CONCORDANCIA Y CONSISTENCIA

ÍNDICE

6.1. Concordancia

6.1.0. Introducción

Este módulo aborda las técnicas más utilizadas para el análisis de concordancia entre observadores y para el análisis de consistencia interna de cuestionarios.

Los parámetros para medir concordancia dependen de la escala de medida; en la siguiente tabla se resumen las opciones incluidas en Epidat 4 con sus características generales:

Coeficiente	Escala	Nº de observadores
Kappa de Cohen	Nominal, ordinal	Dos o más
Coeficiente de correlación intraclase	Continua	Dos o más
Bland-Altman	Continua	Dos

Los dos primeros submódulos de concordancia se dedican a las distintas situaciones en que se puede necesitar calcular el coeficiente kappa de Cohen: dos observadores con dos o más categorías de clasificación y tres o más observadores con dos o más categorías de clasificación. El tercer submódulo presenta un método de contraste de hipótesis de igualdad de kappas. El cuarto submódulo incluye la estimación del coeficiente de correlación intraclase (CCI), que es una medida de concordancia entre una serie de mediciones repetidas con resultado continuo, y que podría considerarse un equivalente al índice kappa para variables continuas. El quinto presenta el método de Bland-Altman, propuesto para analizar gráficamente la concordancia entre dos métodos diagnósticos con resultado continuo. Por último, el sexto submódulo se destina a la técnica más común para analizar la consistencia interna de cuestionarios: el coeficiente alfa de Cronbach.

6.1.1. Concordancia entre observadores

6.1.1.0. Conceptos generales

Se dice que un instrumento o procedimiento es *preciso* si sus resultados son consistentes cuando se aplica más de una vez al mismo individuo bajo las mismas circunstancias. La *precisión* de un procedimiento se ve afectada por dos factores fundamentales: la variación propia del instrumento o procedimiento y la variación del examinador. La primera de ellas tiene que ver con la calidad y calibrado del instrumental de medida y diagnóstico, por ejemplo, el nivel de calidad y mantenimiento de un equipo radiográfico. La variación del observador o examinador está relacionada con su entrenamiento, formación y capacidad, y también se llama error del examinador. A menor variación de éste, mayor precisión se consigue a la hora de realizar una prueba y, por tanto, más válido será el resultado.

La variación de un observador respecto de sí mismo, de un estándar (prueba de oro) o de otros observadores, se puede medir por medio de la concordancia alcanzada al examinar y clasificar una serie de elementos (pacientes, radiografías, muestras biológicas, etc.). Por tanto, la precisión o la exactitud de las observaciones pueden evaluarse de varias formas:

1. Comparando un observador consigo mismo para estudiar el grado de concordancia de sus decisiones (concordancia intraobservador). Por ejemplo, entregándole a un radiólogo dos o más veces la misma serie de radiografías para que las clasifique como sospechosas de tuberculosis o libres de sospecha (precisión).

2. Comparando un observador con otro que se toma como referente. Por ejemplo, en el contexto de la concordancia anátomo-clínica, el diagnóstico histopatológico puede considerarse como el referente real, con respecto al cual se evalúa el diagnóstico del clínico (exactitud).
3. Comparando una o varias pruebas diagnósticas con otra que se toma como prueba confirmatoria o de referencia. Por ejemplo, en el tamizaje de pacientes con disfunción renal, puede evaluarse la concordancia de las distintas tiras diagnósticas comerciales, con la medición del índice albúmina/creatinina medido por nefelometría, que se tomaría como prueba confirmatoria (exactitud).
4. Comparando varios observadores entre sí para medir el grado de acuerdo entre ellos (precisión).

En el análisis de concordancia con datos categóricos la situación más sencilla se tiene cuando son dos los observadores que clasifican a un grupo de individuos en dos o más categorías. Los datos se resumen en una tabla de k filas por k columnas siendo k el número de categorías, con la clasificación por ambos observadores. En esta situación, Epidat 4 calcula alguno de los estadísticos que dan información cuantitativa del grado de concordancia en diferentes situaciones, como son el índice de concordancia o acuerdo observado, el kappa de Cohen y los valores mínimo y máximo de kappa. Para hacer más fluida la explicación se referirá al punto 3 (concordancia entre observadores), puesto que no hay diferencias con el 1 o el 2 a efectos de cálculo.

Acuerdo observado (*Índice de concordancia*). Es la primera aproximación a la concordancia entre observadores; resulta, por tanto, la más intuitiva. Simplemente expresa el porcentaje de acuerdo entre ellos, es decir, en qué medida hubo coincidencia en la clasificación entre los observadores en relación al total de elementos examinados.

El problema que plantea este índice básico es que una parte de ese acuerdo, en principio desconocida, puede deberse exclusivamente al azar. Póngase, como ejemplo extremo, que dos ciudadanos, sin ningún tipo de formación especializada, clasifican una serie de sujetos en sanos o sospechosos de enfermedad, en vez de hacerlo dos especialistas en el tema. Indudablemente, los "diagnósticos" serán coincidentes para cierto número de sujetos, pero no debido a la coincidencia de criterios de los observadores, sino simplemente por azar.

¿Cómo se puede cuantificar el grado de acuerdo una vez eliminada la parte que puede atribuirse solamente al azar? Para ello se dispone del kappa elaborado por Cohen en 1960.

Kappa de Cohen. El índice kappa relaciona el acuerdo que exhiben los observadores, más allá del debido al azar, con el acuerdo potencial también más allá del azar. En esencia, el proceso de elaboración del índice es el siguiente: se calcula la diferencia entre la proporción de acuerdo observado y la proporción de acuerdo esperado por azar; si ésta es igual a cero, entonces el grado de acuerdo que se ha observado puede atribuirse enteramente al azar; si la diferencia es positiva, ello indica que el grado de acuerdo es mayor que el que cabría esperar si solo estuviera operando el azar y viceversa: en el caso (ciertamente improbable) en que la diferencia fuera negativa entonces los datos estarían exhibiendo menos acuerdo que el que se espera solo por concepto de azar. Kappa es el cociente entre esa cantidad y el acuerdo máximo que se puede esperar sin intervención del azar. Este índice cumple las características que debe tener una medida de concordancia según Hirji y Rosove [1]: primero, cuando los observadores son independientes, toma el valor 0; en segundo lugar, alcanza el valor

máximo de 1 sólo si hay acuerdo perfecto entre los observadores y, por último, nunca es menor que –1.

La siguiente pregunta que surge sobre el índice kappa es: ¿qué valor de kappa se puede considerar como indicador de buena concordancia? No hay una respuesta exacta; lo que se considera adecuado o no, depende del problema que se esté estudiando. No se espera la misma concordancia entre psiquiatras o psicólogos, entre cuyos pacientes muchas veces es difícil objetivar síntomas, que entre radiólogos de un programa de detección precoz de cáncer de mama, entre los que el grado de acuerdo debería ser elevado. Landis y Koch propusieron en 1977 [2] una escala de interpretación del valor de kappa que considera como *aceptable* un valor mayor o igual a 0,40 y *excelentes* los valores superiores a 0,75.

Kappa mínimo y máximo. El valor máximo de kappa, 1, se da si hay total coincidencia entre los observadores; es decir, se produce cuando el acuerdo observado es de 100% y sólo en esta situación. Sin embargo, puede darse el caso de que el acuerdo observado sea alto y, en cambio, se obtenga un valor de kappa próximo a cero. Suponga tres situaciones en las que el acuerdo observado es del 90%:

Situación A

		Observador 1	
		+	-
Observador 2	+	1	4
	-	6	89

Situación B

		Observador 1	
		+	-
+		89	8
-		2	1

Situación C

		Observador 1	
		+	-
+		55	10
-		0	35

La situación A se da cuando la prevalencia del fenómeno es baja entre los sujetos observados; en esos casos, el número de verdaderos "negativos" es elevado y es más alta la probabilidad de que los dos observadores clasifiquen a los sujetos como tal, de modo que la coincidencia atribuible al azar será mayor. En consecuencia, dado que kappa elimina la influencia del azar, se obtendrán valores bajos de dicho coeficiente (kappa=0,115). Lo mismo sucede si la prevalencia es alta, como en el ejemplo B (kappa=0,127). En situaciones intermedias, por ejemplo la C, la distribución de los acuerdos es más simétrica y se obtienen valores de kappa más altos (kappa=0,794). Es decir, la "paradoja" de valores altos de acuerdo observado asociados a valores bajos de kappa, descrita por Feinstein y Ciccheti [3], se explica porque, para un valor fijo del acuerdo observado, la magnitud de kappa depende de la prevalencia del fenómeno estudiado. Indudablemente, esta circunstancia configura un defecto del coeficiente que se está tratando. En particular, las comparaciones entre coeficientes kappa estimados en poblaciones con prevalencias muy diferentes pueden resultar conflictivas. Esto afecta también a la pertinencia de las calificaciones sugeridas por Landis y Koch [2]. Entre las soluciones propuestas para este problema está la de Lantz y Nebenzahl [4], quienes sugieren que el índice kappa se acompañe con los estadísticos kappa mínimo y máximo, que corresponden, respectivamente, a los valores mínimo y máximo de kappa para un nivel dado de acuerdo observado. Epidat presenta estos valores en el caso de dos observadores con dos categorías de clasificación.

Tipos de aplicaciones de kappa. La aplicación más simple y común del análisis de concordancia se da en el caso de dos observadores y dos categorías de clasificación. Sin embargo, kappa se puede calcular en situaciones más complejas, como cuando son dos los observadores pero tres o más las categorías de clasificación, cuando son tres o más los observadores y dos las categorías, e incluso el caso más general en que son tres o más, tanto los observadores como las categorías posibles de clasificación. Cuando las categorías son más de dos se puede calcular el coeficiente para conocer el grado de acuerdo entre los observadores en cada una

de las categorías, de manera independiente. Para estudiar más profundamente las diversas opciones citadas, se sugiere acudir al capítulo 13 del texto de Fleiss [5].

El índice kappa puede utilizarse para medir la concordancia entre observadores (sin tomar en cuenta la validez o exactitud de la medición) o la concordancia de uno o más observadores con un referente que representa un valor real o una prueba de oro.. Aunque el desarrollo inicial del estadístico kappa estuvo dirigido a la medición del acuerdo entre observadores, en realidad tiene utilidad para medir, en datos categóricos, otros aspectos como "similitud" o "agrupamiento"; por ejemplo, cuando se quiere determinar el grado de similitud entre controles emparejados en un estudio de casos y controles [5].

Kappa ponderado. Supóngase que las categorías de clasificación son más de dos y están definidas en una escala ordinal, como por ejemplo "sano", "posiblemente enfermo" y "claramente enfermo". A la hora de valorar el grado de discrepancia entre dos observadores, no es lo mismo que uno clasifique a un sujeto como "posiblemente enfermo" y el otro lo declare "sano" a que uno lo clasifique como "sano" y el otro como "claramente enfermo". La "distancia" entre ambas discrepancias no es la misma.

Cuando ciertos tipos de desacuerdo tienen mayor relevancia o peores consecuencias para la calidad de la medición que otros, al investigador le puede interesar tenerlo en cuenta en la construcción del índice kappa. Para ello, se ha sugerido ponderar las diferentes discrepancias, usando una matriz de pesos que pueden variar según el criterio del investigador en función de lo que esté analizando, aunque siempre cumpliendo ciertas restricciones, que por otro lado son bastante intuitivas: puesto que kappa no hace distinción entre los dos observadores, la matriz debe ser simétrica; además, a la diagonal de acuerdos se le asigna el máximo peso, que es 1 y el resto de pesos deben ser inferiores, aunque siempre positivos o iguales a cero.

Epidat 4 incorpora, además de la opción manual, los dos tipos de ponderación más comunes: los pesos cuadráticos y los de Cicchetti [5]. Ambos se basan en las distancias relativas entre las categorías de clasificación, con la única diferencia de que los primeros utilizan diferencias al cuadrado y los otros operan con diferencias en valor absoluto, de modo que los pesos cuadráticos tienden a dar una ponderación mayor a los desacuerdos. La opción manual permite introducir desde el teclado los pesos que desee el usuario con las restricciones mencionadas anteriormente.

6.1.1.1. Limitaciones del estadístico kappa

- El valor de kappa se ve afectado por la prevalencia del rasgo estudiado. Por tanto, es necesario ser cuidadoso a la hora de generalizar los resultados de comparación de observadores en situaciones con prevalencias diferentes; esto quiere decir que kappa es un estadístico descriptivo útil, pero es inadecuado con fines de predicción o inferencia [6].
- Kappa es dependiente del número de categorías. Cuantas más categorías se estén considerando, más difícil será clasificar correctamente los sujetos de observación, lo que habitualmente implica valores de kappa más bajos [7]. Por tanto, debe tenerse en cuenta el número de categorías a la hora de interpretar kappa.
- Para datos ordinales derivados de categorizar variables continuas, el valor de kappa depende fuertemente de las a menudo arbitrarias definiciones que se hacen de las categorías.

- El uso de la ponderación, aunque lógico y atractivo, introduce otro componente de subjetividad.

6.1.1.2. Recomendaciones

- Es insuficiente presentar un único coeficiente o índice; se recomienda la presentación de los datos [7].
- Es aconsejable presentar, junto al índice kappa, sus valores mínimo y máximo, tal como sugieren Lantz y Nebenzahl [4].
- Es necesario recordar que la interpretación de kappa no puede disociarse de su aplicación particular. *Ceteris paribus* un mismo valor de kappa puede considerarse alto en unas circunstancias y bajo en otras.
- Las soluciones para los desacuerdos inter e intra observadores deben buscarse en la estandarización de las mediciones y las reuniones de consenso sobre observaciones clínicas. El conocimiento sobre el origen de los errores ayuda en este proceso. Si la concordancia no puede aumentarse con estas estrategias, la solución puede conseguirse a través de las medidas múltiples. Dependiendo de cual sea la principal fuente de desacuerdo, las medidas deben realizarse por diferentes o por el mismo observador [7].
- Hay otras medidas de concordancia y/o asociación [1][8], que se incluyen en el módulo "Tablas de contingencia" de Epidat 4, como son la τ de Kendall y la γ de Goodman y Kruskal para datos ordinales.
- Debe tenerse en cuenta que un valor de kappa significativo no puede interpretarse como expresión de buena concordancia.

Ejemplos

Ejemplo A. Suponga que a dos radiólogos del programa de tuberculosis se les remiten radiografías de tórax de 170 sujetos que están controlados en una unidad de neumología, y que se quiere estimar el grado de concordancia entre ellos. Los radiólogos A y B tienen que clasificar cada radiografía en una de dos categorías: "positiva" (sospechosa de lesión tuberculosa) o "negativa" (no sospechosa de lesión tuberculosa). Los resultados se muestran a continuación:

		Radiólogo A	
		+	**-**
Radiólogo B	**+**	58	39
	-	12	61

Calcule la concordancia bruta entre los radiólogos y el kappa de Cohen, con sus intervalos de confianza.

Ejemplo B. (Modificado de: Banauch D, Koller PU, Bablok W. Evaluation of Diabur-Test 5000: a cooperative study carried out at 12 diabetes centers. *Diabetes Care* 1983; 6(3): 213-18.)

Se desea estimar la concordancia entre dos pruebas diagnósticas de diabetes, el Diabur-Test 5000® y el Clinitest® en 1.677 muestras de orina. Los valores obtenidos se muestran en la tabla. Calcule el valor de kappa sin ponderar y ponderado por pesos cuadráticos.

		Diabur-Test 5000®					
		Negativo	**Trazas**	**1**	**2**	**3**	**5**
Clinitest®	**Negativo**	452	5				
	Trazas	133	270	28	1	2	
	1	4	36	107	5	2	2
	2		5	53	76	28	4
	3			12	28	81	35
	5			2	11	44	251

Ejemplo C. Suponga ahora que una selección de 25 radiografías correspondientes a otros tantos pacientes del servicio de neumología del ejemplo A, se entregan a un grupo de radiólogos para que las clasifiquen de forma independiente. Por diversas razones, no todos los radiólogos del equipo pudieron emitir juicio sobre todas las radiografías. Los resultados se presentan en la siguiente tabla:

Paciente	Número de radiólogos	Número de positivos
1	4	3
2	3	2
3	4	2
4	5	4
5	3	3
6	4	2
7	4	3
8	5	3
9	5	4
10	5	5
11	3	0
12	2	0
13	4	2

Paciente	Número de radiólogos	Número de positivos
14	4	0
15	3	2
16	5	5
17	5	0
18	3	2
19	4	3
20	4	2
21	3	1
22	2	0
23	5	0
24	4	4
25	4	3

Analice la concordancia entre los radiólogos calculando el valor de kappa y su intervalo de confianza.

Ejemplo D. A usted le interesa afinar más su análisis de concordancia entre los radiólogos del ejemplo anterior y selecciona un subgrupo de 15 pacientes cuyas placas entrega a los 5 radiólogos del equipo para que las clasifiquen, de forma independiente, en una de las siguientes clases: "muy sospechosas de lesión tuberculosa" (Categoría 1), "sospecha ligera de lesión tuberculosa" (Categoría 2) o "sin sospecha alguna de lesión tuberculosa" (Categoría 3). Los resultados se muestran en la siguiente tabla:

Número de radiólogos que clasificaron en cada categoría

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

Calcule ahora el valor de kappa.

6.1.1.3. Manejo del submódulo de concordancia entre dos observadores con dos o más categorías de observación y solución a los ejemplos

El submódulo de análisis de concordancia de dos observadores con dos o más categorías de clasificación sólo permite la entrada de datos desde el teclado. Note que, cuando se seleccionan más de dos categorías, se activa la opción de ponderar, por si se desea obtener kappa ponderado. Epidat 4 da la opción de utilizar pesos cuadráticos, pesos de Cicchetti o definir los pesos que el usuario considere oportunos; para visualizar los valores de los pesos se debe seleccionar la pestaña "Ponderaciones", una vez se haya activado la opción de ponderar.

Resultados del Ejemplo A con Epidat 4:

Datos:

	1	2
1	58	39
2	12	61

Nivel de confianza: 95,0%
Número de categorías: 2
Tipo de ponderación: No ponderar

Resultados:

Acuerdo observado: 0,7000
Acuerdo esperado: 0,4875

Kappa	EE*	IC (95,0%)	
0,4146	0,0655	0,2862	0,5430

*EE: error estándar

Kappa mínimo: -0,1765
Kappa máximo: 0,4495

Prueba de significación:

Estadístico Z	Valor p
5,6855	0,0000

Resultados del Ejemplo B con Epidat 4 sin ponderar:

Datos:

La tabla de datos se omite

Nivel de confianza:	95,0%
Número de categorías:	6
Tipo de ponderación:	No ponderar

Resultados:

Acuerdo observado:	0,7376
Acuerdo esperado:	0,2035

Kappa	EE*	IC (95,0%)	
0,6706	0,0130	0,6450	0,6961

*EE: error estándar

Prueba de significación:

Estadístico Z	Valor p
57,0987	0,0000

Resultados del Ejemplo B con Epidat 4 ponderado por pesos cuadráticos:

Datos:

La tabla de datos se omite

Nivel de confianza:	95,0%
Número de categorías:	6
Tipo de ponderación:	Pesos cuadráticos

Resultados:

Acuerdo observado:	0,9856
Acuerdo esperado:	0,7165

Kappa	EE*	IC (95,0%)	
0,9491	0,0033	0,9427	0,9555

*EE: error estándar

Prueba de significación:

Estadístico Z	Valor p
38,9823	0,0000

6.1.1.4. Manejo del submódulo de concordancia entre tres o más observadores con dos o más categorías de observación y solución a los ejemplos

Tres o más observadores y dos categorías. Esta opción sirve para analizar la concordancia en el caso de dos categorías de observación (una variable dicotómica, en general "Positivo-Negativo"), y tres o más observadores. Hay que tener en cuenta que el número de observadores no tiene porqué ser igual en todos los sujetos que se deben clasificar; es decir, a un sujeto pueden, por ejemplo, clasificarlo cuatro observadores y a otro, solo tres.

Epidat 4 calcula el intervalo de confianza para kappa aplicando la técnica jackknife [9], que permite estimar el error estándar en situaciones de cierta complejidad, como es el caso del coeficiente kappa con múltiples observadores. Este método presenta ciertas ventajas sobre los procedimientos tradicionales: es una técnica sencilla y aplicable en muchos problemas, sin necesidad de hacer hipótesis sobre la distribución de la población. Aunque se han desarrollado métodos como el bootstrap, que mejoran la eficiencia del jackknife, éste continúa siendo una técnica útil y muy usada.

Al optar por la entrada automática, se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que se importen tengan una estructura fija (véase Tabla 1). En este caso, la tabla debe tener tantas filas como sujetos clasificados, y dos variables: una que contenga el número de observadores que clasificaron a cada sujeto y otra, el número de clasificaciones "positivas".

Tabla 1. Formato de tabla preparada para importar datos desde Epidat 4 para el análisis de concordancia entre tres o más observadores y dos categorías de clasificación.

"Sujeto"	"Nº de observadores"	"Clasificaciones"
SUJETO	**RADIOLOGOS**	**CLASIFIC**
1	4	3
2	4	4
3	2	2
4	3	3
...	...	...

Los datos de la Tabla 1 se encuentran en la hoja *2 categorías* del archivo RADIOLOGÍAS-TB.XLS incluido en Epidat 4.

Resultados del Ejemplo C con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Concordancia y consistencia\RADIOLOGÍAS-TB.xls
Tabla: 2 categorías
Variables:
Número de observadores: RADIOLOGOS
Clasificaciones: CLASIFIC

Datos:

Número de sujetos:	25
Nivel de confianza:	95,0%

Resultados:

Kappa	IC (95,0%)	
0,2947	0,0126	0,5753

Intervalo de confianza jackknife.

Prueba de significación:

Estadístico Z	Valor p
3,5255	0,0004

Tres o más observadores y tres o más categorías. Esta opción sirve para analizar la concordancia cuando, tanto las categorías de observación como los observadores, son tres o más. A diferencia de la opción anterior, todos los sujetos han de ser clasificados por el mismo número de observadores.

Epidat 4 calcula el intervalo de confianza para kappa aplicando la técnica jackknife [9], que permite estimar el error estándar en situaciones de cierta complejidad, como es el caso del coeficiente kappa con múltiples observadores. Este método presenta ciertas ventajas sobre los métodos tradicionales: es una técnica sencilla y aplicable en muchos problemas, sin necesidad de hacer hipótesis sobre la distribución de la población. Aunque se han desarrollado métodos como el bootstrap, que mejoran la eficiencia del jackknife, éste continúa siendo una técnica útil y muy usada.

Al optar por la entrada automática se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija (véase Tabla 2). En este caso, la tabla debe tener tantas filas como sujetos clasificados, y una variable por cada categoría de clasificación, hasta un máximo de 100, con el número de observadores que clasificaron a cada individuo en dicha categoría, también 100 como máximo. La suma de todas las clasificaciones de un sujeto debe ser igual al número total de observadores.

Tabla 2. Formato de tabla preparada para importar datos desde Epidat 4 para el análisis de concordancia entre tres o más observadores con tres o más categorías de clasificación.

"Sujeto" "Categoría 1" "Categoría 2" "Categoría 3"

SUJETO	CATEG1	CATEG2	CATEG3
1	2	2	1
2	5	0	0
3	0	1	4
4	1	1	3
5	4	1	0
...	...	...	...

Note que la suma de las clasificaciones en cada sujeto es siempre la misma, en este caso 5.

Los datos de la Tabla 2 se encuentran en la hoja *3 categorías* del archivo RADIOLOGÍAS-TB.XLS incluido en Epidat 4.

Resultados del Ejemplo D con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Concordancia y consistencia\RADIOLOGÍAS-TB.xls
Tabla: 3 categorías
Variables:
Categorías: CATEG1, CATEG2, CATEG3

Datos:

Número de sujetos: 15
Número de categorías: 3
Número de observadores: 5
Nivel de confianza: 95,0%

Resultados:

Categoría	Kappa	IC (95,0%)		Estadístico Z	Valor p
CATEG1	0,3100	-0,0147	0,6303	3,7967	0,0001
CATEG2	0,1136	-0,1273	0,3512	1,3918	0,1640
CATEG3	0,3889	0,1366	0,6378	4,7629	0,0000
Kappa global	0,2804	0,0741	0,4836	4,8234	0,0000

Intervalo de confianza jackknife.

6.1.2. Comparación de índices kappa

6.1.2.0. Conceptos generales

Habitualmente los estudios de concordancia se repiten como parte de la evaluación de un programa; tal es el caso, por ejemplo, del control de calidad de un programa de detección precoz de cáncer de mama. Ante una serie de índices kappa obtenidos en diferentes estudios y momentos, surge la pregunta: ¿Son diferentes estos valores entre sí? Epidat 4 tiene una opción para contrastar la hipótesis de igualdad de kappas [5]. Para dicha prueba de hipótesis se precisan los valores de kappa obtenidos y sus correspondientes errores estándar. Como se recordará, el error estándar está directamente relacionado con el número de sujetos examinados y es útil para determinar el intervalo de confianza del coeficiente. Epidat 4 también obtiene un valor de kappa global, que resume los que se están comparando.

Ejemplo

De usted dependen dos unidades de diagnóstico de tuberculosis (A y B). Cada seis meses remite a cada una de ellas un porcentaje de las tinciones realizadas en el semestre anterior con el objeto de comprobar la concordancia entre ellas. A las dos unidades les envía las mismas muestras, que tienen que ser clasificadas como "positivas" o "negativas". Los valores obtenidos en los dos últimos años se exponen a continuación:

Semestre 1

		Unidad A	
		+	-
Unidad B	+	350	120
	-	70	550

Semestre 2

		Unidad A	
		+	-
Unidad B	+	280	80
	-	60	550

Semestre 3

		Unidad A	
		+	-
Unidad B	+	320	30
	-	120	29

Semestre 4

		Unidad A	
		+	-
Unidad B	+	890	210
	-	290	700

Calcule el kappa global. Con un nivel de confianza del 95%, ¿son diferentes los kappas obtenidos en los diferentes semestres?

6.1.2.1. Manejo del submódulo de comparación de kappas y solución al ejemplo

Este submódulo sirve para contrastar la hipótesis de igualdad de kappas. Para ello es necesario disponer del valor de cada uno de los kappas que han de compararse y de sus respectivos errores estándar. En la salida se presentan una estimación global de kappa con su intervalo de confianza, y una prueba de homogeneidad que contrasta la hipótesis nula de que todos los coeficientes kappa que se comparan son iguales.

Al optar por la entrada automática se abre el asistente para la obtención de datos que permite, a través del botón “examinar”, seleccionar el directorio y el archivo (OpenOffice o

Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija (véase Tabla 3). En este caso, la tabla debe contener tantas filas como número de kappas a comparar, hasta un máximo de 100, y dos variables, una con el valor de kappa y otra con su error estándar.

Tabla 3. Formato de tabla preparada para importar datos desde Epidat 4 para el análisis de la comparación de kappas.

"Kappas"	"Kappa estimado"	"Error estándar"
	KAPPA	**ERROR**
1	0,640	0,024
2	0,687	0,024
3	0,132	0,043
4	0,518	0,019

Para resolver el ejemplo es preciso calcular el kappa y el error estándar de cada tabla, para lo cual puede emplearse el primer submódulo de concordancia. Además, los datos de la Tabla 3 se encuentran en el archivo SEMESTRES-TB-XLS incluido en Epidat 4.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Concordancia y consistencia\SEMESTRES-TB.xls
Tabla: Datos
Variables:
Kappa estimado: KAPPA
Error estándar: ERROR

Datos:

Nº de kappas a comparar: 4
Nivel de confianza: 95,0%

Resultados:

Kappa global	IC (95,0%)	
0,5617	0,5379	0,5855

Prueba de homogeneidad de kappas:

Ji-cuadrado	gl*	Valor p
143,0515	3	0,0000

*gl: grados de libertad

6.1.3. Coeficiente de correlación intraclase

6.1.3.0. Conceptos generales

Aunque es muy frecuente en la práctica clínica la necesidad de valorar la concordancia entre medidas cuantitativas, no es raro que se apliquen métodos erróneos con este objetivo, por ejemplo, el coeficiente de correlación de Pearson o un modelo de regresión lineal. Sin embargo, el coeficiente de correlación de Pearson solo cuantifica la asociación lineal entre dos variables, pero no el grado de acuerdo entre ellas, además de tener otras limitaciones [10][11]. Una alternativa adecuada para este propósito es el llamado coeficiente de correlación intraclase (CCI), que estima el promedio de las correlaciones entre todas las posibles ordenaciones de los pares de observaciones disponibles. En medicina, este coeficiente se usa generalmente para valorar la concordancia entre dos o más mediciones continuas realizadas de forma repetida en una serie de sujetos y puede interpretarse como una medida de reproducibilidad o de fiabilidad.

La correlación intraclase puede ilustrarse de la siguiente forma: supongamos que las observaciones de una variable se ordenan en n grupos que contienen m observaciones cada uno, y que no hay motivos para esperar que haya diferencias en el nivel medio de la variable entre los n grupos; si esas diferencias existen, las observaciones del mismo grupo tenderán a estar positivamente correlacionadas. Esto es lo que se conoce como correlación intraclase [12].

Se han propuesto varios métodos para estimar el CCI, que dependen del diseño del estudio. Müller y Büttner analizan todas las posibles situaciones que pueden darse y los coeficientes que se han propuesto en cada una de ellas [13]. Las opciones consideradas son tres: los observadores pueden ser fijos o constituir una muestra aleatoria de una población mayor; también pueden ser o no indistinguibles unos de otros y, por último, puede ocurrir que todos los observadores midan a todos los sujetos o que esto no se cumpla.

El método implementado en Epidat corresponde a la situación en que los observadores son una muestra aleatoria, no todos los sujetos son necesariamente medidos por todos los observadores y las mediciones son intercambiables. La fórmula para el cálculo se basa en un modelo de análisis de la varianza de efectos aleatorios de un factor (ANOVA); la idea es que la variabilidad total de las mediciones se puede descomponer en dos componentes: la variabilidad debida a las diferencias entre los distintos sujetos (varianza *entre sujetos*) y la debida a las diferencias entre las medidas para cada sujeto (varianza *intra sujetos*). El CCI se calcula, entonces, como la proporción que supone la varianza *entre sujetos* sobre la variabilidad total [10]. Este coeficiente paramétrico puede considerarse el equivalente del estadístico kappa para variables continuas y toma valores entre 0 y 1; está próximo a 1 si la variabilidad observada se debe fundamentalmente a las diferencias entre los sujetos, y no a las diferencias entre los métodos de medición o entre los observadores, y toma el valor 0 cuando toda la concordancia observada es debida al azar. Aunque la interpretación es subjetiva, Fleiss propone una escala para valorar el CCI como medida de reproducibilidad: valores inferiores a 0,4 indican poca reproducibilidad y valores iguales o superiores a 0,75 reproducibilidad excelente; los valores intermedios se consideran adecuados [14].

Müller y Büttner discuten en detalle las limitaciones de los coeficientes revisados [13]. Entre las que afectan al CCI basado en un ANOVA, también apuntadas por otros autores [10], están el incumplimiento de las hipótesis del modelo (normalidad, igualdad de varianzas e independencia de los errores), y la dependencia del rango de variación de la escala de medida y del número de observadores; si, por ejemplo, una medición presenta una variabilidad reducida, puede obtenerse un CCI bajo sin que esto signifique un método poco consistente.

Una alternativa a los coeficientes de correlación intraclase, válida para el caso de dos observadores, es el método gráfico propuesto por Bland y Altman [11], también implementado en Epidat 4 como una novedad del módulo de Concordancia y consistencia, y que se comenta en el siguiente epígrafe.

6.1.3.1. Manejo del submódulo de coeficiente de correlación intraclase

Este submódulo permite estimar el coeficiente de correlación intraclase entre dos o más mediciones continuas realizadas a los mismos sujetos, así como su intervalo de confianza.

Para introducir los datos manualmente, es necesario especificar el número de sujetos y de mediciones (variables) y completar la tabla.

Al optar por la entrada automática se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija. En este caso, la tabla debe contener tantas filas como número de sujetos y tantas variables como mediciones realizadas a cada sujeto, hasta un máximo de 100.

Ejemplo 1 [15]

Las muestras de sangre recogidas a 5 mujeres se dividieron en dos partes alícuotas que fueron enviadas de forma ciega a un laboratorio para determinar su concentración de estradiol plasmático (pg/mL). Los resultados, en escala logarítmica, se muestran en la siguiente tabla.

Sujeto	Muestra 1	Muestra 2
1	3,24	3,41
2	2,41	2,71
3	2,08	2,09
4	3,03	2,83
5	1,76	2,13

El coeficiente de correlación intraclase calculado con estos datos es una medida de reproducibilidad. Para obtenerlo con Epidat, pueden introducirse los datos manualmente indicando el número de sujetos (5) y el número de medidas (2) y completando la tabla.

Resultados con Epidat 4:

Datos:

Número de sujetos:	5
Número de variables:	2
Nivel de confianza:	95,0%

Resultados:

Correlación	IC (95,0%)	
0,9145	0,5039	0,9905

El valor obtenido para el coeficiente CCI indica una buena reproducibilidad de los datos, aunque el intervalo de confianza es muy ancho debido al tamaño reducido de la muestra.

Ejemplo 2

La presión arterial es una medida imprecisa con una elevada variabilidad *intra sujeto*; aunque la técnica es bastante simple, pueden aparecer errores debidos a defectos del aparato utilizado, al estado del paciente y a la objetividad y preparación del observador. Por tanto, la forma de caracterizar la presión arterial de un individuo es realizar mediciones repetidas en un período corto de tiempo y considerar la media de todas ellas como el verdadero valor. Supongamos, por ejemplo, que se ha medido la presión arterial sistólica de 10 sujetos en dos momentos del día y se han obtenido los siguientes resultados hipotéticos:

SUJETO	PAS1	PAS2
1	176	156
2	162	142
3	141	121
4	162	142
5	165	145
6	141	121
7	168	148
8	133	113
9	149	129
10	147	127

El coeficiente de correlación de Pearson entre estas dos medidas es 1 (puede obtenerse con Epidat 4 en el módulo de análisis descriptivo), ya que PAS2=PAS1-20; sin embargo, el coeficiente de correlación intraclase indica poca fiabilidad en estos datos, con un valor inferior a 0,4.

Los datos de la tabla se encuentran en la hoja *Datos1* del archivo PAS.xls incluido en Epidat 4.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Concordancia y consistencia\PAS.xls
Tabla: Datos1
Variables:
Variables: PAS1, PAS2

Datos:

Número de sujetos: 10
Número de variables: 2
Nivel de confianza: 95,0%

Resultados:

Correlación	IC (95,0%)	
0,3285	0,0000	0,7738

6.1.4. Método de Bland y Altman

6.1.4.0. Conceptos generales

Bland y Altman propusieron en 1986 un sencillo método gráfico para evaluar la concordancia entre dos variables cuantitativas, y cuyo uso ha ido en aumento en los últimos años. Así lo ponen de manifiesto Dewitte et al, que revisaron los artículos publicados en Clinical Chemistry entre 1995 y 2001 y observaron una tendencia creciente en el uso del gráfico de Bland y Altman en estudios de comparación de métodos, desde un 8% en 1995 hasta un 31-36% en años más recientes [16].

Este procedimiento, descrito con detalle en 1999, resulta útil en estudios de comparación de dos métodos, o en estudios de fiabilidad de un único método, cuando se realizan dos mediciones repetidas a una serie de sujetos con el método a evaluar [17]. Como primer paso pueden representarse los datos gráficamente en un diagrama de dispersión, junto a la recta diagonal; sin embargo, con este tipo de gráfico no es fácil identificar diferencias entre los métodos. Es habitual, por otra parte, que se calcule el coeficiente de correlación de Pearson entre las dos mediciones, pero este coeficiente no es válido para cuantificar el grado de acuerdo entre ellas [10]. Por ejemplo, si se comparan las mediciones realizadas con dos básculas, una de las cuales pesa sistemáticamente 1 Kg. más que la otra, entonces la correlación entre los dos resultados puede ser muy alta, próxima a 1, a pesar de que la concordancia es nula. Las limitaciones que Bland y Altman apuntan sobre el coeficiente de correlación son [11]:

- La correlación expresa la fuerza de asociación lineal entre dos variables, pero no el acuerdo o concordancia entre ellas. La concordancia perfecta implica la coincidencia sobre la diagonal en un gráfico de dispersión.
- Un cambio en la escala de medida no afecta a la correlación, pero si afecta a la concordancia.
- La correlación depende del rango que la variable de interés tiene en la muestra.
- El test de significación estadística del coeficiente de correlación puede mostrar que los dos métodos están relacionados, pero esto es irrelevante para evaluar concordancia.
- Datos con concordancia pobre pueden presentar una correlación elevada.

El método de Bland y Altman consiste en representar gráficamente, en un diagrama de dispersión, la media de las dos mediciones, como la mejor estimación del verdadero valor, frente a la diferencia absoluta entre los dos valores. El gráfico incluye, además, una línea horizontal en la diferencia media y dos líneas, llamadas *límites de concordancia*, a una distancia de 1,96 desviaciones estándar por arriba y por debajo de la primera. Si las diferencias entre los pares de observaciones siguen aproximadamente una distribución normal y los valores tienden a ser estables en todo el rango de medición, se espera que el 95% de esas diferencias caigan dentro de los límites de concordancia. Esto permite valorar gráficamente, de forma sencilla, el grado de acuerdo entre los dos métodos.

Cuando la variabilidad de las diferencias aumenta, o disminuye, a medida que aumenta la magnitud de la media, Bland y Altman proponen aplicar a las dos mediciones una transformación logarítmica antes de hacer el análisis. También proponen aplicar métodos de regresión cuando la transformación logarítmica no estabiliza las diferencias. Para más detalles y ejemplos se recomienda consultar su trabajo [17].

Por último, para estimar la precisión de la diferencia media y de los límites de concordancia, estos valores deben acompañarse de un intervalo de confianza. Epidat 4 presenta, como resultados de este submódulo, el gráfico de Bland y Altman junto a una tabla con las estimaciones mencionadas.

6.1.4.1. Manejo del submódulo del método de Bland y Altman

Este submódulo permite representar el gráfico propuesto por Bland y Altman para analizar la concordancia entre dos mediciones cuantitativas, así como estimar la media y desviación estándar de las diferencias y los límites de concordancia acompañados de un intervalo de confianza.

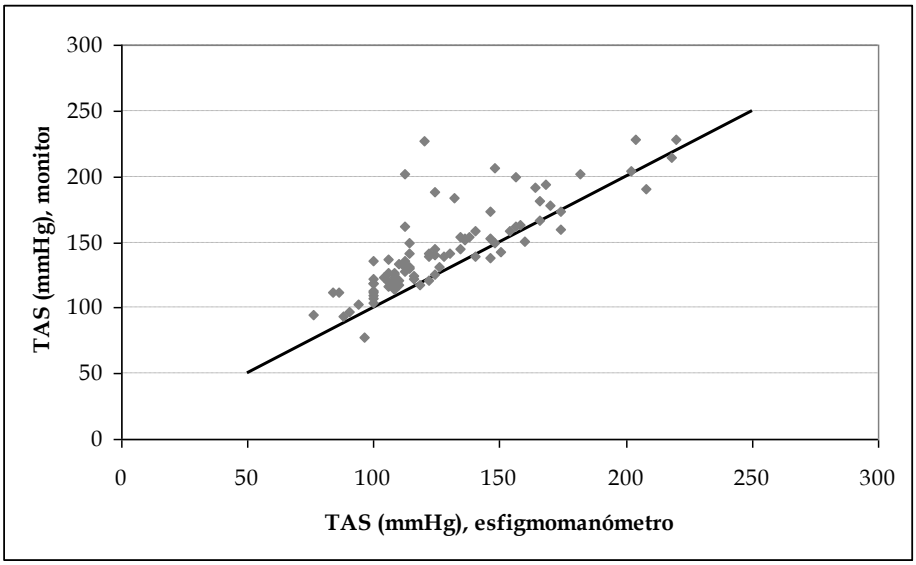
Para introducir los datos manualmente, es necesario especificar el número de sujetos y completar la tabla. El número de mediciones es siempre 2.

Al optar por la entrada automática se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija. En este caso, la tabla debe contener tantas filas como número de sujetos y dos variables que recojan las mediciones realizadas a cada sujeto.

Ejemplo

Para ilustrar la metodología de Bland y Altman se tomaron los datos recogidos en la tabla 1 de su artículo, que consisten en las mediciones de tensión arterial sistólica realizadas a 85 personas por medio de un esfigmomanómetro de mercurio en el brazo y las obtenidas mediante un monitor semiautomático. Los datos se encuentran en la hoja *Datos2* del archivo PAS.xls incluido en Epidat 4.

La correlación entre las dos medidas es alta, con un coeficiente de correlación de Pearson de 0,82; sin embargo, el diagrama de dispersión evidencia diferencias importantes entre los dos métodos:



Estas diferencias se ponen de manifiesto más claramente con el gráfico de Bland y Altman, que se puede obtener con Epidat 4. Como se observa en los resultados, el monitor

semiautomático proporciona valores de la presión arterial más altos que el esfigmomanómetro, con una diferencia media de 16 mmHg; los límites de concordancia indican que los valores del monitor están entre 55 mmHg por encima del esfigmomanómetro y 22 mmHg por debajo. Tales diferencias no son clínicamente aceptables como para considerar equivalentes los dos instrumentos.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Concordancia y consistencia\PAS.xls
Tabla: Datos2
Variables:
Observación 1: PAS_ESFIG
Observación 2: PAS_MONITOR

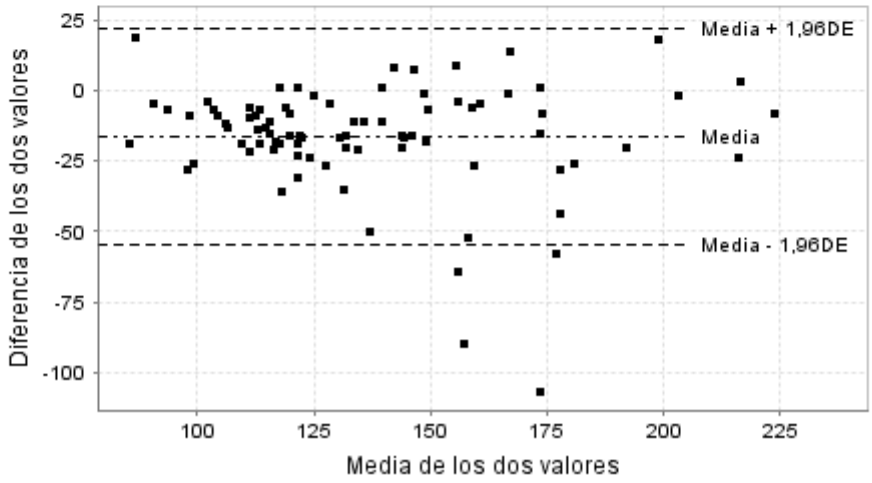
Datos:

Nº de sujetos:	85
Nivel de confianza:	95,0%

Resultados:

	Valor	IC (95,0%)	
Media de las diferencias	-16,2941	-20,5241	-12,0641
DE de las diferencias	19,6110		
Media-1,96DE*	-54,7309	-61,9600	-47,5018
Media+1,96DE*	22,1426	14,9135	29,3718

*DE: desviación estándar



6.2. Consistencia: alfa de Cronbach

6.2.0. Conceptos generales

El coeficiente alfa de Cronbach se emplea para medir lo que ha dado en llamarse "consistencia interna de una escala". Esta expresión exige, para comenzar, algunas precisiones.

Ocasionalmente, los investigadores se ven ante la tarea de construir un indicador capaz de medir cierto concepto abstracto. En esa línea se han desarrollado numerosos procedimientos con los que se intenta cuantificar nociones tales como, la *capacidad de liderazgo*, el *dolor que experimenta un politraumatizado*, la *gravedad de un proceso asmático*, la *discapacidad funcional del anciano* o la *calidad de vida del trasplantado renal*.

Tal proceso es llamado en ocasiones "construcción de una escala". Indudablemente, esta expresión se ha acuñado con bastante firmeza. A nuestro juicio, sin embargo, no es especialmente afortunada, ya que la palabra *escala* está reservada para denominar el tipo de medición que se realiza o la métrica que se emplea (nominal, ordinal, etc.); las escalas, por tanto, no se *construyen* sino que se *usan* en el acto de construcción de una variable o indicador. Consecuentemente, se aludirá en lo sucesivo a la construcción de una *variable sintética* (VS) [18] para referirnos a una función de un conjunto de variables intermedias o ítems, cada una de las cuales contribuye a cuantificar algún rasgo del concepto cuya magnitud quiere sintetizarse.

La creación de una VS para la medición de la salud personal (física y psíquica) por medio del escrutinio múltiple de los sujetos, el cual aportaba puntajes para diferentes aspectos medidos en dichos individuos que producían una única dimensión integrada, fue metodológicamente impulsada en Estados Unidos durante la segunda guerra mundial y en virtud de la necesidad de valorar grandes cantidades de reclutas [19]. Un uso muy extendido de este tipo de variables se produce en el campo de la psicología, disciplina que quizás haya acopiado la mayor experiencia al respecto, tal y como testimonian los múltiples esfuerzos realizados desde la década del 40 bajo el auspicio de la *American Psychological Association*, profusamente citados en artículos clásicos de la época, de los cuales quizás el más connotado sea el de Cronbach y Meehl [20].

En casi todas las áreas, la suma (ocasionalmente ponderada) de las puntuaciones de los ítems individuales es el índice más empleado. Sin embargo, se ha insistido en que tal procedimiento es procedente sólo si ellos miden de algún modo el mismo rasgo. Se suele afirmar que cuando los ítems conciernen a atributos totalmente diferentes, no es en general razonable formar una única variable sintética con ellos.

La materia prima de tal variable integrada suele ser el conjunto de respuestas a un cuestionario, en cuyo caso la VS se construye mediante alguna regla integradora de esas respuestas. La situación típica es similar a la que se produce con las famosas y controvertidas pruebas de inteligencia: tras indicarle la tarea de dar solución a una serie de problemas que se puntúan separadamente, al sujeto se le atribuye un puntaje global, con el que se calcula el polémico cociente de inteligencia conocido como IQ (Inteligence Quotient, por sus siglas en inglés). Otro ejemplo clásico, en este caso de la clínica, es la propuesta de Apgar [21] para cuantificar la vitalidad de un recién nacido en función del pulso cardíaco, el esfuerzo respiratorio, el tono muscular, el color de la piel y la respuesta al estímulo producido por la colocación de un catéter en las fosas nasales.

Las propiedades técnicas fundamentales que se suelen demandar para una variable sintética son que posea fiabilidad (*reliability*) y validez (*validity*). En este submódulo, Epidat 4 se ocupa de una forma concreta de la fiabilidad.

La llamada *fiabilidad externa* (esencialmente consistente en corroborar que se obtienen resultados muy similares cuando se repite la medición) es una demanda cuyo interés para otorgar confianza a la VS es altamente intuitivo. Para medirla se han desarrollado diversos indicadores. Pero existe otra forma de fiabilidad de una VS, la llamada *consistencia interna*. En términos generales, se dice que una VS exhibe consistencia interna cuando hay una alta asociación entre los ítems que la integran.

En muchas situaciones, cuando se está midiendo cierto atributo, se quiere que los componentes que conforman la variable abarquen distintos aspectos de dicho atributo, y no cuestiones aisladas o ajenas entre sí. Por ejemplo, si se está midiendo la habilidad de los estudiantes de medicina para resolver problemas clínico-terapéuticos, entonces cada área, problema o ítem del examen que mide esta habilidad debe estar relacionado con la resolución de este tipo de problemas (no debería, por ejemplo, evaluar el conocimiento que tiene el estudiante sobre demografía o sobre el precio de los fármacos).

En principio, tal condición tiene dos implicaciones: a) que los ítems deben estar correlacionados entre sí y b) que cada ítem debe estar correlacionado con la puntuación total de la VS. La segunda condición parece razonable, pero la primera es muy discutible, pues el atributo global puede desagregarse en componentes que estén, en buena medida, mutuamente incorrelacionados, como se analiza más abajo.

La medición de la consistencia interna ha sido objeto de diversas propuestas. Una de ellas es la llamada *fiabilidad basada en mitades (split-half fiability)*. Este procedimiento exige:

- dividir (ocasionalmente se ha sugerido hacerlo al azar) en dos subgrupos a los ítems que integran la VS,
- con una y otra mitad separadamente evaluar a todos los sujetos (digamos, n) que integren la muestra,
- computar las sumas resultantes al emplear cada mitad en cada sujeto,
- calcular la correlación que exhiben esos n pares de valores.

Esta variante, sin embargo, presenta dos problemas. En primer lugar, el número de formas de hacer la división es enorme y cada uno de ellos arroja un resultado distinto. Concretamente, si los ítems son k, entonces el número de maneras de dividir el conjunto de ítems en mitades es igual al número de subconjuntos de tamaño k/2 que pueden formarse con un conjunto de tamaño k (por ejemplo, si k=10, hay nada menos que 252 formas de producir tal división). Por otra parte, este procedimiento no permite identificar cuáles son los ítems que contribuyen a una ocasional pérdida de fiabilidad.

En cualquier caso, es obviamente atractivo contar con una medida cuantitativa del grado en que los ítems están relacionados entre sí; es decir, del grado de "homogeneidad" interna de la VS.

El indicador más conocido para medir esta forma de fiabilidad es el llamado *coeficiente alfa*, al que se denotará por C_α, propuesto por Cronbach [22]; cuando todos los ítems o variables intermedias son dicotómicos, este coeficiente se reduce al conocido KR-20, *coeficiente de Kuder-Richardson*.

El alfa de Cronbach es una cota inferior de todas las correlaciones que se obtendrían si se aplicara la fiabilidad basada en mitades para todas las maneras posibles de dividir los ítems [23]. En el ejemplo arriba mencionado, sería inferior a los 252 coeficientes de correlación susceptibles de ser obtenidos.

En términos prácticos, se le atribuyen dos usos básicos. En primer lugar, como instrumento para la medición de la homogeneidad interna de la VS mirada globalmente. En ese sentido, tiene un valor intrínseco. Pero también puede usarse como recurso para hacer juicios relativos, lo que tal vez constituya su virtud máxima consistente en que permite analizar la contribución que cada ítem particular hace a la homogeneidad de la VS. Tal contribución se mide mediante el recurso de comparar el valor del coeficiente que se obtiene cuando se emplean todos los ítems con el que resulta de hacer el cálculo luego de haberse eliminado dicho ítem. Así, si alfa aumenta significativamente tras eliminar un ítem específico, esto indicaría que la exclusión de este último aumentaría la homogeneidad de la escala y viceversa.

El coeficiente alfa, sin embargo, es muy controvertido. Llama la atención, de hecho, que medio siglo después de su creación, aún sigan apareciendo artículos sugiriendo posibles interpretaciones y llamando la atención sobre interpretaciones presuntamente erróneas. Repárese, por ejemplo, en los trabajos de Gardner y de Cortina [24] [25][26]. Por ejemplo, Rubin [27] menciona la posible existencia de "subconstructos" que no necesariamente estén relacionados entre sí. Siendo así, según este autor, un valor bajo de C_α puede producirse para una VS fiable (lo ilustra con una VS construida para medir el consumo diario de proteínas: un ítem puede medir el consumo de pescado y otro el de carne; la correlación entre estos dos puede ser baja, ya que quien consume una cosa suele no consumir la otra el mismo día; y sin embargo, la VS puede medir bien el constructo global de interés. Rubin también ilustra el caso en que una VS fiable puede exhibir tanto un C_α alto como un C_α bajo, y finalmente ofrece un ejemplo en que solo un C_α alto puede esperarse de una VS que sea realmente fiable. Tal debate, sin embargo, remite al examen del marco concreto en que se da cada problema, y desborda el alcance de la presente exposición. Una discusión bastante completa puede hallarse en Streiner y Norman [28].

El primer problema que presenta C_α es que depende no sólo de la magnitud de la correlación entre los ítems, sino también del número de ítems involucrados. Se podría conseguir que una variable sintética pareciera más homogénea simplemente duplicando (en general aumentando) el número de ítems, incluso aunque la correlación entre ellos permaneciera constante.

Debe decirse que en relación con la interpretación de este coeficiente hay bastante confusión. Por ejemplo, no es infrecuente hallar textos que afirman que el valor de C_α varía entre 0 y 1 [29]. Sin embargo, esto es falso. Este indicador puede alcanzar valores negativos de alfa si los ítems no están positivamente correlacionados entre sí. De hecho, puede probarse que C_α no solo puede ser negativo sino que puede alcanzar cualquier valor inferior a cero (es decir, no está acotado inferiormente). Es recomendable que todos los ítems tengan el mismo sentido para evitar correlaciones negativas, de modo que siempre se obtengan valores de alfa entre 0 y 1. Algunos paquetes estadísticos, como Stata, invierten automáticamente el sentido de los ítems negativos antes de calcular el coeficiente C_α [30].

La conveniencia de que C_α sea elevado es controversial, ya que una alta asociación tras la maniobra de recalcularlo eliminando un ítem reflejaría algún grado de redundancia en la información que se registra; consecuentemente, es lógico aspirar a que los componentes de la VS recorran dimensiones que sean, en buena medida, independientes.

Es por ello que se suele plantear que, si alfa es demasiado alto, ello pudiera estar sugiriendo un elevado nivel de redundancia entre los ítems. Por tanto, desde el punto de vista práctico, si bien es atractivo que el coeficiente alfa sea alto (por ejemplo, superior a 0,7), sería deseable que ello no ocurra en demasía (no superar el valor 0,9). Esto es evidente, ya que si todos los ítems miden exactamente lo mismo, entonces C_α=1. En cualquier caso, es obvio que el

empleo de este indicador puede ser polémico; el usuario de este recurso debe examinar por sí mismo su problema y decidir qué uso hará de él y qué conclusiones sacará de los resultados.

6.2.1. Manejo del submódulo de consistencia: Alfa de Cronbach

Este submódulo permite calcular el valor del coeficiente alfa de Cronbach para un conjunto de ítems. También presenta el valor obtenido al eliminar, sucesivamente, cada uno de los ítems, excepto en el caso de que se tengan solamente dos ítems.

Para introducir los datos manualmente, es necesario especificar el número de sujetos y de ítems y completar la tabla.

Al optar por la entrada automática se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija; en este caso, cada columna de la tabla contiene los valores individuales de un ítem.

Ejemplos

Considérese el siguiente ejemplo. Supóngase que se quiere medir el *grado en que un inmigrante se ha adaptado a su nuevo medio*. Supóngase que se consideran 5 dimensiones (ítems) las cuales se asume que se asocian al grado de adaptación y que todas ellas pueden tomar valores del 1 al 5. Las preguntas (ítems) que se consideran son:

A: grado en que maneja el idioma del país de acogida
B: nivel de satisfacción que tiene con el trabajo que realiza
C: interés que muestra por regresar definitivamente a su país de origen
D: acceso a los servicios de salud en el país de acogida
E: grado en que ha conseguido legalizar su situación

Supóngase, finalmente, que las cinco preguntas se miden mediante una escala ordinal del modo siguiente: 1-Nulo; 2-Escaso; 3-Adecuado; 4-Muy bueno; 5-Excelente.

Supóngase que se ha aplicado este cuestionario a diez individuos y que los resultados son los siguientes:

Sujeto	Ítem A	Ítem B	Ítem C	Ítem D	Ítem E
1	3	4	5	1	4
2	3	2	5	1	3
3	4	4	4	4	4
4	4	5	4	1	2
5	2	4	5	5	5
6	5	4	5	1	4
7	4	4	5	4	4
8	4	4	4	1	4
9	5	5	1	1	2
10	1	1	1	1	2

Los datos de esta tabla se encuentran en el archivo CUESTIONARIO.XLS, incluido en Epidat 4. Al hacer el cálculo con Epidat se obtienen los siguientes resultados:

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Concordancia y consistencia\CUESTIONARIO.xls
Tabla: Datos
Variables:
Ítems: ITEM-A, ITEM-B, ITEM-C, ITEM-D, ITEM-E

Datos:

Número de ítems: 5
Número de observaciones: 10

Resultados:

Alpha de Cronbach: 0,6616
Covarianza media: 0,5367

Ítem eliminado	Alpha de Cronbach
1	0,6901
2	0,5947
3	0,5584
4	0,6596
5	0,5274

Una interpretación típica sería, en principio, que el ítem A podría (o debería) ser eliminado del cuestionario, ya que su supresión incrementa la fiabilidad interna de la VS.

Si un ítem tuviera para todos los sujetos un valor constante, el valor de C_α se reduciría. Por ejemplo, si el ítem E pasara a valer 4 para todos los sujetos, el valor que se obtendría pasaría a ser C_α = 0,4944. Es obvio que un ítem poco informativo como éste debería eliminarse. Notar que, sin él, el valor que se obtiene es superior (0,5274).

A modo de curiosidad, obsérvese en el siguiente ejemplo que el valor del coeficiente puede ser negativo y, además, enorme.

Sujeto	Ítem A	Ítem B	Ítem C	Ítem D	Ítem E
1	3	3	5	1	4
2	3	3	5	1	3
3	4	2	4	2	4
4	4	2	4	2	2
5	2	4	5	1	5
6	5	1	5	1	4
7	4	2	5	1	4
8	4	2	4	2	4
9	5	1	1	5	2
10	1	5	1	5	2

En tal situación se tendría nada menos que C_α = -8,9904

Bibliografía

1 Shoukri MM. Measurement of agreement. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons; 1998. pp. 103-17.

2 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

3 Feinstein AR, Cicchetti DV. High agreement but low kappa. I. The problems of two paradoxes. J Clin Epidemiol. 1990;43:543-9.

4 Lantz CA, Nebenzahl E. Behavior and interpretation of the k statistic: Resolution of the two paradoxes. J Clin Epidemiol. 1996;49(4):431-4.

5 Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley & Sons; 1981.

6 Thompson WD, Walters SD. A reapraisal of the kappa coefficient. J Clin Epidemiol. 1998;41(10):949-58.

7 De Vet H. Observer reliability and agreement. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 4. Chichester: John Wiley & Sons; 1998. pp. 3123-7.

8 Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther. 1981;29(1):111-23.

9 Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

10 Pita Fernández S, Pértega Díaz S. La fiabilidad de las mediciones clínicas: el análisis de concordancia para variables numéricas. Atención Primaria en la red. Fisterra.com. [Actualizado 12 Ene 2004]. Disponible en : http://www.fisterra.com/mbe/investiga/conc_numerica/conc_numerica.asp

11 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;i:307-10.

12 Armitage P. Correlation. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons; 1998. pp. 974-5.

13 Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med. 1994;13:2465-76.

14 Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.

15 Rosner B. Fundamentals of biostatistics. 5ª ed. Belmont, CA: Duxbury Press; 2000.

16 Dewitte K, Fierens C, Stöckl D, Thienpont LM. Application of the Bland-Altman plot for interpretation of method-comparison studies: A critical investigation of its practice [carta]. Clinical Chemistry. 2002;48(5):799-801.

17 Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8:135-60.

18 Silva LC. Cultura estadística e investigación científica en el campo de la salud: una mirada crítica. Madrid: Díaz de Santos; 1997.

19 Dowell I, Newell C. Measuring health. New York: Oxford University Press; 1987.

20 Cronbach L, Meehl P. Construct validity in psychological test. Psychol Bull. 1955;52:281-302.

21 Apgar V. Proposal for method of evaluation of newborn infant. Anesthesiology and Analgesics. 1953;32:260-7.

22 Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-334.

23 Dukes K. Cronbach's alpha. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons; 1998. pp. 1026-8.

24 Gardner PL. Measuring attitudes to science: Unidimensionality and internal consistency revisited. Research in Science Education. 1995;25:283-9.

25 Gardner PL. The dimensionality of attitude scales: A widely misunderstood idea. Int J Sci Educ. 1996;18:913-9.

26 Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78:98-104.

27 Rubin HR. Psychometrics or psycho metrics? Alpha abuse. 2002.

28 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. New York: Oxford University Press; 1989.

29 Santos JR. Cronbach's alpha: A tool for assessing the reliability of scales. Journal of Extension [serie en Internet]. Abr 1999;37(2):[aprox. 5 p.]. Disponible en: http://joe.org/joe/1999april/tt3.html

30 StataCorp. Stata Statistical Software: release 10. College Station, TX: StataCorp LP; 2007.

Anexo 1: Novedades del módulo de concordancia y consistencia

Novedades de la versión 4.0 con respecto a la versión 3.1:

- En el submódulo de concordancia se incluyen dos opciones nuevas para valorar la concordancia entre medidas con resultado continuo: el coeficiente de correlación intraclase y el método gráfico de Bland y Altman.

- Alfa de Cronbach: los valores ausentes de la tabla de datos no se identifican con el valor -9, o con otro valor numérico, como se hacía en la versión 3.1, sino que se dejan en blanco.

- Alfa de Cronbach: en los resultados se presenta la covarianza media entre los ítems.

Novedades de la versión 4.0.1 con respecto a la versión 4.0:

- Concordancia entre dos observadores: era incorrecto el cálculo de kappa cuando el número de categorías era mayor que 2. Se corrige este error en el programa y se corrigen, en la ayuda, los resultados del ejercicio B de este submódulo.

Novedades de la versión 4.1 con respecto a la versión 4.0.1:

- Método de Bland y Altman: se modifica la fórmula para el cálculo de los límites de concordancia: en la expresión Media±2DE se sustituye el valor 2 por el percentil de orden 97,5 de la distribución normal estándar.

- Método de Bland y Altman: en el gráfico estaban intercambiados los textos de las etiquetas "Media-2DE" y "Media+2DE", actualmente "Media-1,96DE" y "Media+1,96DE", respectivamente. Se corrige este error.

Anexo 2: Fórmulas del módulo de concordancia y consistencia

Esquema del módulo

1. Concordancia
   - 1.1. Dos observadores
   - 1.2. Tres o más observadores
     - 1.2.1. Dos categorías
     - 1.2.2. Tres o más categorías
   - 1.3. Comparación de kappas
   - 1.4. Coeficiente de correlación intraclase
   - 1.5. Método de Bland y Altman
2. Consistencia
   - 2.1. Alfa de Cronbach

1.1. CONCORDANCIA. DOS OBSERVADORES

Acuerdo observado, acuerdo esperado y kappa [Fleiss (1981, p. 223-225)]:

Acuerdo observado:

$$p_o = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{ij}$$

Acuerdo esperado:

$$p_e = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i.} p_{.j}$$

Kappa:

$$\hat{k} = \frac{p_o - p_e}{1 - p_e}$$

Intervalo de confianza con nivel de confianza (1-α)%:

$$\left(\hat{k} - z_{1-\alpha/2} EE(\hat{k}) \, , \, \hat{k} + z_{1-\alpha/2} EE(\hat{k})\right)$$

Error estándar de kappa para el intervalo de confianza:

$$EE(\hat{k}) = \frac{1}{(1-p_e)\sqrt{n}} \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[w_{ij} - (\overline{w}_{i.} + \overline{w}_{.j})(1-\hat{k})\right]^2 - \left[\hat{k} - p_e(1-\hat{k})\right]^2}$$

Prueba de significación para kappa [Fleiss (1981, p. 224)]:

Estadístico para contrastar H_0: k=0 frente a H_1: k≠0:

$$z = \frac{\hat{k}}{EE_0(\hat{k})} \approx N(0,1)$$

Error estándar de kappa para el contraste:

$$EE_0(\hat{k}) = \frac{1}{(1-p_e)\sqrt{n}} \sqrt{\left(\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i.}p_{.j}\left[w_{ij} - (\overline{w}_{i.} + \overline{w}_{.j})\right]^2 \right) - p_e^2}$$

Kappa mínimo y máximo para un acuerdo observado p_0 (si k=2) [Lantz & Nebenzahl (1996)]:

Kappa mínimo:

$$\hat{k}_{min} = \frac{p_0 - 1}{p_0 + 1}$$

Kappa máximo:

$$\hat{k}_{max} = \frac{p_0^2}{(1-p_0)^2 + 1}$$

Donde:

- k es el número de categorías,
- x_{ij} es el número de sujetos clasificados en la categoría i por el observador 1 y en la categoría j por el observador 2, i,j=1,...,k,
- $n = \sum_{i=1}^{k}\sum_{j=1}^{k} x_{ij}$ es el número total de sujetos,
- $p_{ij} = \frac{x_{ij}}{n}$ es la proporción de sujetos clasificados en la categoría i por el observador 1 y en la categoría j por el observador 2, i,j=1,...,k,
- $p_{i.} = \sum_{j=1}^{k} p_{ij}$ y $p_{.j} = \sum_{i=1}^{k} p_{ij}$, i,j=1,...,k,
- w_{ij} es el peso asignado a la celda (i,j) de la tabla, i,j=1,...,k,:
  - Si no hay ponderación: $w_{ij} = \begin{cases} 1 & si \quad i = j \\ 0 & si \quad i \neq j \end{cases}$,
  - Pesos de Cicchetti: $w_{ij} = 1 - \frac{|i-j|}{k-1}$,

- Pesos cuadráticos: $w_{ij} = 1 - \left(\frac{i-j}{k-1}\right)^2$,

- Ponderación manual: $w_{ii} = 1$; $0 \leq w_{ij} < 1$ y $w_{ij} = w_{ji}\ \forall\ i \neq j$

- $\overline{w}_{i.} = \sum_{j=1}^{k} w_{ij} p_{.j}$ y $\overline{w}_{.j} = \sum_{i=1}^{k} w_{ij} p_{i.}$, i,j=1,…k,
- $z_{1-\alpha/2}$ es el percentil de la distribución normal estándar, N(0,1), que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$,
- 1-α es el nivel de confianza.

1.2. CONCORDANCIA. TRES O MÁS OBSERVADORES

1.2.1 DOS CATEGORÍAS

Kappa e intervalo de confianza:

Kappa [Fleiss (1981, p. 226-227)]:

$$\hat{k} = 1 - \frac{\sum_{i=1}^{n} \frac{x_i(m_i - x_i)}{m_i}}{n(\overline{m}-1)\overline{p}(1-\overline{p})}$$

Donde:

- n es el número de sujetos,
- m_i es el número de observadores para el sujeto i-ésimo, i=1,...,n,
- $\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ es el número medio de observadores por sujeto,
- x_i es el número de clasificaciones positivas del sujeto i-ésimo, i=1,...,n,
- $\overline{p} = \frac{1}{n\overline{m}}\sum_{i=1}^{n} x_i$ es la proporción global de clasificaciones positivas.

Intervalo de confianza jackknife para kappa con nivel de confianza (1-α)% [Abraira (1999), Efron & Tibshirani (1993, p. 145)]:

$$\left(J(k)-t_{n-1,1-\alpha/2}S\,,\,J(k)+t_{n-1,1-\alpha/2}S\right)$$

A partir de una muestra de n individuos, la i-ésima muestra jackknife, de tamaño n-1, se obtiene eliminando el individuo i.

- $J(k)=\frac{1}{n}\sum_{i=1}^{n}\hat{k}_{(i)}$ es la estimación jackknife de kappa,
- $S=\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{k}_{(i)}-\tilde{k}\right)^2}$ es el error estándar de J(k),
- $\hat{k}_{(i)}$ es la estimación de kappa obtenida con la submuestra que no contiene al individuo i, y $\hat{k}$ es la estimación de kappa con toda la muestra,
- $\tilde{k}_{(i)}=n\hat{k}-(n-1)\hat{k}_{(i)}$, i=1,…n,
- $\tilde{k}=\frac{1}{n}\sum_{i=1}^{n}\tilde{k}_{(i)}$,
- $t_{n-1,1-\alpha/2}$ es el percentil de la distribución t de Student con n-1 grados de libertad que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$,
- 1-α es el nivel de confianza.

<u>Prueba de significación para kappa</u> [Fleiss (1981, p. 228)]:

Estadístico para contrastar H_0: k=0 frente a H_1: k≠0:

$$z=\frac{\hat{k}}{EE_0(\hat{k})}\approx N(0,1)$$

Error estándar de kappa para el contraste:

$$EE_0(\hat{k})=\frac{1}{(\bar{m}-1)\sqrt{n\bar{m}_H}}\sqrt{2(\bar{m}_H-1)+\frac{(\bar{m}-\bar{m}_H)(1-4\bar{p}(1-\bar{p}))}{\bar{m}\bar{p}(1-\bar{p})}}$$

Donde:

- $\overline{m}_H = \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{m_i}}$ es la media armónica de los valores m_i, i=1,...,n,

1.2.2. TRES O MÁS CATEGORÍAS

Kappa de la categoría j e intervalo de confianza:

Kappa de la categoría j [Fleiss (1981, p. 230)]:

$$\hat{k}_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)\overline{p}_j\overline{q}_j}$$

Intervalo de confianza jackknife para el kappa de la categoría j con nivel de confianza (1-α)% [Abraira (1999), Efron & Tibshirani (1993, p. 145)]:

$$\left(J(k_j) - t_{n-1,1-\alpha/2} S_j \, , \, J(k_j) + t_{n-1,1-\alpha/2} S_j \right)$$

Prueba de significación para el kappa de la categoría j [Fleiss (1981, p. 231)]:

Estadístico para contrastar H_0: $k_j=0$ frente a H_1: $k_j \neq 0$:

$$z = \frac{\hat{k}_j}{EE_0(\hat{k}_j)} \approx N(0,1)$$

Error estándar de kappa para el contraste:

$$EE_0(\hat{k}_j) = \sqrt{\frac{2}{nm(m-1)}}$$

Donde:

- n es el número de sujetos,
- m es el número de observadores por sujeto,
- k es el número de categorías,
- x_{ij} es el número de clasificaciones del sujeto i en la categoría j, i=1,...,n, j=1,…,k,

- $\bar{p}_j = \frac{1}{nm}\sum_{i=1}^{n} x_{ij}$ es la proporción global de clasificaciones en la categoría j, j=1,…k, y $\bar{q}_j = 1 - \bar{p}_j$,
- $J(k_j)$ es la estimación jackknife de kappa, y S_j es el error estándar de $J(k_j)$, que se obtienen por el procedimiento descrito en el epígrafe 2.1,
- $t_{n-1,1-\alpha/2}$ es el percentil de la distribución t de Student con n-1 grados de libertad que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$,
- 1-α es el nivel de confianza.

Kappa global e intervalo de confianza:

Kappa global [Fleiss (1981, p. 229)]:

$$\hat{k} = \frac{\sum_{j=1}^{k} \bar{p}_j(1-\bar{p}_j)\hat{k}_j}{\sum_{j=1}^{k} \bar{p}_j(1-\bar{p}_j)}$$

Intervalo de confianza jackknife para el kappa global con nivel de confianza (1-α)% [Abraira (1999), Efron & Tibshirani (1993, p. 145)]:

$$\left(J(k) - t_{n-1,1-\alpha/2}S\,,\, J(k) + t_{n-1,1-\alpha/2}S\right)$$

Prueba de significación para el kappa global [Fleiss (1981, p. 231)]:

Estadístico para contrastar H_0: k=0 frente a H_1: k≠0:

$$z = \frac{\hat{k}}{EE_0(\hat{k})} \approx N(0,1)$$

Error estándar de kappa para el contraste:

$$EE_0(\hat{K}) = \frac{\sqrt{2}}{\sum_{j=1}^{k} \bar{p}_j \bar{q}_j \sqrt{nm(m-1)}} \sqrt{\left(\sum_{j=1}^{k} \bar{p}_j \bar{q}_j\right)^2 - \sum_{j=1}^{k} \bar{p}_j \bar{q}_j (\bar{q}_j - \bar{p}_j)}$$

Donde:

- $\hat{k}_j$ es el kappa de la categoría j, j=1,…k.
- J(k) es la estimación jackknife de kappa, y S es el error estándar de J(k), que se obtienen por el procedimiento descrito en el epígrafe 2.1.

1.3. CONCORDANCIA. COMPARACIÓN DE KAPPAS

Kappa global e intervalo de confianza [Fleiss (1981, p. 222)]:

Kappa global:

$$\hat{k} = \frac{\sum_{j=1}^{g} W_j \hat{k}_j}{\sum_{j=1}^{g} W_j}$$

Intervalo de confianza para el kappa global con nivel de confianza (1-α)%:

$$\left(\hat{k} - z_{1-\alpha/2} EE(\hat{k}) \, , \, \hat{k} + z_{1-\alpha/2} EE(\hat{k})\right)$$

Error estándar del kappa global:

$$EE(\hat{k}) = \sqrt{\frac{1}{\sum_{j=1}^{g} W_j}}$$

Prueba de homogeneidad [Fleiss (1981, p. 222)]

Estadístico χ^2 para contrastar H_0: $k_1 = k_2 = ... = k_g$:

$$\chi^2 = \sum_{j=i}^{g}\left(\frac{\hat{k}_j - \hat{k}}{EE(\hat{k}_j)}\right)^2$$,que sigue una distribución χ^2 con g-1 grados de libertad,

Donde:

- g es el número de kappas que se comparan,
- $\hat{k}_j$ es el j-ésimo valor de kappa que se compara, j=1,...,g,
- $W_j = \frac{1}{\left(EE(\hat{k}_j)\right)^2}$ es el peso de $\hat{k}_j$,
- $EE(\hat{k}_j)$ es el error estándar de $\hat{k}_j$, j=1,...,g,
- $z_{1-\alpha/2}$ es el percentil de la distribución normal estándar, N(0,1), que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$,
- 1-α es el nivel de confianza.

1.4. CONCORDANCIA. COEFICIENTE DE CORRELACIÓN INTRACLASE

Coeficiente de correlación intraclase [Rosner (2000, p. 558-564)]:

$$\hat{\rho}_I = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}^2}$$

Intervalo de confianza con nivel de confianza (1-α)%:

$$\left(\max\left\{\frac{F_I - 1}{k_0 + F_I - 1}, 0\right\}, \max\left\{\frac{F_S - 1}{k_0 + F_S - 1}, 0\right\}\right)$$

Donde:

- n es el número de sujetos,
- k_i es el número de observaciones del sujeto i, i=1,…,n,

- y_{ij} es la observación j del i-ésimo sujeto, j=1,…,k_i, i=1,…,n,
- $\hat{\sigma}_A^2 = \max\left[\frac{CME-\hat{\sigma}^2}{k_0},0\right]$ es la estimación de la varianza entre-sujetos en un modelo ANOVA de efectos aleatorios con un factor,
- $\hat{\sigma}^2 = \frac{1}{K-n}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(y_{ij}-\overline{y}_i\right)^2$ es la estimación de la varianza intra-sujetos,
- $CME = \frac{1}{n-1}\sum_{i=1}^{n}k_i\left(\overline{y}_i-\overline{\overline{y}}\right)^2$ es el cuadrado medio entre-sujetos,
- $\overline{y}_i = \frac{1}{k_i}\sum_{j=1}^{k_i}y_{ij}$, $\overline{\overline{y}} = \frac{1}{K}\sum_{i=1}^{n}k_i\overline{y}_i$, $K = \sum_{i=1}^{n}k_i$,
- $k_0 = \frac{1}{n-1}\left(\sum_{i=1}^{n}k_i - \frac{\sum_{i=1}^{n}k_i^2}{\sum_{i=1}^{n}k_i}\right)$ y $k_0 = k$ si $k_1 = \ldots = k_n = k$
- $F_I = \frac{F}{F_{n-1,K-n,1-\alpha/2}}$, $F_S = \frac{F}{F_{n-1,K-n,\alpha/2}}$ y $F = \frac{CME}{\hat{\sigma}^2}$,
- $F_{n-1,K-n,1-\alpha/2}$ es el percentil de la distribución F de Snedecor con n-1 y K-n grados de libertad que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$,
- 1-α es el nivel de confianza.

1.5. CONCORDANCIA. MÉTODO DE BLAND Y ALTMAN [Bland & Altman (1986)]

Media de las diferencias:

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n}d_i$$

Desviación estándar de las diferencias:

$$s_d = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i-\overline{d}\right)^2$$

Intervalo de confianza para la media de las diferencias con nivel de confianza (1-α)%:

$$\left(\overline{d} - t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}}\,,\ \overline{d} + t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}}\right)$$

Límites de concordancia:

Inferior: $d_I = \overline{d} - 1{,}96 \cdot s_d$

Superior: $d_S = \overline{d} + 1{,}96 \cdot s_d$

Intervalo de confianza para los límites de concordancia con nivel de confianza (1-α)%:

Para el límite inferior: $\left(d_I - t_{x,n-1}\frac{s_d}{\sqrt{n}}\sqrt{1+\frac{(1{,}96)^2}{2}}\,,\ d_I + t_{x,n-1}\frac{s_d}{\sqrt{n}}\sqrt{1+\frac{(1{,}96)^2}{2}}\right)$

Para el límite superior: $\left(d_S - t_{x,n-1}\frac{s_d}{\sqrt{n}}\sqrt{1+\frac{(1{,}96)^2}{2}}\,,\ d_S + t_{x,n-1}\frac{s_d}{\sqrt{n}}\sqrt{1+\frac{(1{,}96)^2}{2}}\right)$

Donde:

- n es el número de sujetos,
- y_{ij} es la observación j del sujeto i, i=1, …, n y j=1,2,
- $d_i = y_{i1} - y_{i2}$ es la diferencia entre las observaciones del sujeto i, i=1, …, n,
- $m_i = \frac{y_{i1} + y_{i2}}{2}$ es la media de las observaciones del sujeto i, i=1, …, n,
- $t_{n-1,1-\alpha/2}$ es el percentil de la distribución t de Student con n-1 grados de libertad que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$,
- 1-α es el nivel de confianza.

2.1. CONSISTENCIA. ALFA DE CRONBACH

Alfa de Cronbach [Nunnally & Bernstein (1994, p. 232-234)]:

$$\alpha_1 = \frac{k\bar{c}}{\bar{v} + (k-1)\bar{c}}$$

Otra fórmula equivalente [Bland & Altman (1997)]:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} c_{ii}}{s^2}\right)$$

Donde:

- k es el número de ítems,
- n es el número de sujetos,
- x_{ij} es el valor del ítem j en el sujeto i, i=1,…,n, j=1,...,k,
- $\bar{v} = \frac{\sum_{j=1}^{k} n_{jj} c_{jj}}{\sum_{j=1}^{k} n_{jj}}$ es la varianza media y $\bar{c} = \frac{\sum_{j=2}^{k}\sum_{t=1}^{j-1} n_{jt} c_{jt}}{\sum_{j=2}^{k}\sum_{t=1}^{j-1} n_{jt}}$ es la covarianza media,
- $c_{jt} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{it} - \bar{x}_t)}{n_{jt} - 1}$ es la covarianza entre los ítems j y t, j,t=1,…,k,
- $\bar{x}_j = \frac{1}{n_{jj}}\sum_{i=1}^{n} x_{ij}$ es la media del ítem j, j=1,...,k,
- $s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m-1}$ es la varianza de la suma de todos los ítems,
- $x_i = \sum_{j=1}^{k} x_{ij}$, i=1,…,m y $\bar{x} = \frac{1}{m}\sum_{i=1}^{n} x_i$,
- n_{jj} es el número de observaciones válidas correspondientes al ítem j, j=1,...,k,
- n_{jt} es el número de pares válidos (sin valores perdidos) utilizados para calcular la covarianza c_{jt} o la correlación r_{jt}.
- m es el número de observaciones válidas en todos los ítems.

Bibliografía

- Abraira V, Pérez de Vargas A. Generalization of the kappa coefficient for ordinal categorical data, multiple observers and incomplete designs. Qüestiió.1999;23:561-71.

- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;i:307-10.

- Bland JM, Altman DG. Statistical notes: Cronbach's alpha. BMJ. 1997;314:572.

- Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

- Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley & Sons; 1981.

- Lantz CA, Nebenzahl E. Behavior and interpretation of the k statistic: Resolution of the two paradoxes. J Clin Epidemiol. 1996;49(4):431-4.

- Nunnally JC, Bernstein IH. Psychometric theory. 3ª ed. New York: McGraw-Hill; 1994.

- Rosner B. Fundamentals of biostatistics. 5ª ed. Belmont, CA: Duxbury Press; 2000.