3

Error de medida

1. Concepto de error de medida

1. Si efectuamos N mediciones empíricas de una misma dimensión de un mismo objeto con un mismo instrumento de medida, en lugar de obtener N **puntuaciones empíricas X** iguales entre sí, obtendremos una distribución de N puntuaciones.

Las N puntuaciones pueden ser muy semejantes, pero aun en ese caso aparecen pequeñas diferencias si N es suficientemente grande y se observan subunidades suficientemente pequeñas.

Dado que partimos del axioma de que una misma dimensión de un mismo objeto (que no ha sufrido cambios entre medición y medición) ha de medir lo mismo en sucesivas mediciones, el hecho de que al medir N veces una misma dimensión de un mismo objeto aparezca una distribución de puntuaciones provoca perplejidad y lleva inmediatamente a la cuestión de cuál es realmente la medida del objeto en esa dimensión. ¿Qué valor de los obtenidos en la distribución representa lo que "realmente" mide esa dimensión de ese objeto? Por otra parte, ¿a qué pueden atribuirse las variaciones observadas?

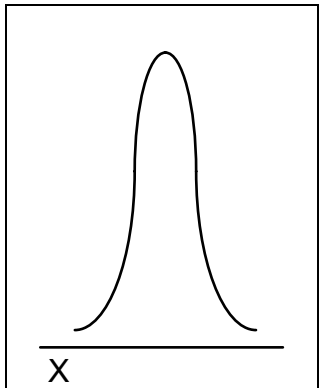

Figura 1. Distribución de puntuaciones empíricas X, obtenidas al medir N veces una misma dimensión de un mismo objeto.

El "modelo lineal clásico", con sus conceptos de puntuación verdadera y de error de medida es ***un*** intento de resolver estas cuestiones.

2. Si la distribución de puntuaciones empíricas obtenida tiene la forma de la figura 1, podemos tomar como la mejor representación de esas puntuaciones la media $\bar{X}$ de todas ellas, que denominaremos **puntuación verdadera V**.

$$\bar{X} = V$$

Parece razonable pensar que N mediciones empíricas de una misma dimensión de un mismo objeto producirán una distribución simétrica y apuntada, como la que representa la Figura 1.

En una distribución de este tipo la media aritmética es una buena representación de los valores obtenidos (Figura 2).

El modelo lineal clásico identifica ese estadístico de posición central con la puntuación "verdadera" V, la verdadera medida del objeto.

Una vez que se admite que la media es la puntuación verdadera hay que admitir que cualquier diferencia entre el valor de una medición concreta X y esa puntuación V no puede ser medición "verdadera".

3. La diferencia entre cualquier puntuación X y la puntuación verdadera V se denomina **error de medida E**.

$$E = X - V$$

4. Una puntuación empírica obtenida de un acto de medición puede considerarse compuesta por la puntuación verdadera V y el error de medida E.

$$X = V + E$$

Esta concepción de la puntuación empírica es lo que se conoce como *modelo lineal aditivo.*

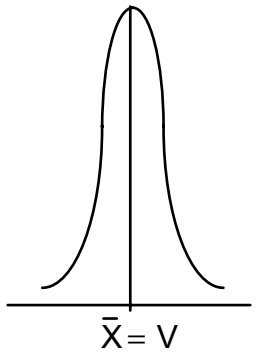

Figura 2. Designación de la media de la distribución de puntuaciones empíricas como puntuación verdadera.

Observaciones:

a. El concepto de error de medida y esta concepción de la puntuación empírica se desprenden de la decisión de considerar la media de las puntuaciones empíricas como puntuación verdadera.

b. Para una medición concreta, tomada al azar entre las N realizadas, el valor y el signo de E puede

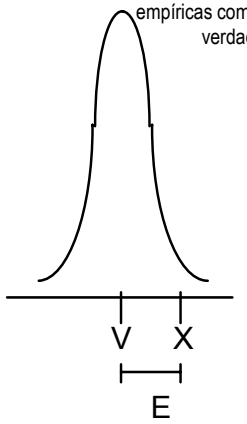

Figura 3. Concepto de Error de medida: La diferencia entre cualquier puntuación empírica X y la puntuación verdadera V.

interpretarse como aleatorio.

Podemos encontrar errores E de diferentes magnitudes, y con signo positivo o negativo.

Si E tiene signo positivo entonces la medición X ha sobrestimado V. Si E tiene signo negativo entonces la medición X ha infraestimado V.

c. Dado que la puntuación verdadera se ha definido como la media de las puntuaciones empíricas, los errores de medida pueden considerarse como *puntuaciones diferenciales:*

$$x = X - \bar{X} = X - V = E$$

d. Por tanto, la distribución de puntuaciones empíricas y la de sus errores de medida son *iguales;* dado que cada puntuación empírica difiere de la verdadera únicamente en su error de medida. (Recuérdese que estamos hablando de la distribución de N puntuaciones obtenidas de medir N veces una misma dimensión de *un* mismo objeto no expuesto a cambio).

e. Por ello, los *estadísticos de dispersión* que obtengamos para la distribución de puntuaciones empíricas serán, en este caso, los mismos que los que obtengamos para la distribución de errores de medida.

f. La media de la distribución de los errores de medida es cero, necesariamente, debido a la definición de la puntuación verdadera.

Dada la definición de puntuación verdadera que lleva a la de error de medida, la media de los errores de medida es la media de un conjunto de puntuaciones diferenciales, cuya suma es necesariamente 0.

$$\bar{E} = \frac{\sum E}{N} = \frac{\sum (X - V)}{N} = \frac{\sum (X - \bar{X})}{N} = 0$$

El último paso se debe a que:

$$\frac{\sum (X - \bar{X})}{N} = \frac{\sum X}{N} - \frac{N\bar{X}}{N} = \frac{\sum X}{N} - \bar{X} = \frac{\sum X}{N} - \frac{\sum X}{N} = 0$$

g. Muchos consideran razonable hacer el supuesto adicional de que la distribución de puntuaciones empíricas de N mediciones de una misma dimensión de un mismo objeto con un mismo instrumento de medida es una distribución normal con media en V.

Este supuesto implica que la distribución de errores de medida de N mediciones de una misma dimensión de un mismo objeto con un mismo

instrumento de medida es una distribución normal, con media 0.

Algunas observaciones críticas adicionales:

a. La definición de puntuación verdadera como media de las puntuaciones empíricas de N mediciones de una misma dimensión de un mismo objeto es el "motor" del modelo. Tomada esa decisión el modelo esta servido.

La definición de puntuación verdadera implica una definición de error de medida y la definición de error de medida implica una definición de puntuación verdadera. Esto es más importante de lo que parece dado que supone que *no puede elaborarse un modelo de medición que considere el error de medida sin que ello implique alguna clase de definición de puntuación verdadera,* se llame como se llame a esa puntuación verdadera.

La definición de puntuación verdadera como media de las empíricas es una definición de puntuación verdadera arbitraria, aunque parece razonable *si* la distribución de puntuaciones cumple una serie de condiciones.

Para que la definición de puntuación verdadera V como media de las empíricas X resulte razonable, la distribución de las empíricas X debe ser:

1. Como mínimo razonablemente simétrica.

La media no representa bien distribuciones asimétricas. Si un instrumento produjera, por ejemplo, una distribución J (con una moda alta y una cola asimétrica hacia un solo lado) la elección de la moda, por ejemplo, podría ser más razonable. No necesariamente todos los instrumentos habrán de producir la misma distribución de errores de medida (y por tanto de puntuaciones empíricas). Es más razonable creer que la forma de esa distribución es una cuestión a establecer empíricamente, pero en Teoría Clásica de Tests se supone *sin contraste empírico* que la media es la mejor representación posible de *cualquier distribución* de puntuaciones que se obtenga con *cualquier instrumento* de medida de *cualquier variable* psicológica. Quizás son muchas suposiciones para tan poca evidencia. El problema se extiende fácilmente a cualquier modelo de medición

que considere error de medida y suponga que la media es la mejor representación de todo conjunto de puntuaciones empíricas de una misma dimensión de un mismo objeto que no varía.

2. Es más razonable si además de simétrica es unimodal.

   La distribución podría ser simétrica pero no necesariamente unimodal. No todas las distribuciones simétricas están bien representadas por una media. Aunque una distribución simétrica inadecuada para ser resumida por una media no parece un comportamiento esperable para las N mediciones de un mismo instrumento, la cuestión es que no hay nada que garantice formalmente a priori que la media sea una buena representación de la distribución, ni siquiera si sólo puede garantizarse que la distribución es simétrica.

3. Es más razonable si, además, la distribución no presenta valores extremos.

   La distribución podría ser simétrica y unimodal y todavía la presencia de valores extremos podría requerir atención en la elección del estadístico de representación. Por ejemplo, en algunos casos una media recortada podría ser más aceptable que una media aritmética.

4. Es más razonable si, además, presenta la mínima dispersión posible (poca desviación típica).

   Sólo una distribución razonablemente concentrada está bien representada por un único valor. Debe tenerse en cuenta que estamos utilizando ese valor como decisión de qué mide el objeto. Si la dispersión es poca ésta puede ser una solución razonable. Pero si la dispersión es muy grande ¿por qué creer que esa es la puntuación verdadera?.

   Obsérvese que si la dispersión es mucha no es sólo que no podemos confiar es cualquier medida X porque puede contener mucho E. El problema es más serio. Si la dispersión es mucha ¿qué razón hay para pensar que la media de las X es V?. Y si no tenemos argumento para afirmar qué valor es V entonces ¿qué error E hay en cada medición X?

5. Es más manejable (para desarrollos posteriores) si, además, es una distribución normal.

   Si la distribución de los errores es una distribución normal entonces disponemos de la ventaja de poder estimar fácilmente la probabilidad de cualquier zona de error (en magnitud y signo) con sólo conocer la desviación típica de esa distribución de errores (dado que la media, en este modelo, es necesariamente 0).

b. Los cinco supuestos anteriores sobre la forma de la distribución de las puntuaciones empíricas (y por tanto sobre la forma de la distribución de los errores) son graduales. Es decir, van desde el supuesto mínimo al máximo. Del menos exigente al más exigente. Del más fácil al más difícil de cumplir.

Los cinco supuestos pueden ponerse a prueba empíricamente si se cumplen dos condiciones:

1. Que exista un objeto con una dimensión susceptible de ser medida N veces *sin que el objeto cambie en esa dimensión*, y

2. Que exista un instrumento de medida susceptible de medir N veces esa dimensión del mismo objeto *sin afectarla y sin resultar afectado*.

c. Si estas dos condiciones operativas no se cumplen, los cinco supuestos anteriores no pueden ser comprobados. Si los cinco supuestos no pueden ser comprobados no sabemos si es o no razonable definir la puntuación verdadera V *como el promedio* de las N mediciones empíricas X de una misma dimensión de un mismo objeto con un mismo instrumento de medida.

d. La definición anterior de puntuación verdadera, "núcleo" de esta teoría, se enfoca sobre la comparación de sucesivas mediciones de una misma dimensión de un mismo objeto que no ha cambiado con un mismo instrumento, pero, precisamente sabemos si el objeto ha cambiado o no mediante la medición de esa dimensión. Esto produce un ciclo paradójico: para definir puntuación verdadera y saber si una medición contiene error de medida hay que disponer de un objeto que no cambie en N mediciones, pero para saber que un objeto no cambia en N mediciones hay que medirlo y constatar que no ha cambiado.

Si medimos N veces una misma dimensión de un mismo objeto con un mismo instrumento, y no hay variación, suponiendo que el instrumento sea sensible,

podemos considerar que el objeto no varía y que el instrumento no introduce variación en el resultado. Pero si obtenemos variación, sin otras fuentes externas de información, no podemos saber con certeza si las variaciones en torno a la puntuación definida como verdadera se deben al instrumento, al objeto medido, o a ambos. Este problema no es despreciable cuando pretendemos medir inobservables psicológicos, especialmente cuando la definición operativa del inobservable la produce el instrumento y no hay otra vía de acceso al inobservable que mediciones de este u otros conceptos inobservables producidas por instrumentos expuestos todos a este mismo problema.

e. El foco podría trasladarse a la comparación entre mediciones efectuadas por diferentes instrumentos de medida que se supone midan lo mismo de igual modo. Este es un enfoque clásico de la cuestión: se suponen N instrumentos paralelos -- mas adelante estudiaremos con algún detalle este concepto-- aplicados sobre un objeto que no varía, ni por sí, ni debido a las mediciones. Esta aproximación facilita otros modos de abordar el problema pero no parece resolver plenamente las dificultades anteriores

f. La teoría razona *como si* el caso de N mediciones de una misma dimensión de un mismo objeto fuera viable en psicología cuando obviamente no lo es.

Esta aproximación "semántica" generó un número de críticas al modelo clásico de teoría de tests, especialmente a lo que se ha denominado la concepción platónica de la puntuación verdadera V. Se han hecho diversos ensayos para superar estas dificultades, especialmente efectuando una definición "sintáctica" de puntuación verdadera que niega estar describiendo un proceso real en el mundo real e instala el concepto como un formalismo matemático. Aunque resuelve formalmente algunas dificultades esta concepción "sintáctica", abanderada por Lord y Novick (1968) y en la que desde entonces se han situado la mayoría de los textos posteriores, necesita inmediatamente una operativización "semántica", si aspira a servir a los datos reales de tests reales, con lo que el hechizo formal ha de desvanecerse unas páginas después. Dicho de otro modo nuestro interés en el modelo es función del grado en que puede describir y operar razonablemente sobre datos reales; las deficiencias formales resueltas por modelos formales en la medida en que se alejan de las operaciones reales de medición resultan desde esta perspectiva secundarias.

g. El modelo define la puntuación verdadera como un determinado estadístico de posición central,

la media. Otras definiciones de puntuación verdadera darían dar lugar a otros modelos. Por ejemplo, con la lógica de minimizar el sumatorio de las distancias absolutas respecto al estadístico podría haberse optado por la mediana, que presentaría otras ventajas y también otros inconvenientes. (La media es un estadístico que minimiza el sumatorio de las distancias al cuadrado; es decir, minimiza la suma de los errores al cuadrado, o suma de los errores cuadráticos).

También es posible pensar que un modelo no orientado a un estadístico de tendencia central podría ser útil. Por ejemplo, si pudiéramos medir N veces la inteligencia de un sujeto en una escala de 1 a 100 obteniendo mediciones en el rango 73 a 78, podría tener sentido definir su inteligencia como la puntuación máxima que ha sido capaz de obtener (especialmente si no puede sostenerse que ese máximo se debe al azar). Esta es la lógica que se sigue en muchas competiciones deportivas, y tiene sentido si lo que se pretende es reflejar el *rendimiento máximo* que el sujeto es capaz de generar. Una lógica de puntuación verdadera como un "máximo" o como un "mínimo" también tendría sentido en otros campos. Por ejemplo, si estamos interesados en establecer la cantidad una de sustancia que provoca a un sujeto una reacción anafiláctica probablemente el mínimo que alguna vez pueda causar ésta es más interesante que el promedio que causa ésta. La puntuación verdadera de N mediciones en este caso quizás sería mejor representada por el mínimo.

El modelo clásico define la puntuación verdadera como el rendimiento medio, pero podría estar igualmente justificado interesarse por otros conceptos de puntuación verdadera en muchas variables psicológicas. Debe notarse que la elección del estadístico no sólo depende de la forma de la distribución. También depende de los propósitos que se persiguen.

El concepto de puntuación verdadera, que usualmente se asume acríticamente, no es neutro respecto al tipo de variable y al propósito para el que se mide.

La cuestión del cuarto elemento o el problema de la concepción del objeto. Un problema que quizá no es exclusivo pero si representativo de la medición psicológica y que afecta a todo modelo de medición, aunque quizá particularmente al modelo clásico, es la concepción del objeto medido.

Todo modelo de medición supone cuatro elementos:

1) Un *sujeto* activo (aquel quien mide), o en su defecto quien diseña, valida, implementa, utiliza o activa el dispositivo de medida.

2) Unas *condiciones* en las que se ha de efectuar la medición, que han de ser estables en sus características relevantes para todo acto de medición comparable.

3) Un *instrumento* de medida, que ha de ser también estable en sus características relevantes entre medidas.

4) Un *objeto* medido.

La medición es un acto que vincula al objeto con el sujeto, aportándole información seleccionada mediante un instrumento de medida en unas condiciones conocidas.

Si pensamos en tests psicológicos los tres primeros elementos se puede considerar que se pueden comportar (no sin esfuerzo) como se supone implícitamente en los modelos de medición. El problema está en el cuarto elemento.

¿Qué cualidades implícitas se esperan del cuarto elemento en los modelos de medición? Del objeto medido, implícitamente, se supone que es objeto *pasivo* de la medición, que *no cambia durante la medición* o debido a la medición, que *no decide activamente resultados* de la medición, que *no adopta estrategias*, que *no hace suposiciones sobre el acto de medición* o sobre el sujeto que mide. En suma que es pasivo, que es un objeto que "se deja medir".

Esto no tiene nada que ver con lo que pasa en psicología. En psicología los sujetos no "*son medidos"* [en voz pasiva] por los tests. Ni siquiera es suficiente decir que los sujetos "*responden"* a los tests. Los sujetos hacen mucho más que eso. Los sujetos piensan, deciden, actúan, interpretan lo que el evaluador quiere y como les afecta…y sobre todo, los sujetos producen, crean, generan activamente la respuesta a los reactivos que se les proponen. La "medición" psicológica es un proceso donde, curiosamente, el evaluado es plenamente activo y el evaluador "pasivo" (en el sentido de que no puede tomar decisiones y está sujeto a una rutina de medición).

La medición psicológica cambia a los sujetos (p.e., aprenden) y los sujetos cambian durante la medición (p.e., se fatigan y también aprenden). La "medición" psicológica implica que

el sujeto crea, produce un producto que no existía antes y que no puede volver a existir después del mismo modo. Nada que ver con el hipotético objeto que se deja medir N veces pasivo e inalterable.

Es sorprendente que no se haya destacado que la medición psicológica implica ante todo un acto de interacción social entre un sujeto y un contexto demandante y otro que se comporta, dentro de ciertas reglas, para satisfacer esas demandas. Hay un rol del "testeur" (quien aplica las pruebas, quien evalúa), pero hay también un rol del sujeto evaluado que este aprende y desempeña, con ciertos grados de libertad.

Lo que observamos en un acto de "medición" psicológica son productos que no estaban antes en el sujeto y que han sido producidos por el sujeto a partir de (pero no necesariamente como "respuesta a" en el sentido mecánico) el test, el cuestionario o la prueba.

En la "medición" psicológica convencional mediante tests y cuestionarios no es cierto que observemos al sujeto. No observamos al sujeto. Tampoco observamos ninguna dimensión o faceta tangible del sujeto. Generalmente ni siquiera observamos la conducta del sujeto, ni la cognitiva, ni la afectiva, ni siquiera la motora. Por lo general sólo observamos ciertos productos de esa conducta. Unos productos restringidos en los que, (¿intencionadamente?) no pueden reflejarse la mayor parte de las conductas cognitivas (¿cómo lo ha hecho? ¿qué decisiones ha tomado? ¿por qué razones? ¿por ejemplo, ha acertado por un conjunto de razones adecuadas o ha acertado por una combinación particular de razones o procesos erróneos como sucede en muchos tests y pruebas con frecuencia desconocida?) ni afectivas (¿qué deseaba? ¿qué sentía y por qué?) ni volitivas (¿a qué aspira el sujeto al actuar así? ¿qué clase de resultados desea obtener en realidad?).

Por supuesto un psicólogo o una psicóloga en un acto de evaluación psicológica sí observa cuidadosamente a la persona y sí trata de obtener el máximo de informaciones útiles acerca de los aspectos que son relevantes para sus objetivos. Lo hace y quizás conviene decir que debe hacerlo. Lo que trato de expresar aquí es que la medición psicológica en sí misma, mediante un test o un cuestionario no hace nada de esto. Deliberadamente ignora toda faceta de la conducta excepto algún aspecto delimitado y

definido, aquel al que se refiere la prueba, por ejemplo, el simple número de respuestas acertadas.

Aunque es el lenguaje común en psicología, para ser más exactos no deberíamos decir que los sujetos "son medidos" por los tests, sería más preciso decir que los sujetos "se comportan", realizan conductas, ante los tests. El problema del objeto proviene del hecho de que no tenemos objetos estables e inertes sino sujetos activos con intenciones, y los instrumentos no "miden" una cualidad del sujeto que se observa, más bien activan al sujeto para que genere conductas de ciertos tipos de los que se considera alguna faceta de su producto.

2. Concepto de error típico de medida.

1. Un instrumento de medida será mejor cuanto menos dispersa sea la distribución de puntuaciones empíricas que produce al medir N veces una misma dimensión de un mismo objeto (siempre supuesto que el objeto no ha cambiado en esa dimensión).

Podemos calcular el grado de dispersión que presenta la distribución de puntuaciones empíricas mediante la varianza o la desviación típica.

2. Varianza de las puntuaciones empíricas o varianza de error:

$$s_x^2 = \frac{\sum(X - \bar{X})^2}{N} = \frac{\sum(X - V)^2}{N} = \frac{\sum E^2}{N} = s_e^2$$

Esta es una definición de varianza de error *para N mediciones empíricas de una misma dimensión de un mismo objeto.*

3. Desviación típica de las puntuaciones empíricas o, también, desviación típica de la distribución de errores de medida:

$$s_x = \sqrt{\frac{\sum(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum(X - V)^2}{N}} = \sqrt{\frac{\sum E^2}{N}} = s_e$$

Se denomina error típico de medida es la desviación típica de los errores de medida. La fórmula anterior es una definición del error típico de medida *para N mediciones empíricas de una misma dimensión de un mismo objeto.*

4. Un instrumento de medida es mejor si produce errores de medida menores. Es decir, si, midiendo una misma dimensión de un mismo objeto N veces, produce una distribución de puntuaciones empíricas poco dispersa. Es

decir, un instrumento de medida es mejor cuanto menor error típico de medida presente.

Si la distribución de los errores de medida es una distribución normal con media 0 y desviación típica s_e, cada magnitud (tamaño) del error tiene una probabilidad conocida.

Hasta aquí hemos descrito algunos conceptos básicos considerando N mediciones empíricas de una misma dimensión de un mismo objeto que no cambia y no se ve alterado por las sucesivas mediciones. Por supuesto se trata solo de una suposición para comprender mejor los conceptos básicos de la Teoría Clásica de Tests.

Este "modelo" resulta inviable para las variables psicológicas de interés de las que se ocupan los tests y otros instrumentos de medida en psicología. Por su propia naturaleza estas variables no pueden ser sometidas a N mediciones, al menos sin esperar que cambien por sí o por efecto de las mediciones mismas. Característicamente las variables de interés psicológico son de carácter "reactivo", es decir, se expresan como respuestas ante cambios estimulares de las condiciones ambientales y del medio social. El modelo supone variables "inertes", estables por sí e indiferentes al proceso de la medición. Ha de observarse que el proceso de medición con instrumentos de medida psicológica clásicos *es explícitamente un cambio de las condiciones ambientales y sociales.* El método clásico de medir en psicología consiste explícitamente en *someter a los sujetos a reactivos de diversa índole.* Aunque esto no puede hacernos concluir que el modelo carece de interés, la situación paradigmática con que se define el modelo (medir N veces una misma variable estable) no es adecuada al menos para variables psicológicas tales como habilidades, aptitudes, personalidad, rendimientos, etc. En general no es adecuada para describir la medición de inobservables psicológicos mediante tests, cuestionarios, etc.

En los apartados siguientes consideraremos el caso de N mediciones sobre N objetos distintos, que describe de modo mucho más realista el tipo de datos de medición de que puede disponer un psicólogo. No obstante el modelo teórico operará para ese caso a partir de las definiciones y suposiciones que acabamos de ver.

Aunque la situación de medir a N personas distintas es realista en términos psicológicos, el modelo de la teoría clásica interpretara cada medición concreta como uno de los posibles resultados obtenidos si hubiéramos podido medir N veces a cada objeto. Esto implica que el modelo no

realista descrito en estos dos primeros apartados subyace a todos los desarrollos posteriores. De hecho, sin este modelo no realista el mismo concepto de puntuación verdadera y error de medida carecen de sentido. Vistas así las cosas los dos conceptos más básicos del modelo clásico, el error de medida y la puntuación verdadera, provienen de un modelo imposible.

3.Indice de Fiabilidad y Coeficiente de Fiabilidad.

1. En los apartados anteriores hemos considerado N mediciones de un mismo objeto, una situación bien poco real en psicología que, no obstante, nos ha ayudado a definir y comprender los conceptos centrales del modelo de puntuación verdadera.

A partir de ahora, consideremos N mediciones efectuadas sobre una misma dimensión de *N objetos diferentes* con un mismo instrumento de medida. (Es decir, cada objeto es medido una sola vez). Una situación que puede considerarse realista en términos de medición psicológica.

Cada objeto tendrá su puntuación empírica X diferente, fruto de la medición. Puede suponerse de acuerdo con el modelo

$$X = V + E$$

que cada puntuación empírica X podría descomponerse en su puntuación verdadera (diferente en cada caso) y su error de medida (diferente en cada caso).

Es decir, tendríamos:

$$\begin{array}{lllll} s_1 & X_1 & = & V_1 & + & E_1 \\ s_2 & X_2 & = & V_2 & + & E_2 \\ s_3 & X_3 & = & V_3 & + & E_3 \\ \vdots & \vdots & & \vdots & & \vdots \\ s_N & X_N & = & V_N & + & E_N \end{array}$$

Las puntuaciones verdaderas y los errores de medida son aquí constructos inobservables, solo las puntuaciones empíricas son observables. Se supone que estas se pueden descomponer aditivamente en aquellas, según el modelo que hemos visto en los apartados anteriores.

2. Cuanto más correlacionen las puntuaciones empíricas con las verdaderas tanto mejor será la medición que proporciona el instrumento de medida.

(El lector quizás se pregunte aquí que cómo podemos averiguar esta correlación dado que las

puntuaciones verdaderas no son observables. Algunas páginas más adelante dedicaremos mucha atención a esta cuestión.)

La correlación r_{xv} entre las puntuaciones empíricas y las verdaderas se denomina **índice de fiabilidad** del instrumento de medida.

3. Si se eleva al cuadrado esa correlación podremos interpretarla en términos de qué proporción de la varianza de las puntuaciones empíricas es explicada por las puntuaciones verdaderas.

Se denomina **coeficiente de fiabilidad** y se representa por r_{xx} al cuadrado del índice de fiabilidad.

$$r_{xx} = r_{xv}^2$$

El coeficiente de fiabilidad por tanto puede interpretarse como la proporción de varianza de las puntuaciones empíricas debida a las verdaderas.

Por ejemplo, si para un test determinado r_{xx} = 0,86 puede decirse que el 86% de la varianza de las puntuaciones empíricas de ese test (las puntuaciones que arroja al medir sujetos) se debe a varianza verdadera (es decir, a varianza de las puntuaciones verdaderas).

Cómo r_{xv} es un coeficiente de correlación de Pearson en teoría su valor puede encontrarse entre -1 (máxima correlación negativa) y +1 (máxima correlación positiva) pasando por 0 (ausencia de relación lineal). Sin embargo, dada la naturaleza de la relación entre puntuaciones verdaderas y puntuaciones empíricas, únicamente adquieren sentido índices de fiabilidad positivos (es decir, cuanto mayor es una puntuación verdadera mayor es la puntuación empírica que aparece). No tendría sentido un r_{xv} negativo que significaría que a mayor puntuación verdadera correspondería una menor puntuación empírica.

Como el coeficiente de fiabilidad es un coeficiente de determinación (es decir, un coeficiente de correlación elevado al cuadrado) puede adoptar valores entre 0 y 1. Cuanto más cercano a 1 sea el coeficiente de fiabilidad (o su raíz cuadrada el índice de fiabilidad) más fiable es el instrumento de medida.

Observaciones:

a. Cuando tenemos N mediciones de N objetos diferentes conocemos sólo las puntuaciones empíricas. Las puntuaciones verdaderas y los errores de medida son inobservables.

b. No sabemos exactamente la cuantía ni el signo de los errores de medida, pero si $N \to \infty$, es razonable suponer que siendo los errores de naturaleza aleatoria se compensarán, de modo que

$$\bar{E} = \frac{\sum E}{N} = 0$$

Es decir, es razonable suponer que el promedio de los errores de medida de N mediciones de la misma dimensión de N objetos distintos es cero. Es decir, es razonable suponer que los errores de medida, en promedio, dada su naturaleza aleatoria, se compensarán.

(Desde luego para muestras concretas de N mediciones, especialmente si N no es muy grande esta afirmación no será estrictamente cierta.)

c. Si los errores de medida son aleatorios no presentarán ninguna variación sistemática en función de V, de modo que parece razonable suponer que:

$$r_{VE} = 0$$

Este supuesto razonable se desprende también de la definición inicial de puntuación verdadera, y puede reexpresarse de diversos modos. Como la correlación entre dos variables es igual a su covarianza partida por el producto de sus desviaciones típicas, tenemos que:

$$r_{VE} = \frac{s_{VE}}{s_V s_E}$$

despejando el término covarianza:

$$s_{VE} = r_{VE} s_V s_E$$

de donde, dado que

$$r_{VE} = 0$$

se desprende que:

$$s_{VE} = 0$$

Como la covarianza entre dos variables es igual al sumatorio del producto de sus puntuaciones diferenciales partido por N, podemos afirmar que:

$$s_{VE} = \frac{\sum ve}{N} = 0$$

Es decir, la covarianza entre las puntuaciones verdaderas y los errores es cero. Esto es una consecuencia como se ve de la aleatoriedad del error.

d. Del mismo modo es razonable suponer, debido a la aleatoriedad del error de medida, que los errores de medida de dos instrumentos de medida aplicados sobre los mismos N casos será también nula:

$$r_{E_1 E_2} = 0$$

e. Otra consecuencia de la aleatoriedad del error es que, la media de las puntuaciones verdaderas (de N objetos distintos medidos en la misma dimensión por un mismo instrumento de medida) es igual a la media de las puntuaciones empíricas. En efecto:

$$\bar{X} = \frac{\sum X}{N} = \frac{\sum (V + E)}{N} = \frac{\sum V}{N} + \frac{\sum E}{N}$$

A su vez, como la media de los errores es 0, tenemos que:

$$\bar{X} = \frac{\sum V}{N} + 0 = \bar{V}$$

f. Por otra parte, si $r_{xx} = r^2_{xv}$ entonces necesariamente:

$$r_{xx} \leq r_{xv}$$

El coeficiente de fiabilidad es menor o igual que el índice de fiabilidad. En realidad podría decirse que el coeficiente de fiabilidad es menor que el índice de fiabilidad excepto en el caso ideal de que el índice y el coeficiente sean igual a 1, cosa que no sucede en la práctica.

Por ejemplo, si el índice de fiabilidad de un test es 0,80, entonces su coeficiente de fiabilidad será 0,80 * 0,80 = 0,64. Al revés, si nos dicen que el coeficiente de fiabilidad de un test es 0,64 y hemos de obtener el índice de fiabilidad calcularemos la raíz cuadrada de 0,64 que es 0,80 dado que:

$$r_{XV} = \sqrt{r_{XX}}$$

Observaciones adicionales:

La aleatoriedad del error es un supuesto interesante que puede analizarse desde diversos puntos de vista. Este modo de razonar permite, como hemos visto, un conjunto de deducciones operativas que contribuyen al desarrollo de la teoría. Sin este supuesto no podrían darse después algunos pasos clave en la teoría de la fiabilidad.

La aleatoriedad del error significa que el tamaño del error no puede describirse como una función de las puntuaciones verdaderas ni de ninguna otra variable. Esto significa, por ejemplo, que el tamaño del error no depende del tamaño de la puntuación verdadera.

Sin embargo, este supuesto puede contradecir nuestra experiencia cotidiana. En efecto, al menos para algunas dimensiones físicas, parece razonable esperar que cuanto mayor es el tamaño de lo medido mayor es el error que se puede cometer al medir. Aunque esto es principalmente un problema de la tecnología de la medición disponible, nadie espera que una medición de distancias entre galaxias presente el resultado con un error de más menos unas pocas micras, aunque si esperamos un error tan pequeño para una buena medición de objetos relativamente pequeños. El tamaño promedio de la variación aleatoria en torno a la puntuación verdadera si parece que depende del tamaño de lo medido.

Sin embargo, si circunscribimos la medición a una escala determinada, es decir, si estamos hablando de la medición de objetos cuya magnitud está dentro de un cierto rango acotado, el supuesto puede tener sentido. No obstante, la noción de qué es un rango suficientemente acotado para que el error no dependa de la magnitud de lo medido debería ser una cuestión empírica a dilucidar empíricamente más que un supuesto.

Desde el modo clásico de considerar este problema podría describirse cualquier variación (error) que puede ser explicada por una función de las puntuaciones verdaderas como *error sistemático intrínseco* y cualquier variación que pueda ser descrita como una función de cualquier otra variable ajena al instrumento como *error sistemático extrínseco.* Además de estos errores sistemáticos todavía podría postularse la existencia del error de medida aleatorio.

Si se introducen estos errores en el modelo este podría representarse así:

$$X_i = V_i + Ei_i + Ex_i + E_i$$

donde E_i simboliza el error sistemático intrínseco y Ex el error sistemático extrínseco. Formalmente E_i es una función f definida sobre las puntuaciones verdaderas

$$Ei_i = f(V_i)$$

y E_x una función g definida sobre otra u otras variables Z ajenas al instrumento

$$Ex_i = g(Z_i)$$

Si el error sistemático es constante a través de los sujetos entonces, de hecho, será indistinguible teóricamente de la puntuación verdadera e irrelevante para el desarrollo de la teoría de la fiabilidad y para una medición orientada a normas de grupo[1]. Aunque no puede decirse que sea irrelevante para una medición orientada a criterios[2].

Si el error sistemático no es constante a través de sujetos y las mediciones empíricas X se analizan exclusivamente en función de V y de E, como en el modelo clásico, el error sistemático podría afectar de forma desconocida y dependiente de cada función específica tanto a las puntuaciones verdaderas como a los errores.

Si se reflexiona el tema con referentes psicológicos el asunto es probablemente más complicado. Es difícil asumir que el modelo no deba incorporar diferentes funciones para diferentes errores sistemáticos intrínsecos y extrínsecos cuya determinación es parcialmente un problema de investigación empírica.

[1] Una **medición orientada a normas de grupo** es aquella en la que la puntuación final del sujeto expresa su ubicación relativa en un grupo normativo. Por ejemplo este es el uso con normas percentiles. La medición sólo indica, al final, una posición relativa. Un **grupo normativo** es una "buena" muestra que se utiliza como punto de referencia y comparación, en la que se calcularon los baremos o equivalencias entre posiciones en la muestra y puntuaciones directas en el test. Después la puntuación directa de cualquier sujeto en el test se interpreta en función de que posición significa en esa muestra.

[2] Una **medición referida a criterios** es aquella que permite conocer el significado de una puntuación directa en el test en términos absolutos de contenido de la variable, sin referencia al desempeño de otros sujetos.

4. Definición del Coeficiente de Fiabilidad como proporción de la varianza empírica debida a varianza verdadera.

Como hemos visto, el coeficiente de fiabilidad es el índice de fiabilidad al cuadrado. Es decir, el coeficiente de fiabilidad es un coeficiente de determinación, y como tal puede interpretarse en el sentido de proporción de la varianza de una variable explicada por otra.

Específicamente puede definirse e interpretarse el **coeficiente de fiabilidad** como la proporción de la varianza empírica (varianza de las puntuaciones empíricas) debida a varianza verdadera (varianza de las puntuaciones verdaderas).

Esto puede mostrarse por varios caminos. Por ejemplo, el índice de fiabilidad, como cualquier coeficiente de correlación puede expresarse como cociente entre la covarianza de las variables y el producto de sus desviaciones típicas:

$$r_{XV} = \frac{s_{XV}}{s_X s_V}$$

A su vez, cualquier covarianza puede expresarse como el promedio del sumatorio del producto de puntuaciones diferenciales:

$$r_{XV} = \frac{s_{XV}}{s_X s_V} = \frac{\frac{\sum xv}{N}}{s_X s_V}$$

A partir de este punto, se sigue que:

$$\frac{\sum xv}{N s_X s_V} = \frac{\sum xv}{N} \cdot \frac{1}{s_X s_V} = \frac{\sum (v+e)v}{N} \cdot \frac{1}{s_X s_V} =$$

$$= \frac{\sum v^2 + ev}{N} \cdot \frac{1}{s_X s_V} = \left(\frac{\sum v^2}{N} + \frac{\sum ve}{N}\right) \cdot \frac{1}{s_X s_V}$$

Dado que:

$$\frac{\sum v^2}{N} = \frac{\sum (V - \bar{V})^2}{N} = s_V^2$$

y, como ya vimos anteriormente,

$$\frac{\sum ve}{N} = \frac{\sum (V - \bar{V})(E - \bar{E})}{N} = s_{VE} = 0$$

podemos simplificar la expresión anterior:

$$(\frac{\sum v^2}{N} + \frac{\sum ve}{N}) \cdot \frac{1}{s_X s_V} = (s_V^2 + 0) \cdot \frac{1}{s_X s_V} =$$

$$= \frac{s_V^2}{s_X s_V} = \frac{s_V}{s_X}$$

En conclusión podemos decir que el **índice de fiabilidad** puede definirse como el cociente entre la desviación típica de las puntuaciones verdaderas y la desviación típica de las puntuaciones empíricas:

$$r_{XV} = \frac{s_V}{s_X}$$

Por tanto, un índice de fiabilidad puede interpretarse en términos de que proporción de la desviación típica de las puntuaciones empíricas se debe a la desviación típica de las puntuaciones verdaderas.

Como el coeficiente de fiabilidad es el índice de fiabilidad al cuadrado,

$$r_{XX} = r_{XV}^2$$

entonces se sigue que:

$$r_{XX} = \frac{s_V^2}{s_X^2}$$

Esta expresión se considera la **definición formal del coeficiente de fiabilidad**.

La definición del coeficiente de fiabilidad como cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas es un punto central de la Teoría Clásica de Tests. Como todas las definiciones y ecuaciones obtenidas hasta aquí esta definición es tautológica y se sigue de la definición inicial de puntuación verdadera.

Se llega a ella por diversos caminos. Por ejemplo:

Dado que, como ya hemos visto, el índice de fiabilidad puede definirse simultáneamente como:

$$r_{XV} = \frac{s_{XV}}{s_X s_V} \qquad r_{XV} = \frac{s_V}{s_X}$$

podemos igualar ambas expresiones, de donde acaba deduciéndose que la covarianza entre empíricas y verdaderas es igual a la varianza de las verdaderas:

$$\frac{s_{XV}}{s_X s_V} = \frac{s_V}{s_X}$$

$$s_{XV} s_X = s_X s_V s_V$$

$$s_{XV} = s_V^2$$

Teniendo en cuenta este resultado, dada la definición del coeficiente de fiabilidad como cuadrado del índice de fiabilidad, tenemos inmediatamente:

$$r_{XX} = r_{XV}^2 = \frac{s_{XV}^2}{s_X^2 s_V^2} = \frac{s_V^2}{s_X^2}$$

5. Descomposición de la varianza empírica en varianza verdadera y varianza de error.

1. Si tenemos N mediciones efectuadas con un mismo instrumento sobre una misma dimensión de N objetos diferentes:

$$X_i = V_i + E_i$$

Donde el subíndice i se refiere a los N casos. La variable X puntuación empírica es la suma de las variables V puntuación verdadera y E error de medida.

Si una variable es suma de otras dos, entonces su varianza es igual a la suma de las varianzas de las variables sumadas más el doble de su covarianza *(ver explicación adicional al final del apartado)* por tanto:

$$s_X^2 = s_V^2 + s_E^2 + 2\, s_{VE}$$

Ahora bien, al ser el error de medida aleatorio no presentará ninguna relación sistemática con las puntuaciones verdaderas pudiendo suponerse razonablemente según el modelo, como ya hemos visto que:

$$s_{VE} = 0$$

Por tanto:

$$s_X^2 = s_V^2 + s_E^2$$

Esta fórmula expresa la descomposición de la **varianza empírica** de un test en **varianza verdadera** y **varianza de error**.

De la expresión anterior se deduce inmediatamente que

$$s_V^2 = s_X^2 - s_E^2$$

2. Lo que permite reexpresar el coeficiente de fiabilidad. Como el coeficiente de fiabilidad es el cociente entre varianza verdadera y la varianza empírica,

$$r_{XX} = \frac{s_V^2}{s_X^2}$$

entonces:

$$r_{XX} = \frac{s_V^2}{s_X^2} = \frac{s_X^2 - s_E^2}{s_X^2} = \frac{s_X^2}{s_X^2} - \frac{s_E^2}{s_X^2}$$

$$r_{XX} = 1 - \frac{s_E^2}{s_X^2}$$

Explicación adicional:

La varianza de una variable X puede expresarse como:

$$s_X^2 = \frac{\sum (X - \bar{X})^2}{N}$$

y, en puntuaciones diferenciales:

$$s_X^2 = \frac{\sum x^2}{N}$$

Si una variable X es la suma de otras dos Y y Z,

$$X = Y + Z$$

entonces su varianza, utilizando puntuaciones diferenciales para mayor sencillez, puede expresarse como:

$$s_X^2 = \frac{\sum x^2}{N} = \frac{\sum (y+z)^2}{N} = \frac{\sum (y^2 + z^2 + 2yz)}{N} = \frac{\sum y^2}{N} + \frac{\sum z^2}{N} + 2\frac{\sum yz}{N}$$

es decir,

$$s_X^2 = s_Y^2 + s_Z^2 + 2s_{YZ}$$

la varianza de una variable suma de otras dos es igual a la suma de las varianzas de éstas más el doble de su covarianza.

6. Definición del Error Típico de Medida a partir del Coeficiente de Fiabilidad.

Se denomina **error típico de medida** a la desviación típica de los errores de medida, es decir:

$$s_E = \sqrt{\frac{\sum (E - \bar{E})^2}{N}} = \sqrt{\frac{\sum E^2}{N}}$$

Esta fórmula, que responde literalmente a la definición teórica anterior es poco operativa, por ello en su lugar se utilizará otra deducida a partir del coeficiente de fiabilidad.

Sabemos que:

$$r_{XX} = \frac{s_V^2}{s_X^2}$$

de donde:

$$s_V^2 = r_{XX} s_X^2$$

Por otra parte hemos visto que:

$$s_X^2 = s_V^2 + s_E^2$$

de donde es inmediato que:

$$s_E^2 = s_X^2 - s_V^2$$

sustituyendo el valor de s_V^2 tenemos:

$$s_E^2 = s_X^2 - r_{XX} s_X^2$$

sacando factor común:

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

de donde:

$$s_E = s_X \sqrt{1 - r_{XX}}$$

que es la nueva expresión del error típico de medida que buscábamos, permite calcular éste conocida la desviación típica de las puntuaciones empíricas y el coeficiente de fiabilidad.

7. Supuesto de Constancia del error típico de medida a través de muestras.

Si se concibe el error típico de medida como una característica del instrumento de medida, parece razonable suponer que un mismo test aplicado a dos muestras distintas mantendrá el mismo error típico de medida:

$$s_{E_1} = s_{E_2}$$

Por tanto:

$$s_{X_1} \sqrt{1 - r_{XX_1}} = s_{X_2} \sqrt{1 - r_{XX_2}}$$

A partir de este supuesto, al igualar las fórmulas del error típico de medida en ambas muestras, es posible, conocido el error típico de medida en la muestra 1, efectuar pronósticos acerca del coeficiente de fiabilidad esperado en la muestra 2 si se cuenta con una estimación de la desviación típica en esa muestra.

Esos pronósticos podrán contrastarse aplicando el test en esa muestra una vez conocidos los procedimientos para la estimación del coeficiente de fiabilidad.

Fue Kelley (1921) quien originariamente sugirió que podría considerarse el error típico de medida como una característica invariante del test, independiente por tanto de la homogeneidad-heterogeneidad de la muestra en que se calcule. Otis (1922) siguió está sugerencia y Gulliksen (1950) la consagró al dedicarle el capítulo 10 de su influyente obra.

No obstante este es un caso claro donde el supuesto puede ser expuesto a contraste empírico. Es decir, debe considerarse hipótesis sujeta a prueba si un test determinado presenta o no el mismo error típico de medida en dos muestras distintas.

CONCEPTOS BASICOS DE LA TEORIA CLASICA DE LA FIABILIDAD

Caso: N mediciones de un mismo objeto.

Modelo:

$\bar{X} = V$

$E = X - V$ $\quad X = V + E$

$\bar{E} = 0$

Caso: N mediciones de N objetos distintos.

Modelo:

$X_i = V_i + E_i \qquad \bar{E} = 0$

Consecuencias:

$r_{VE} = 0 \qquad r_{E_1E_2} = 0$

$\bar{X} = \bar{V} \qquad s_X^2 = s_V^2 + s_E^2$

$r_{XV} = \frac{s_V}{s_X} \qquad r_{XX} = \frac{s_V^2}{s_X^2}$

$s_E = s_X\sqrt{1 - r_{XX}}$

Ejemplo resuelto 1. Fórmulas básicas.

Tenemos un test cuya desviación típica es 6 y cuyo coeficiente de fiabilidad es 0´81.

Se desea calcular el índice de fiabilidad, el error típico de medida, la varianza verdadera, la varianza de error.

Datos:

$s_x = 6$

$r_{xx} = 0'81$

Solución:

Indice de fiabilidad:

$$r_{xv} = \sqrt{r_{xx}} = \sqrt{0'81} = 0'9$$

Error típico de medida:

$$s_e = s_x\sqrt{1 - r_{xx}} = 6\sqrt{1 - 0'81} = 2'6153$$

Varianza Empírica:

$$s_x^2 = 6^2 = 36$$

Varianza de Error:

$$s_e^2 = 2'6153^2 = 6'8398$$

Varianza Verdadera:

$$s_x^2 = s_v^2 + s_e^2 \rightarrow s_v^2 = s_x^2 - s_e^2 = 36 - 6'8398 = 29,1602$$

Coeficiente de Fiabilidad (Comprobación):

$$r_{xx} = \frac{s_v^2}{s_x^2} = \frac{29'1602}{36} = 0'81$$

Ejemplo resuelto 2. Constancia del Error Típico de Medida.

En la muestra 1 el Test X presenta una varianza de 25 y un coeficiente de fiabilidad de 0'9. Se sabe que en la muestra 2 el test presenta una varianza de 36. ¿Cuál será el coeficiente de fiabilidad que se espera en esa segunda muestra si se sostiene el supuesto de constancia del error típico de medida?

Datos:

$s_x^2 = 25 \qquad r_{xx} = 0'9$

Solución:

Desviación típica empírica en la muestra 1:

$$s_x = \sqrt{s_x^2} = \sqrt{25} = 5$$

Error típico de medida en la muestra 1:

$$s_e = s_x\sqrt{1 - r_{xx}} = 5\sqrt{1 - 0'9} = 1'5811$$

Desviación típica en la muestra 2:

$$s_x = \sqrt{s_x^2} = \sqrt{36} = 6$$

Supuesto de constancia del error típico de medida:

$$s_{e_1} = s_{e_2}$$

$$1'5811 = 6\sqrt{1 - r_{xx_2}}$$

Coeficiente de fiabilidad en la segunda muestra

$$1'5811 = 6\sqrt{1 - r_{xx_2}} \rightarrow 0'0694 = 1 - r_{xx_2} \rightarrow r_{xx_2} = 0'9305$$

Comprobación:

$$1'5811 \approx 6\sqrt{1 - 0'9305}$$