# TEMA 5. LA FIABILIDAD

1. Estadísticos de correlación

    1.1. Coeficiente de correlación de Pearson
    1.2. Coeficiente de determinación
    1.3. Otros coeficientes de correlación en función de la naturaleza de las variables

2. Concepto de Fiabilidad

3. Procedimiento para la estimación de la fiabilidad

    3.1. Método de los tests paralelos o de las formas paralelas de un test
    3.2. Método test-retest
    3.3. Consistencia interna de un test
        - Método de las dos mitades
        - Coeficiente α de Cronbach
    3.4. Fiabilidad entre calificadores o evaluadores

4. Factores que afectan a la fiabilidad de un test

5. Factores que afectan a la segunda medición

Bibliografía

# 1. ESTADÍSTICOS DE CORRELACIÓN

## 1.1. Coeficiente de correlación de Pearson

Se basa en el cálculo previo de la **covarianza ($S_{xy}$)**, que es la variación conjunta de dos variables:

$$S_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{N} \qquad \text{o} \qquad S_{xy} = \frac{\Sigma xy}{N} - \bar{x}\bar{y}$$

Si $S_{xy} > 0$ → la relación es positiva
Si $S_{xy} = 0$ → no existe relación
Si $S_{xy} < 0$ → la relación es negativa

El <u>inconveniente</u> de la covarianza es que no tiene límite superior porque depende de las unidades de medida. No sabemos si la relación es intensa o moderada, sólo si es positiva o negativa. Una manera de compensar este problema es a través del **coeficiente de correlación de Pearson**, ya que es independiente de las unidades de medida:

$$r_{xy} = \frac{S_{xy}}{s_x s_y} \qquad \text{donde} \qquad -1 \leq r_{xy} \leq 1$$

Si $r_{xy} > 0$ → la relación es positiva
Si $r_{xy} = 0$ → no existe relación
Si $r_{xy} < 0$ → la relación es negativa

***Factores a tener en cuenta:***

1. Los límites oscilan entre –1 y 1.
2. Clasificación de <u>Guilford</u>:

| <u>Positiva</u> | | | | <u>Negativa</u> |
|---|---|---|---|---|
| $0 \leq r_{xy} < 0'20$ | ↔ | Muy baja | ↔ | $0 \geq r_{xy} > -0'20$ |
| $0'20 \leq r_{xy} < 0'40$ | ↔ | Baja | ↔ | $-0'20 \geq r_{xy} > -0'40$ |
| $0'40 \leq r_{xy} < 0'60$ | ↔ | Media | ↔ | $-0'40 \geq r_{xy} > -0'60$ |
| $0'60 \leq r_{xy} < 0'80$ | ↔ | Alta | ↔ | $-0'60 \geq r_{xy} > -0'80$ |
| $0'80 \leq r_{xy} < 1$ | ↔ | Muy alta | ↔ | $-0'80 \geq r_{xy} > -1$ |

3. Naturaleza de las variables: si trabajamos con variables muy semejantes (ej. dos formas paralelas de un test) las correlaciones deberán ser muy altas, pero si trabajamos con variables de naturaleza muy diversa (ej. habilidad verbal y extraversión) una correlación menor supondrá una relación alta.
4. Coeficientes de correlación que han obtenido con anterioridad otros investigadores.
5. Fiabilidad de los instrumentos de medida, ya que si existe correlación, al aumentar la fiabilidad de los instrumentos, suele también aumentar la correlación.
6. Variabilidad de los datos: cuanto mayor sea la covarianza, mayor será también la correlación de Pearson.

***Condiciones para poder aplicar el coeficiente de correlación de Pearson:***

1. Se debe de tratar de una relación simple, es decir, entre dos variables.
2. La relación entre esas dos variables ha de ser lineal.
3. Las variables tienen que estar, como mínimo, en escala de intervalo y además tienen que ser continuas.
4. La relación entre las variables ha de ser una relación bivariada normal, es decir, ambas variables se distribuyen de forma normal o parecida a la normal.
5. Entre las variables tiene que aparecer la homocedasticidad (variación homogénea): La variación de *y* para cada valor de *x* debe ser constante, y viceversa. La homocedasticidad se da cuando las dos variables son aproximadamente simétricas.

## 1.2. Coeficiente de determinación ( $r^2_{xy}$ )

Es el cuadrado del coeficiente de correlación de Pearson $r^2_{xy}$ ó $\eta^2_{xy}$. Es la proporción de la varianza de una variable asociada a la variación de otra variable.

Ejemplo: x = introversión y = inteligencia

$r_{xy}$ = 0’64 ..... Puntuación alta en x supone puntuación alta en y.

$r^2_{xy}$ = 0’40 ..... El 40% de los datos de x están asociados a los datos y.

## 1.3. Otros coeficientes de correlación en función de la naturaleza de las variables

- ***Estadísticos para variables ordinales: coeficiente ρ de Spearman***

  Es, en realidad una aplicación directa de la fórmula de Pearson pero calculada sobre rangos u órdenes. Su interpretación es igual que la de Pearson, pudiendo oscilar entre –1 y 1.

  La aplicación óptima es cuando tenemos variables ordenadas en rangos. También se puede utilizar con variables en escala de intervalos o razón si los transformamos en órdenes.

- ***Estadísticos para variables nominales: $\chi^2$***

  Para poder aplicarlo no necesitamos tener un número concreto de categorías en las variables, sino que sirve para cualquier número de categorías. El signo siempre será positivo (no nos indica el sentido de la relación) y no tiene límite superior. Para solucionar este problema se utiliza el ***coeficiente de contingencia ( C )*** que sí que tiene límite superior.

- ***Relación entre variables dicotómicas y dicotomizadas:***
  - *Coeficiente biserial puntual ($r_{bp}$):* se utiliza cuando una variable es contínua y permanece contínua, y la otra variable es dicotómica.
  - *Coeficiente Phi ($\varphi$)*: se utiliza cuando las dos variables con las que trabajamos son dicotómicas
  - *Coeficiente biserial ($r_b$)*: se utiliza con dos variables continuas. Una de ellas ( x ) permanece continua y la otra ( y ) la hemos dicotomizado
  - *Coeficiente de correlación tetracórico ($r_t$):* Se aplica con dos variables continuas y distribuidas normalmente que han sido dicotomizadas.

| | | | | | |
|---|---|---|---|---|---|
| **Nominal** | Chi-cuadrado | | | | |
| **Dicotómica** | | Phi | | | Biserial-puntual |
| **Dicotomizada** | | | Tetracórico | | Biserial |
| **Ordinal** | | | | Spearman | |
| **Continua** | | | | | Pearson |
| | **Nominal** | **Dicotómica** | **Dicotomizada** | **Ordinal** | **Continua** |

## 2. CONCEPTO DE FIABILIDAD

Una de las principales características que debe cumplir un test es la de ***Fiabilidad***. La fiabilidad de un test es el grado o la precisión con que el test mide un determinado rasgo psicológico, independientemente del hecho de si es capaz o no de medirlo (validez). Es decir, se dice que un test es fiable cuando "mide bien aquello que está midiendo". Se refiere a la constancia de la medida, al grado en que un instrumento de medida psicológica no deformará el resultado de una medición debido a cambios, fluctuaciones o variaciones del instrumento mismo.

La fiabilidad tiene dos grandes componentes:

- La ***consistencia interna:*** se refiere al grado en que los distintos ítems, partes o piezas de un test miden la misma cosa. Significa la constancia de los ítems para operar sobre un mismo constructo psicológico de un modo análogo.

- La ***estabilidad temporal:*** se refiere al grado en que un instrumento de medida arrojará el mismo resultado en diversas mediciones concretas midiendo un objeto o sujeto que ha permanecido invariable.

Un test totalmente fiable sería aquel con el que se pudiera medir, es decir, situar a un individuo en el baremo sin ningún error. Aunque, en la práctica, ningún instrumento de medida es totalmente fiable, ni siquiera aquellos que miden características físicas. Es decir, si medimos un mismo objeto repetidas veces con el mismo instrumento obtenemos medidas ligeramente diferentes. Por tanto, toda puntuación se compone de la puntuación verdadera más el error cometido, es decir:

$$X = V + E$$

De esta manera, podemos definir la fiabilidad como la proporción de la varianza verdadera de las puntuaciones de un test; lo que significa que la fiabilidad disminuirá a medida que aumente la varianza de error:

$$r_{tt} = 1 - \frac{S_E}{S_V}$$

## 3. PROCEDIMIENTOS PARA LA ESTIMACIÓN DE LA FIABILIDAD

El concepto de fiabilidad se ha definido de manera operativa de diferentes formas:

- Fiabilidad de formas paralelas
- Fiabilidad test-retest
- Fiabilidad de consistencia interna
- Fiabilidad entre calificadores o evaluadores

### 3.1. Método de los tests paralelos o de las formas paralelas de un test

Este método consiste en:

1. Elaborar dos formas paralelas de un mismo test, o lo que es lo mismo, dos tests paralelos.

2. Aplicar una forma del test a la muestra de interés, y tras un lapso de tiempo que no sea relevante para la aparición de cambios en los sujetos, aplicar la segunda forma del test a la muestra.

3. Calcular el coeficiente de correlación entre las puntuaciones empíricas obtenidas por los sujetos en las dos ocasiones. Si las formas son paralelas esa correlación es el *coeficiente de fiabilidad* del test.

***Paso 1: Elaborar formas paralelas***

Hay dos tipos de criterios que dos tests han de cumplir para que los consideremos paralelos:

1. *Criterio estadístico*: Las dos formas presentan medias iguales y varianzas iguales tanto en sus puntuaciones empíricas, como verdaderas y errores (mediciones paralelas) u obtienen las mismas puntuaciones verdaderas, pero no se requiere igual varianza de error (tau-equivalentes).

2. *Criterios de formato y contenido:* En la práctica dos tests paralelos consisten en dos conjuntos distintos de ítems referidos a una misma variable o constructo psicológico, habitualmente con las mismas instrucciones y el mismo formato de prueba y de ítems. Las formas paralelas pretenden muestrear el mismo contenido con cuestiones formuladas de manera distinta.

No puede considerarse formas paralelas aquéllas en las que la diferencia consiste en que se ha variado el orden de los ítems o el orden de las alternativas.

***Paso 2: La aplicación de las formas del test***

1. Las dos formas deben ser administradas bajo las mismas condiciones, o, al menos, bajos los mínimos cambios posibles en las condiciones. Se trata de no introducir factores que puedan provocar cambios en los resultados.

2. Respecto al tiempo, debe utilizarse un lapso entre ambas formas lo suficientemente corto como para que los sujetos no hayan cambiado en la variable de interés y lo suficientemente largo para que factores de memoria, fatiga, o entrenamiento tengan el mínimo efecto.

* *Tipos de tests adecuados para este método*

Es adecuado para tests de potencia y para tests de velocidad en todas las áreas de medición psicológica con instrumentos de lápiz y papel y también, con ciertos tests manipulativos.

***Paso 3: Cálculo del coeficiente de correlación***

Una vez se han administrado las dos formas paralelas se dispondrá de una tabla de datos con N sujetos por 2 variables, la puntuación en la forma A y en la forma B para cada sujeto. Se procede entonces a calcular el coeficiente de correlación de Pearson.

El resultado obtenido puede estar entre –1 y +1, pasando por 0 (ausencia de relación lineal). En realidad, como se trata de formas paralelas, no tiene sentido esperar

correlaciones negativas debiendo estar el resultado entre 0 y +1, incluso cabría esperar valores positivos alejados de 0.

Si A y B son formas paralelas entonces la correlación es el coeficiente de fiabilidad. Para considerar el test fiable, el coeficiente de correlación obtenido deber ser alto, de modo que una gran proporción de la varianza de las puntuaciones se deba a varianza verdadera.

> Es decir, si obtenemos un coeficiente de fiabilidad de 0'75 diremos que tres cuartas partes de la varianza empírica del test se deben a varianza verdadera, o lo que es lo mismo, que un 25% de la varianza empírica es varianza de error.

### 3.2. Método test-retest

Está indicado para estimar la fiabilidad de un test del que sólo disponemos una forma. Consistiría en:

1. Administrar el mismo test en dos ocasiones diferentes separadas por cierto lapso temporal a una misma muestra de sujetos.

2. Calcular el coeficiente de correlación entre las puntuaciones obtenidas por los sujetos en las dos ocasiones.

El método evalúa la estabilidad de los resultados a través de cierto tiempo. Por ello, al coeficiente de fiabilidad que obtiene se le denomina *coeficiente de estabilidad temporal.*

Respecto al tiempo que debe transcurrir:

- A menor tiempo mayor efecto de la memoria de las respuestas dadas, del aprendizaje debido al propio test y de la fatiga producida por el propio test (si la segunda medición sucede de un modo más o menos inmediato).

- A mayor tiempo, mayor posibilidad de que los sujetos hayan cambiado realmente en la variable de interés debido a múltiples factores permanentes o circunstanciales: aprendizaje, cambios evolutivos, experiencias emocionales, enfermedad, condiciones ambientales y sociales, etc.

Por todo esto, las estimaciones por el método test-retest son más apropiadas para tests que miden rasgos poco afectables por los efectos de la práctica y que son estables a lo largo del intervalo de tiempo transcurrido, como son los tests de rapidez perceptiva, discriminación sensorial, verificación rápida de cálculos numéricos, etc.

## 3.3. Consistencia interna de un test

En muchas situaciones no es posible llevar a cabo dos aplicaciones del test. El objetivo aquí, es establecer hasta qué punto se puede generalizar del conjunto específico de ítems al dominio o universo de contenidos. Una forma de llevar a cabo esta estimación es valorando el grado de consistencia con el que los examinados responden los ítems o subconjuntos de ítems del test, en una única aplicación del mismo. Cuando los sujetos tienen un rendimiento consistente en los distintos ítems, decimos que el test tiene *homogeneidad de ítems.* Para que un grupo de ítems sea homogéneo debe medir el mismo constructo o el mismo dominio de contenidos.

### *3.3.1. Métodos de las dos mitades*

***Mediante la fórmula de corrección de Spearman-Brown***

1. Administrar el test a una muestra de sujetos una sola vez.

2. Descomponer el test en dos partes de modo que tengan el mismo número de ítems y que puedan ser consideradas paralelas. Calcular la puntuación total en cada una de estas partes. (Es común comparar la primera mitad del test con la segunda, o comparar los ítems pares con los impares).

3. Obtener la correlación entre las partes. Esa correlación, si las formas son paralelas, podría considerarse la fiabilidad de un test con la mitad de ítems.

4. Aplicar sobre esa correlación la corrección de Spearman-Brown para longitud doble:

$$r_{xx} = \frac{2r}{1+r}$$

Esta corrección estima la correlación que se hubiera obtenido entre las partes si hubiesen tenido el mismo número de ítems que el test completo.

***Mediante la fórmula de Rulon***

1. Administrar el test a una muestra de sujetos una sola vez.

2. Descomponer el test en dos partes de modo que tengan el mismo número de ítems y que puedan ser consideradas paralelas. Calcular la puntuación total en cada una de estas partes.

3. Calcular para cada sujeto la *diferencia entre las puntuaciones* que ha obtenido en las partes: $d = X_1 - X_2$

4. Obtener la varianza del total y la varianza de la nueva variable d. Aplicar la fórmula de Rulon:

$$r_{xx} = 1 - \frac{s_d^2}{S_X^2}$$

***Mediante la fórmula $L_4$ de Guttman***

1. Administrar el test a una muestra de sujetos una sola vez.

2. Descomponer el test en dos partes de modo que tengan el mismo número de ítems y que puedan ser consideradas paralelas. Calcular la puntuación total en cada una de estas partes.

3. Calcular para cada sujeto la varianza que ha obtenido en cada una de las partes así como la varianza total.

4. Aplicar la fórmula $L_4$ de Guttman:

$$r_{xx} = 2\left(1 - \frac{s_1^2 + s_2^2}{s_T^2}\right)$$

La fórmula de Guttman puede considerarse una reexpresión de la fórmula de Rulon, por ello ambas darán el mismo resultado bajo cualquier situación. Ambas, a su vez, equivalen a Spearman-Brown cuando la varianza de las puntuaciones en ambas partes es igual. Si no son iguales, entonces las fórmulas de Rulon y de Guttman darán un valor inferior a la fórmula de Spearman-Brown

### 3.3.2. El coeficiente α de Cronbach

Alfa representa la consistencia interna del test, el grado que todos los ítems del test covarían entre sí. Salvo que tengamos un interés expreso en conocer la consistencia entre dos o más partes de un test (ej. primera mitad y segunda mitad; ítems pares e impares) será preferible calcular el coeficiente α, a aplicar métodos de dos mitades. Éstos únicamente ofrecen información sobre la consistencia entre las partes, mientras que alfa tiene en cuenta la covariación entre cualquier par de ítems.

$$r_{xx} = \frac{n}{n-1}\left(1 - \frac{\Sigma s_i^2}{s_T^2}\right)$$

Donde:
n = número de ítems

$s^2_i$ = varianza de cada ítem

$s^2_T$ = varianza del test total

El coeficiente α oscila entre 0 y 1. Cuanto más próximo esté a 1, los ítems serán más consistentes entre sí. Hay que tener en cuenta que a mayor longitud del test, mayor será alfa.

En el caso de que estemos trabajando con ítems valorados dicotómicamente se utilizarán las fórmulas de Kuder-Richardson (KR –20 y KR -21). Cuando los ítems tienen diferentes índices de dificultad se utiliza la fórmula KR –20. En el caso de que el índice de dificultad sea igual, utilizaremos KR –21.

$$KR-20=\frac{n}{n-1}\left(1-\frac{\Sigma p_i q_i}{s_T^2}\right)$$

$$KR-21=\frac{n}{n-1}\left(1-\frac{x_T-x_T^2/n}{s_T^2}\right)$$

Donde:
n = número de ítems del test

$s^2_T$ = varianza total de las puntuaciones

p = proporción de sujetos que acierta el ítem

q = 1 – p = proporción de sujetos que no aciertan el ítem

$x_T$ = suma de las medias de los ítems. Para ítems dicotómicos: $x_T = n\ p_i$

### 3.4. Fiabilidad entre calificadores o evaluadores

En los tests no estructurados, aunque no exclusivamente en ellos, es necesario determinar si dos o más resultados obtenidos por dos o más evaluadores distintos o por el mismo evaluador en momentos diferentes son coincidentes. En estos casos estaremos hablando de **Fiabilidad intrajuez** o **Fiabilidad interjueces**.

Se calcula a través de un índice de concordancia entre evaluadores, siendo la fórmula más utilizada el índice Kappa:

$$K = \frac{P_0 - P_C}{1 - P_C}$$

Donde:

$P_o$ = proporción de acuerdo observado (suma de los acuerdos conseguidos en cada categoría dividida por el número de registros)

$P_c$ = proporción de acuerdo esperado al azar (suma de la probabilidad de acuerdo por azar de cada categoría).

## 4. FACTORES QUE AFECTAN A LA FIABILIDAD DEL TEST.

1. *Según el método de estimación de la fiabilidad que utilicemos.*
2. *Según las condiciones concretas seleccionadas para aplicar el método:* la fiabilidad variará en función del lapso de tiempo elegido o del número de formas paralelas que apliquemos sobre una muestra.
3. *Características y tamaño de la muestra*: cuanto más homogéneas sean las muestras habrá menos variabilidad y, por tanto, la fiabilidad será menor. En cambio, si las muestras son más heterogéneas, la fiabilidad será mayor.
4. *Longitud del test*: es decir, el número de ítems que presenta el test. Cuanto más largo es un test, mayor es su fiabilidad.

## 5. FACTORES QUE AFECTAN A LA SEGUNDA MEDICIÓN

La segunda medición ha de realizarse en condiciones constantes respecto a las de la primera. Esas condiciones constantes implican ausencia de cambio en los sujetos y ausencia de cambio en las condiciones de administración:

| **Factores que pueden introducir cambios en los sujetos** | **Factores que pueden introducir cambios en las condiciones de administración** |
|---|---|
| - Maduración<br>- Aprendizaje e influencia general debida al medio social<br>- Actividad anterior a la administración de la prueba.<br>- Factores que influencian el estado de ánimo de los sujetos.<br>- Cansancio debido a otras actividades.<br>- Estado de salud de los sujetos.<br>- Fatiga debida a la primera prueba<br>- Memoria de la primera prueba.<br>- Aprendizaje debido a la primera prueba.<br>- Conocimiento de los resultados de la primera prueba. | - El administrador de la prueba.<br>- El local y sus condiciones ambientales.<br>- La hora del día.<br>- El día de la semana.<br>- Sucesos no previstos durante la administración de la pruebas.<br>- Pequeños errores o variaciones en las instrucciones o en los tiempos límite. |

Además, hay otros efectos que son importantes pero que no pueden agruparse fácilmente en estas dos categorías:

1. *Mortalidad experimental* o pérdida de sujetos entre la primera y la segunda medición por las razones que sean.
2. *El fenómeno de regresión a la media:* una persona con una puntuación extrema en la primera medición tenderá a presentar su puntuación en la segunda medición más próxima a la media del grupo.

**BIBLIOGRAFÍA**

Amón, J. (1999). *Estadística para psicológos I. Estadística descriptiva*. Madrid, España: Pirámide.

Botella, J., León, O. G., San Martín, R. y Barriopedro, M. I. (2001). *Análisis de datos en psicología I. Teoría y ejercicios.* Madrid, España: Pirámide.

Meliá, J. L. (1990). *La construcción de la psicometría como ciencia teórica y aplicada.*. Valencia, España: Cristóbal Serrano.

Meliá, J. L. (1993). *Apuntes sobre teoría clásica de tests.* Valencia, España: Cristóbal Serrano.

Navas, M. J. (2002). La fiabilidad como criterio métrico de la calidad global del test. En M. J. Navas (Coord.), *Métodos, diseños y técnicas de investigación psicológica* (pp. 213-261). Madrid, España: UNED.

Pérez Juste, R., García Llamas, J. L., Gil Pascual, J. A. y Galán González, A. (2009). *Estadística aplicada a la educación.* Madrid, España: UNED/Pearson-Prentice Hall.