d = \frac{X_1 - X_2}{s_d}$$

o, lo que es lo mismo:

$$z_d = \frac{X_1 - X_2}{\sqrt{s^2_{e_1} + s^2_{e_2}}}$$

2. Nivel de significación de la diferencia. En el segundo paso, se determina el nivel de significación de la diferencia que acabamos de tipificar.

> El nivel de significación de una diferencia z_d es la probabilidad con que una diferencia así o mayor aparece debido al azar cuando de verdad no hay diferencias.

Se determina el nivel de significación viendo en una curva normal con qué probabilidad aparecen diferencias de ese tamaño o mayores. Es decir, viendo que porcentaje de casos queda en la región de la curva entre $\pm z_d$ y los extremos.

Cómo se busca y obtiene el nivel de significación a partir de la tabla de curva normal depende de que tipo de tabla de curva normal tengamos.

Si disponemos de una tabla de "colas izquierdas de la distribución normal tipificada", es decir, que acumula la distribución desde menos infinito (tipo la que aparece en el manual de Santisteban, 1990; pag. 547 ó en el manual de Amón, 1979; pag. 370) entonces la probabilidad hallada en la tabla se sustrae de 1, y el resultado se multiplica por 2 para hallar el nivel de significación.

> Ejemplo: Para z=1'88 la tabla señala una cola izquierda de 0'97. 1-0'97=0'03; 2 por 0'03=0'06. El nivel de significación de la diferencia 1'88 es 0'06.

Si se dispone de una tabla que acumule desde el medio (desde la z=0) se determina la proporción (probabilidad) acumulada entre z=0 y la z_d, se multiplica ese valor por 2 (la curva es simétrica) y se resta de 1.

> Ejemplo: Para z=1'88 la tabla señala 0'4699. Este valor por dos es 0'9398. Y el resultado sustraído de 1 da: 0'0602, que es el nivel de significación e la diferencia estandarizada 1'88.

Si tenemos una tabla de curva normal que acumula desde la derecha hacia la izquierda (de más z a menos) se busca la probabilidad que corresponde a esa z_d en la tabla y se multiplica por 2. (Ejemplo de este tipo de tabla de "colas derechas de la curva normal tipificada" es la que aparece en San Martín, Espinosa y Fernández, 1.987; pag. 355).

Ejemplo: Para z=1'88 la tabla señala una cola derecha de 0'0301. Este valor por dos da 0'0602, que es el nivel de significación.

Una calculadora de sobremesa o un programa informático pueden dar directamente el nivel de significación del estadístico o un resultado análogo al de alguno de estos tres tipos de tablas que habrá que tratar del modo señalado para ese caso.

Ejemplo: Para una z=1'88, mostrando 6 decimales, obtenemos una cola derecha de 0'03005. Multiplicando por 2 vemos que el nivel de significación es, más exactamente, 0'060108. Los resultados obtenidos por las tablas son más que aceptablemente exactos para el uso corriente.

Direccionalidad del contraste de hipótesis. Debe tenerse en cuenta que el contraste de hipótesis clásico a este respecto es bidireccional, aunque a veces se hacen salvedades. Un contraste bidireccional supone que la región de rechazo de la hipótesis nula se sitúa a ambos lados de la curva normal, de modo que la hipótesis que se contrasta es si la segunda puntuación es tan grande *o tan pequeña* respecto a la primera que la diferencia no parezca razonable atribuirla al azar. La determinación del nivel de significación que hemos explicado es consistente con un enfoque de contraste bidireccional, que es el que, en general, parece recomendable hacer. Si estuviéramos efectuando un contraste unidireccional admitiríamos como nivel de significación la cola derecha de la distribución (sin multiplicar por 2) a partir de la diferencia tipificada.

Aunque puede parecer intuitivamente más razonable tratar algunos de los casos de contrastes enunciados como unidireccionales y el tema puede ser objeto de opinión y controversia, en general, en mi opinión, es preferible que efectuemos contrastes bidireccionales. La discusión pormenorizada de esta

cuestión nos llevaría muy lejos del tema que nos ocupa y no entraremos en ella. El lector puede encontrar en Welkowitz, Ewen y Cohen (1981; pags 174 a 178) algunas explicaciones sobre los argumentos principales que sustentan esta opinión.

Juicio sobre el nivel de significación;. Con un enfoque de contraste de hipótesis, si el nivel de significación es igual o menor a 0'05, se dice que la diferencia es estadísticamente significativa y se rechaza la hipótesis nula (que establece que las diferencias se deben al azar). Es decir, una diferencia así o mayor sucede tan raras veces cuando de verdad no hay diferencias que nos inclinamos a pensar que esa diferencia es real y no efecto del azar.

Con un enfoque de contraste de hipótesis, si el nivel de significación es mayor que 0'05, no rechazamos la hipótesis nula y decimos que la diferencia encontrada no es estadísticamente significativa. Es decir, una diferencia de ese tamaño no es lo suficientemente rara, cuando de verdad las dos personas no son diferentes, como para rechazar la idea de que puede deberse al azar.

Precisamente a un nivel $\alpha = 0'05$ corresponde $Z_C = 1'96$ con un contraste bidireccional, es decir, si la región de rechazo se establece a ambos lados de la curva.

A veces se adopta como nivel convencional para rechazar la hipótesis nula 0'01, lo que implica ser más exigente con una diferencia para empezar a pensar que no es cosa de azar.

A un nivel $\alpha = 0'01$ corresponde $Z_C = 2'5758$ con un contraste bidireccional, es decir, si la región de rechazo se establece a ambos lados de la curva. Para hacerlo más sencillo a veces se redondea diciendo que para $\alpha = 0'01$ corresponde una $Z_C = 2'58$

Por otra parte debe tenerse en cuenta que hay tablas de curva normal para todos los gustos. Las hay que acumulan la distribución desde el lado izquierdo, desde el lado derecho, y desde la media, como hemos visto. Y las hay que expresan las cantidades de distribución

acumuladas en proporciones, mientras que otras lo hacen en porcentajes. Lo mismo pasa con las calculadores. Esto puede desorientar al principio, pero una vez que se conoce bien la tabla o la máquina que se está manejando estos detalles no son relevantes.

4.4. DSM: Diferencia Significativa Mínima

Puede determinarse para un test cual es la *Diferencia Significativa Mínima* (DSM) entre puntuaciones a un nivel de confianza elegido, por ejemplo al nivel de confianza del 95%.

Como la z_d es igual a la diferencia entre puntuaciones partido por s_d,

$$z_d = \frac{X_1 - X_2}{s_d}$$

Si hacemos $z_d = z_c$, es decir, si igualamos el quebrado a la z significativa mínima, entonces la diferencia entre puntuaciones será la diferencia entre puntuaciones mínima para obtener significación estadística

$$z_c = \frac{DSM}{s_d}$$

Por tanto, la diferencia significativa mínima DSM será igual a:

$$DSM = z_c \cdot s_d$$

Para ese error típico de la diferencia cualquier diferencia igual o mayor a DSM será significativa; cualquier diferencia menor no lo será.

Puede apreciarse que cuanto mayor es el error típico de la diferencia mayor ha de ser una diferencia entre puntuaciones para ser significativa.

4.5.Ejemplos de comparación de puntuaciones

Vemos ahora algunos *ejemplos* para aclarar el uso de estos conceptos.

Ejemplo 1.

Supongamos que aplicamos a un sujeto un test y se lo volvemos a aplicar a los 15 días. Observamos una diferencia de 3 puntos entre las dos puntuaciones y nos preguntamos si esta diferencia habrá aparecido por azar. Sabemos que el coeficiente de fiabilidad del test es 0'93 y que la desviación típica de las puntuaciones del test es 4'15, de lo que se deduce

enseguida que el error típico de medida del test es 1’098

$$s_e = 4'15\sqrt{1 - 0'93} = 1'098$$

Paso 1.

Por tanto, el error típico de la diferencia es:

$$s_d = \sqrt{s_{e_1}^2 + s_{e_2}^2} = \sqrt{1'098^{\,2} + 1'098^{\,2}} = 1'5528$$

La diferencia de 3 puntos,

$$X_1 - X_2 = 3$$

hay que estandarizarla dividiendo por 1’5528, lo que es igual a:

$$z_d = 3/\,1'5528 = 1'932$$

Es decir, la diferencia de 3 puntos una vez tipificada vale 1’932.

Paso 2.

Para una z de 1’932 el nivel de significación bidireccional es 0’0536.

Explicaciones del paso 2. Este paso pretende contestar la cuestión ¿Cuán probable es que aparezca por azar una diferencia de ese tamaño o superior?

Bajo el supuesto, razonable, de que las diferencias entre puntuaciones se distribuyen según la curva normal, y con un enfoque bidireccional, esta pregunta se traduce operativamente en esta otra ¿que proporción de la distribución normal queda entre $\pm 1'932$ y las colas?

En una tabla de la curva normal vemos que entre una z=1’93 y la media de la distribución queda el 47’32% de la distribución. Por tanto entre $\pm 1'932$ queda un 94.64% aproximadamente. Por tanto entre $\pm 1'932$ y las colas queda 0’0536 en términos de proporciones. El nivel de significación de la diferencia es, por tanto, 0’0536. -Si en lugar de operar con tablas se obtiene el nivel de significación con una calculadora tenemos que es, más exactamente, igual a 0’05336, lo que no afecta la cuestión-

Si el nivel de significación requerido para rechazar la hipótesis nula es como máximo 0’05 entonces, estrictamente no podríamos rechazar la hipótesis nula, es decir, no nos atreveríamos a decir que el resultado no se debe al azar. No obstante, un resultado tan próximo al nivel de significación arbitrario podría considerarse como casi significativo. Algunos textos hablan de “marginalmente significativo”, un lenguaje laxo que suele querer decir que el

estadístico no es significativo, aunque se aproxima mucho al punto de corte que decide la significación estadística. Con rigor estas aproximaciones algo ambiguas no pueden ser bien acogidas.

En este ejemplo el enfoque orientado a obtener la probabilidad $\alpha = 1 - P$ asociada al resultado ha mostrado su interés, al revelar que la diferencia observada de tres puntos aunque estrictamente hablando no es significativa está tan próxima al nivel de significación que casi podría rechazarse la hipótesis nula y aceptar que la diferencia es significativa.

Con un enfoque tradicional orientado a la llamada Z_C o zeta crítica, hubiéramos sabido que, para un nivel $\alpha = 0'05$ bidireccional corresponde una $Z_C = 1'96$, como ese valor no es superado por la z empírica 1'932, la diferencia de 3 puntos observada entre ambas puntuaciones no es significativa o puede atribuirse al azar, sin que el método permita más matiz.

Una observación más, tal como está enunciado el problema no sabemos si la diferencia de 3 puntos es a favor de la segunda o primera medición, lo que en la práctica sería un dato esencial para interpretar su significado psicológico.

Ejemplo 2

Dos personas han contestado un test clásico de inteligencia, cuya media es 100 y cuya desviación típica es 16, con un coeficiente de fiabilidad de 0'90.

La persona A ha obtenido una puntuación de 110 y la B de 97. ¿Podemos decir que hay una diferencia significativa entre estas puntuaciones?

Paso 1.

Como

$$s_E^2 = s_X^2(1 - r_{XX})$$

tenemos que:

$$s_E^2 = 16^2(1 - 0'90) = 25,6$$

El error típico de la diferencia es, pues,

$$s_d = \sqrt{s^2_{e_A} + s^2_{e_B}} = \sqrt{25,6 + 25,6} = 7,155418$$

La diferencia entre puntuaciones es:

$$X_1 - X_2 = 110 - 97 = 13$$

Esta diferencia es la que hay que estandarizar:

$$z_d = 13 / 7,155418 = 1,816805$$

Paso 2.

A una z de 1'816805 le corresponde un nivel de significación de 0'069247, superior al nivel máximo de admisión de diferencias significativas 0'05. Por tanto, no rechazamos la hipótesis nula o la diferencia no es significativa.

DSM.

A un nivel de confianza del 95%, para este test tendremos que:

$$DSM = 1,96 \cdot 7,155418 = 14,024619$$

Para este test hace falta una diferencia de casi una desviación típica para alcanzar significación estadística.

Desde un punto de vista práctico esto significa que no podemos rechazar la hipótesis nula de que una puntuación de 100 y otra de 87, por ejemplo, son iguales y la diferencia aparente podría deberse al azar introducido por el error de medida.

Ejemplo 3

Una persona ha obtenido en el test A de inteligencia una puntuación de 103 puntos, y en el test de inteligencia B una puntuación de 36 puntos.

El test A tiene una media de 100, una desviación típica de 16 y un coeficiente de fiabilidad de 0'916. El test B tiene una media de 35, una desviación típica de 4 y un coeficiente de fiabilidad de 0'87.

Dado que los test están en escalas obviamente distintas (vistas sus medias y desviaciones típicas) hay que proceder primero a convertir en típicas las puntuaciones del sujeto en cada test para que la diferencia entre puntuaciones pueda calcularse.

Paso 0. Conversión de directas a típicas antes de hallar la diferencia.

$$z_A = (103 - 100)/16 = 0,1875$$

$$z_B = (36 - 35)/4 = 0,25$$

Paso 1.

Dado que en típicas, la varianza de los tests vale 1, la varianza de error es igual a:

$$s^2_{Ez} = 1 - r_{XX}$$

Por tanto:

$$s^2_{Ez_A} = 1 - 0,916 = 0'084$$

$$s^2_{Ez_B} = 1 - 0,87 = 0'13$$

El error típico de la diferencia es:

$$s_{dz} = \sqrt{s^2_{Ez_A} + s^2_{Ez_B}} = \sqrt{0,084 + 0,13} = 0'462601$$

Y la diferencia tipificada, por tanto, es:

$$z_d = (0,1875 - 0,25)/0,462601 = -0'135106$$

Que el resultado sea negativo no tiene importancia alguna, significa únicamente que la puntuación tipificada que hemos colocado primero (0,1875) es menor que la segunda (0,25). Si fueran dos medidas con el mismo test o tests paralelos en dos tiempos distintos es usual restar del segundo el primero, aunque es indiferente con tal de que se sepa interpretar la diferencia. Aquí como son dos tests distintos no hay en principio razón alguna para preferir un orden dado.

Paso 2.

El nivel de significación es 0'892528.

Por supuesto hay que interpretarlo en el sentido de que no hay diferencias significativas

5. Comparaciones entre la puntuación y un nivel prefijado

Esta es una cuestión importante que surge en la práctica frecuentemente.

El caso más común consiste en comparar una puntuación -que generalmente expresa un rendimiento- con un nivel de aceptación prefijado. Esto es típico de los tests orientados al criterio. Un ejemplo característico es él las pruebas objetivas de evaluación donde para "aprobar", "pasar la prueba" o "ser admitido" hay que superar determinada puntuación. La comparación en si no presenta mayor

dificultad, aunque la cuestión de como se elabora la prueba, como se puntúa y, particularmente, como se fija el nivel de competencia a superar son temas que requieren consideración a parte.

Un segundo caso se plantea cuando se establecen niveles de competencia a superar o puntos de corte, que representan una decisión distinta para el sujeto, en tests interpretados con una orientación normativa. Por ejemplo, recomendar a los sujetos por debajo del percentil k' determinado tratamiento; por ejemplo admitir a las personas por encima de determinado percentil o puntuación t. Aunque los puntos de corte -niveles de competencia incluidos- son fijados a partir de la distribución normativa de puntuaciones, una vez fijada la puntuación de corte la comparación puede efectuarse entre puntuaciones, como en el caso de los tests orientados a criterio.

La consideración sencilla, en ambos casos, de si una puntuación supera o no, otra determinada como punto de corte, puede realizarse como una simple comparación. Este proceder sencillo suele ser el más usual. En ambos casos, tests orientados al criterio y tests orientados a normas de grupo, la cuestión se complica un poco más si deseamos considerar el error de medida que acompaña a la medición.

Supongamos, por ejemplo, una prueba objetiva corriente formada por 30 items que el sujeto puede acertar o fallar. Supongamos que ha sido administrada a 250 estudiantes, la desviación típica es 4, la media ha sido 16, y el coeficiente de fiabilidad 0'94. El profesor había establecido, como es convencional, el "aprobado" en 15 puntos. Una persona ha obtenido una puntuación de 17 puntos. Una comparación tradicional sencilla dice que 17 está por encima de 15 y por tanto está aprobada.

Ahora bien ¿Podría deberse esa diferencia de dos puntos al error de medida? Esta es una pregunta incómoda que podría intentar resolverse trazando un intervalo confidencial en torno a la puntuación de corte (con idénticos resultados podría resolverse trazando un intervalo confidencial en torno a cualquier resultado concreto). Con los datos anteriores el error típico de medida es 0'9798 y el error máximo, con un nivel de confianza de 95%, es 1'92. Es decir, solo las puntuaciones por debajo de 13'08 y por encima de 16'92 podemos decir que son realmente distintas de 15 con una confianza del 95%. Inmediatamente surgen numerosas dudas que, a mi juicio, cuestionan la utilidad del enfoque de los intervalos confidenciales en este tema. Por ejemplo, ¿qué hacemos con todos los sujetos entre 13'08 y 16'92? Los que están por debajo de 13'08 parece bastante seguro que no están aprobados.

Algún estudiante propondría aquí colocar el punto de aprobado en 13'08, pero, definido ese valor como punto de corte tendríamos el mismo problema,

además de aumentar espectacularmente los falsos positivos (aprobar a quien no supera en realidad el punto de corte 15). Con este instrumento, por cierto "muy fiable" según su coeficiente de fiabilidad, para cada puntuación p podemos dudar razonablemente si el verdadero valor del sujeto está entre p-2 y p+2, aproximadamente. Si bien, a falta de otros datos, la puntuación p es la mejor (y la única) estimación que tenemos de la verdadera posición del sujeto.

En estas condiciones lo más razonable parece seguir siendo aplicar la simple comparación con el punto de corte, desear suerte a los estudiantes y, a ser posible, hacer alguna recuperación con buenos propósitos didácticos. Son posibles otras aproximaciones a esta cuestión, más o menos complejas, pero merecen consideración especial aparte y no entraremos ahora en ellas.

Ejemplos

1. Método general de contraste de puntuaciones individuales

Caso 1. Puntuaciones en escalas distintas

Una persona ha sido evaluada en dos tests que pretenden medir dos aptitudes, distintas pero razonablemente relacionadas, la precisión en el desempeño mecánico manual (Aptitud A) y la rapidez en ese desempeño (Aptitud B), obteniendo en ambas 101 puntos del test. El primer test presenta una media de 95, una desviación típica de 8 y un coeficiente de fiabilidad de 0'8. El segundo test presenta una media de 105, una desviación típica de 7 y un coeficiente de fiabilidad de 0'9. ¿Puede afirmarse que existen diferencias significativas entre ambas puntuaciones a un nivel alfa 0'05?

Datos:

	Test A	Test B
Medias:	$\bar{A} = 95$	$\bar{B} = 105$
Des.Tip.:	$s_a = 8$	$s_b = 7$
Co. Fiab.:	$r_{aa} = 0'8$	$r_{bb} = 0'9$
Punt.:	$A = 101$	$B = 101$

Resultados:

¿Cuándo hay que convertir a puntuaciones típicas las puntuaciones antes de efectuar el contraste?

La primera cuestión ante este tipo de problemas es plantearse si las unidades en que están medidas las puntuaciones son conmensurables y comparables desde la

misma escala. Si las mediciones pertenecen a tests distintos en distinta escala de medida conviene efectuar la tipificación antes de comparar. No siempre es del todo claro si las unidades pueden ser comparadas sin más, y pueden haber algunos casos discutibles, según que criterio se adopte. En general, si las mediciones provienen de tests distintos con distinta media y/o distinta desviación típica conviene tipificar las puntuaciones antes de proceder al contraste.

En este caso no hay duda, las puntuaciones a comparar provienen de tests distintos con distinta media y desviación típica, por tanto, antes de comparar las puntuaciones en el contraste las convertiremos en típicas.

Transformación de las puntuaciones directas a típicas:

$$z_a = \frac{A - \bar{A}}{s_a} \quad \rightarrow \quad z_a = \frac{101 - 95}{8} = 0'75$$

$$z_b = \frac{B - \bar{B}}{s_b} \quad \rightarrow \quad z_b = \frac{101 - 105}{7} = -0'571429$$

Paso 0. Calcular el error típico de la diferencia.

$$s_d = \sqrt{s_{e_a}^2 + s_{e_b}^2}$$

Para calcular el error típico de la diferencia es necesario obtener la varianza de error de cada test. Puede recordarse que la varianza de error de un test no es más que su error típico de medida al cuadrado. En este caso, como vamos a comparar las puntuaciones una vez tipificadas hemos de considerar que la varianza (y la desviación típica) de cualquier variable en típicas vale siempre 1.

$$s_{e_a}^2 = s_a^2 (1 - r_{aa}) \quad \rightarrow \quad s_{e_a}^2 = 1(1 - 0'8) = 0'2$$

$$s_{e_b}^2 = s_b^2 (1 - r_{bb}) \quad \rightarrow \quad s_{e_b}^2 = 1(1 - 0'9) = 0'1$$

$$s_d = \sqrt{s_{e_a}^2 + s_{e_b}^2} \quad \rightarrow \quad s_d = \sqrt{0'2 + 0'1} = 0'547723$$

Paso 1. Tipificar la diferencia entre las puntuaciones

$$z_d = \frac{X_1 - X_2}{s_d} \quad \rightarrow \quad z_d = \frac{0'75 - (-0'571429)}{0'547723} = 2'412586$$

Paso 2. Nivel de significación de la diferencia

Este paso puede darse con distintos enfoque de trabajo que pasamos a comentar.

Enfoque de nivel de significación.

Utilizando una tabla de curva normal (o una calculadora que tenga esta función) (Ver la "Tabla de Nivel de Significación Bidireccional (P) de una puntuación típica (z) en valor absoluto") se determina cual es el área que queda entre las colas y la z_d. (se toman ambas colas para un contraste bidireccional).

En este caso ese valor es 0'015840 (El valor obtenido en una tabla será una aproximación al anterior y puede diferir ligeramente. En adelante redondeamos a 0'016 por simplicidad).

Este valor representa la probabilidad de encontrar una diferencia tan grande o mayor que 2'412586 cuando la hipótesis nula es cierta (es decir, cuando realmente ambas puntuaciones no difieren entre sí y toda variación se debe al error de medición).

Es decir, una diferencia tan grande o mayor que 2'41 todavía aparecería por mero azar, que se atribuye al error de medida, aproximadamente 16 veces de cada mil con tests de estas características métricas.

(El valor 0'016 es la probabilidad de lo que se suele denominar error tipo I: es decir, la probabilidad de rechazar la hipótesis nula cuando en realidad es cierta. El valor 0'016 es la probabilidad de error tipo I asociada a una diferencia z_d de 2'412586).

En resumen, mediante una tabla de curva normal determinamos que a una z_d de 2'412586 le corresponde un nivel de significación de 0'016 (aproximadamente).

Contraste de hipótesis utilizando el nivel de significación

Como se ha pedido una decisión al nivel alfa 0'05 rechazaremos la hipótesis nula siempre que la probabilidad asociada a la diferencia bajo la hipótesis nula sea menor o igual a 0'05.

En este problema, dado que, efectivamente 0'015840 < 0'05, hemos de rechazar la hipótesis nula y considerar la diferencia entre ambas puntuaciones como significativa.

Contraste de hipótesis sin obtener el nivel de significación exacto de la z_d

En este caso, como el problema se ha planteado en términos de efectuar un contraste de hipótesis sin que se haya solicitado el nivel de significación asociado a la z_d, podíamos haber efectuado efectivamente el contraste de hipótesis sin pasar por la obtención del nivel de significación de la z_d por medio de una tabla o de una calculadora.

Simplemente, como sabemos que cualquier puntuación típica mayor (en términos de valores absolutos) que 1'959964 (que suele redondearse por simplicidad a 1'96) deja a ambos lados de la distribución normal menos del 5% de los casos, entonces, en términos prácticos, es suficiente con comparar la z_d obtenida con el valor de la puntuación típica "crítica" (z_c). La puntuación típica crítica (z_c) expresa la menor típica posible (en valor absoluto) que representa una diferencia significativa. En este caso, para nivel alfa 0'05 bidireccional, esa puntuación sera $z_c = 1'96$.

En resumen, comparamos z_d con z_c y si $|z_d| \geq |z_c|$ entonces rechazamos la hipótesis nula y decimos que hay diferencias significativas. (Como en este problema).

Por el contrario, si $|z_d| < |z_c|$ entonces no podemos rechazar la hipótesis nula y decimos que no hay diferencias significativas.

Obsérvese que en aquellos casos en que decimos que "no podemos rechazar la hipótesis nula", se quiere afirmar que las diferencias detectadas no son lo suficientemente grandes (o "raras" en términos de muestreo) para afirmar que están más allá de las que podrían suceder "fácilmente" por azar, pero esto no equivale a establecer la igualdad entre ellas. No debe entenderse que se ha *probado* que las puntuaciones son iguales (ó, en contrastes estadísticos con muestras, que se ha *probado* que en la población no hay diferencias).

Caso 2. Puntuaciones en la misma escala.

Se administra a una persona un test que mide la variable X. Después de unos meses trabajando para mejorar la posición de la persona en esa variable (lo que significar conseguir reducir su puntuación) volvemos a medirla con el mismo test. El test en cuestión tiene una media de 30, una desviación típica de 5 y un coeficiente de fiabilidad de 0'9. La primera vez la persona obtuvo 35 puntos, y la segunda logró reducir su puntuación hasta 30. ¿Ha habido un cambio significativo estadísticamente en la evolución de la persona? (Responded la cuestión utilizando un nivel alfa 0'01 bidireccional).

Datos:

$\bar{X} = 30 \quad s_x = 5 \quad r_{xx} = 0'9 \qquad X_1 = 35 \quad X_2 = 30$

Solución:

Obviamente ambas mediciones están en la misma escala, por lo que no es necesario convertirlas previamente a típicas.

(No obstante, aunque es innecesario, si se convierte a típicas se ha de obtener exactamente el mismo resultado al final del contraste -es decir, se obtendrá justo la misma z_d y por tanto justo la misma probabilidad-. Queda como ejercicio planteado comprobar esta afirmación.)

Paso 0. Calcular el error típico de la diferencia.

Para obtener el error típico de la diferencia es necesario obtener la varianza de error:

$$s_e^2 = s_x^2(1 - r_{xx}) \rightarrow s_e^2 = 25(1 - 0'9) = 2'5$$

$$s_d = \sqrt{s_{e_1}^2 + s_{e_2}^2} \rightarrow s_d = \sqrt{2'5 + 2'5} = 2'236068$$

Paso 1. Tipificar la diferencia entre las puntuaciones

$$z_d = \frac{X_1 - X_2}{s_d} \rightarrow z_d = \frac{35 - 30}{2'236068} = 2'236068$$

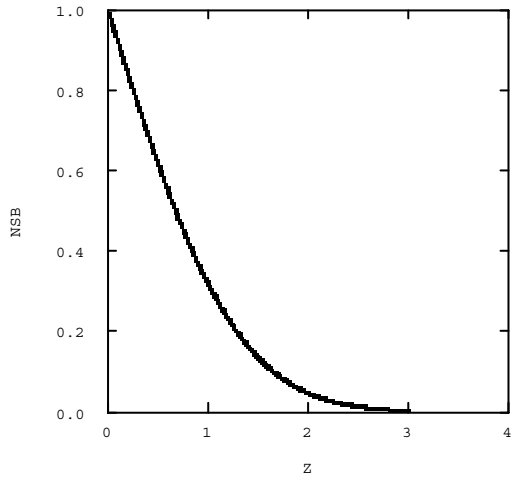
Paso 2. Nivel de significación de la diferencia

Nivel de significación. A la puntuación 2'236068 le corresponde un nivel de significación (alfa) de 0'025347. (Mediante una tabla de curva normal se obtendrá una aproximación a este número).

Contraste. Por tanto, a un nivel alfa 0'01 bidireccional no puede rechazarse la hipótesis nula (dado que la probabilidad obtenida es mayor que 0'01).

Analizado desde el valor de las puntuaciones típicas, la z_d es menor que la $z_c = 2'58$ por lo que no puede rechazarse la hipótesis nula (la diferencia entre las puntuaciones no es significativa a un nivel alfa 0'01 bidireccional).

Gráfico. Nivel de significación bidireccional (NSB) para cada típica (Z)



Caso 3. Diferencia Significativa Mínima (DSM)

He introducido el concepto de DSM -que es una novedad en este campo de las comparaciones individuales bajo teoría clásica-, inspirado por analogía en tests de Fisher para comparar múltiples muestras bajo unas mismas condiciones.

Este estadístico sirve para evitar tener que realizar un contraste de diferencias individuales para cada par de puntuaciones cuando estamos interesados en comparar muchos pares de puntuaciones individualmente (par a par) más que como grupo.

Por ejemplo, supongamos que medimos a una clase de estudiantes de 5º grado con un amplio test de vocabulario orientado a los objetivos de ese grado educativo. Para cada persona efectuamos dos mediciones, una a comienzo de curso y otra al final. El grupo está formado por 30 estudiantes, y estamos interesados en discernir si existen diferencias significativas (alfa 0'01 bidireccional) entre la primera y la segunda medición *de cada uno de los 30 estudiantes individualmente.* El test tiene una media de 97, una deviación típica de 6'8 y un coeficiente de fiabilidad de 0'97. En esta situación habría que hacer 30 contrastes individuales (dado que el foco del problema son las personas individuales y no el grupo como un todo). En lugar de ello tiene sentido preguntarse ¿cuál es la diferencia mínima entre puntuaciones que resulta significativa?

Datos:

$\bar{X} = 97 \quad s_x = 6'8 \quad r_{xx} = 0'97 \quad N = 30$

Respuesta:

La diferencia significativa mínima (DSM) es:

$$DSM = z_c \cdot s_d$$

A su vez:

$$s_e^2 = s_x^2(1 - r_{xx}) \rightarrow s_e^2 = 6'8^2(1 - 0'97) = 1'3872$$

$$s_d = \sqrt{s_{e_1}^2 + s_{e_2}^2} \rightarrow s_d = \sqrt{1'3873 + 1'3872} = 1'665653$$

Por tanto:

$$DSM = z_c \cdot s_d \rightarrow DSM = 2'575829 \cdot 1'665653 = 4'290438$$

Resultado que se interpreta en el sentido de que una diferencia entre dos puntuaciones individuales obtenidas con este test será significativa al nivel alfa 0'01 si es igual o mayor a 4'29 puntos. Esto permite contrastar las diferencias individuales entre las puntuaciones a ese nivel de significación cómodamente.

Comentario:

El problema anterior es una versión simplificada de las dificultades reales. Si estuviéramos interesados en conocer si existen diferencias significativas entre las mediciones del *grupo* antes y después el problema es muy sencillo. Bastaría con tomar los datos de la primera medición y los de la segunda medición y compararlos mediante una prueba t para muestras dependientes (una t de Gosset o una t de Welch, según se asumiera o no igualdad de varianzas entre ambas mediciones). Esto permitiría comparar el comportamiento de los dos grupos de puntuaciones a través de sus medias y decidir si existen o no diferencias significativas.

Sin embargo, tal y como está enfocado el problema, orientados a averiguar para cada persona individual si puede hablarse de cambio significativo en sus puntuaciones no hay más remedio que abordar la comparación desde una perspectiva individual. Esto supone algunas dificultades adicionales, pero realmente estas 30 preguntas individuales son para el psicólogo orientado a informar sobre el curso de la evolución de cada una de estas personas tan o más importantes que la cuestión sobre el grupo como un todo.

Las dificultades prácticas son de diversa índole.

Primero, por supuesto, en la práctica es altamente improbable que ambas mediciones presenten la misma media y la misma desviación típica, y si se calcula el coeficiente de fiabilidad separadamente con los datos de cada una de ellas probablemente también diferiría. Estas diferencias en medias (y desviaciones típicas) son la base que permite que la pregunta sobre la comparación de los grupos tenga sentido pero plantean preguntas para la comparación de puntuaciones individuales Dado que la media y la desviación típica de ambas mediciones no son iguales ¿debemos tipificar las puntuaciones antes del contraste? Puede responderse a esta cuestión de diferentes modos desde diferentes criterios. En primer lugar, si los grupos no difieren estadísticamente pueda obtenerse la media y la desviación típica para las 60 puntuaciones, dado que puede argumentarse que ambas mediciones pertenecen a una misma población de mediciones y dado que una estimación conjunta de media y de varianza puede considerarse en ese caso más adecuada. En este caso no sería necesario tipificar antes de comparar puntuaciones individuales. En segundo lugar, si los grupos difieren significativamente en media (prueba t) o en varianza (test de Levene, p.e.) no puede sostenerse que pertenezcan a una misma población de mediciones y no puede recomendarse obtener una estimación conjunta de media o varianza. En ese caso para cada medición corresponde calcular su varianza de error separadamente y convendría tipificar las puntuaciones antes del contraste individual.

Segundo, un problema inherente a la realización de muchos contrastes entre puntuaciones individuales es que se acumula error tipo I, es decir, la probabilidad de rechazar la hipótesis nula cuando en realidad debería aceptarse. Si, pongamos por caso, para un contraste a nivel alfa 0'05 tenemos que de cada 100 veces que hiciéramos el contraste, en promedio, 5 rechazaríamos la hipótesis nula equivocadamente, entonces, si efectuamos 30 contrastes puede esperarse que 1 ó 2 de ellos presenten diferencias significativas por mero azar, es decir, diferencias que nos llevarían rechazar la hipótesis nula cuando en realidad es cierta. Pueden ensayarse diversas soluciones para esta cuestión aunque pueden resultar discutibles. En primer lugar, si el número de contrastes a realizar puede establecerse de antemano puede pensarse en un procedimiento tipo Bonferroni, dividiendo el nivel de significación por el número de contrastes para garantizar el nivel de significación originario. En la práctica, si el número de contrastes es muy elevado este procedimiento puede ser demasiado exigente con las diferencias para se consideradas significativas, y, desde un punto de vista de las diferencias para un caso individual, aceptar hipótesis nulas que en realidad son falsas, llevando a juicios sobre el comportamiento de las puntuaciones de personas concretas equivocados. En segundo lugar, puede pensarse en "proteger" los contrastes individuales mediante un contraste

estadístico general a nivel de grupos. Esta aproximación puede no ser razonable para algunos de los variados casos en que puede aplicarse el contraste individual de puntuaciones, para empezar simplemente porque no este claro de que grupos se habla en algunos casos. De todas formas aun en los casos en que sea razonable aplicar el método, como en el ejemplo del problema donde se podría aplicar previamente una t para muestras dependientes, es difícil considerar, en términos psicométricos de comparación entre puntuaciones individuales que si no puede rechazarse la hipótesis nula a nivel de grupos necesariamente no pueda rechazarse para un caso individual: un razonamiento así sería una petición de principio que desaconsejaría en cualquier caso efectuar cualquier contraste que no fuera de grupos de puntuaciones.

En síntesis, de la discusión anterior puede concluirse que el contraste estadístico de puntuaciones individuales es procedimiento que debe tomarse con precaución muy particularmente cuando hay que efectuar múltiples contrastes individuales.