Psicothema
2000. Vol. 12, Supl. nº 2, pp. 335-339

ISSN 0214 - 9915 CODEN PSOTEG

Fiabilidad entre observadores con datos categóricos mediante el Anova

José Luis Losada y Jaime Arnau
Universidad de Barcelona

El estudio de la conducta humana requiere, en la mayoría de los casos, instrumentos creados para la situación objeto de estudio (ad hoc). Una característica importante que deben cumplir estos instrumentos es tener alta fiabilidad. En la Metodología Observacional para estudiar la fiabilidad del observador se debe asumir que cada medida se divide en dos partes: una parte conocida o verdadera, y una parte desconocida o error. Cuando los datos son cuantitativos las pruebas sobre los sesgos entre observadores y las medidas de acuerdos entre ellos, se obtienen a partir del modelo ANOVA mixto estándar o a través de los modelos aleatorios. En estos casos la correlación intraclase es el índice de fiabilidad más utilizado. Por ejemplo cuando tenemos varios observadores y se quiere conocer su fiabilidad, generalmente se utiliza el coeficiente *intraclase de Berck* (1979), que detecta la concordancia y el error sistemático de unos observadores respecto a otros (ρ^2). Existen numerosas versiones de la correlación intraclase, y para cada situación específica hay una forma apropiada, aunque conceptualmente todas se centran en el estudio de la fiabilidad. Cuando los datos son categóricos, o cuando la variable de respuesta se clasifica de acuerdo con una escala nominal o multinomial, una medida de acuerdo entre observadores, similar a la correlación intraclase, es el índice *kappa* de Cohen (1960). La presente comunicación plantea el estudio de la fiabilidad entre observadores mediante el índice de Kappa con el procedimiento del ANOVA. Si se verifica que no existe sesgo, la aplicación de un ANOVA unidimensional es suficiente para la estimación del coeficiente. Si, por el contrario, existiese sesgo entre los observadores, se deberán considerar como alternativas el ANOVA bidimensional de efectos aleatorios, o el modelo mixto de dos dimensiones.

Reliability among observers when the data are categorical. The study of the human behavior requires, in most of the cases, instruments created for the situation study object (ad hoc). In the Observational Methodology, to study the observer's reliability it should be assumed that each measure is divided in two parts: a well-known or true part, and an unknown part or error. When the data are quantitative the tests on the biases between observers and the measures of agreements among them, are obtained starting from the pattern standard mixed ANOVA or through the aleatory models. In these cases the correlation intra is the used index of reliability. When the data are categorical, or when the answer variable is classified of a greement with a nominal scale or multinomial, an agreement measure among observers, similar to the correlation intra, it is the index Kappa of Cohen (1960). The present communication outlines the study of the reliability among observers by means of the index Kappa with the procedure of the ANOVA. If it is verified that bias, the application of an ANOVA unidimensional doesn't exist it is enough for the estimate of the coefficient. If, on the contrary, exists bias among the observers, they will be considered as alternative the two-dimensional ANOVA of aleatory effects, or the mixed pattern of two dimensions.

El estudio de la conducta humana requiere, en la mayoría de los casos, instrumentos creados para la situación objeto de estudio (ad hoc). Una característica importante que deben cumplir estos instrumentos es tener alta fiabilidad.

El instrumento debe entenderse como aquel mecanismo representacional a través del cual se obtienen los registros, de tal forma que se puede considerar como instrumento un sistema de categorías, un observador, etc. En la Metodología Observacional, se utilizan términos como fiabilidad del observador, acuerdo entre observadores, que deben diferenciarse de términos como 'estimadores estadísticos', que hacen referencia a índices de grupo y a la exactitud de la precisión de las medidas. Para estudiar la fiabilidad del observador se debe asumir que cada medida se divide en dos partes: una parte conocida o verdadera, y una parte desconocida o error. Cuando los datos son cuantitativos las pruebas sobre los sesgos entre observadores y las medidas de acuerdos entre ellos, se obtienen a partir del modelo ANOVA mixto estándar o a través de los modelos aleatorios. En estos casos la correlación intraclase es el índice de fiabilidad más utilizado. Por ejemplo cuando tenemos varios observadores y se quiere conocer su fiabilidad, generalmente se utiliza el coeficiente *intraclase de Berck* (1979), que de-

Correspondencia: José Luis Losada
Facultad de Psicología
Universidad de Barcelona
08035 Barcelona (Spain)
E-mail: jlosada@psi.ub.es

tecta la concordancia y el error sistemático de unos observadores respecto a otros (ρ^2). Existen numerosas versiones de la correlación intraclase, y para cada situación específica hay una forma apropiada, aunque conceptualmente todas se centran en el estudio de la fiabilidad.

Cuando los datos son categóricos, o cuando la variable de respuesta se clasifica de acuerdo con una escala nominal o multinomial, una medida de acuerdo entre observadores, similar a la correlación intraclase, es el índice *kappa* de Cohen (1960). El índice kappa es un estadístico de concordancia que corrige el azar. Fleiss, Cohen y Everitt (1969) han descrito la distribución de muestreo de kappa.

La evaluación de esta concordancia entre observadores cumple más de una función. Cuando interesa demostrar que los observadores son precisos, se agrupan datos de diferentes tablas de concordancia en una sola tabla, calculando e interpretando un único valor de kappa. De esta forma se obtienen marginales más realistas. Sin embargo, cuando el objetivo es calibrar y entrenar observadores (competencia), el índice kappa debe calcularse individualmente para cada tabla de concordancia.

La matriz de confusión es la estructura más adecuada para controlar los acuerdos y desacuerdos entre dos observadores, pero cuando tenemos más de dos observadores, las posibles combinaciones dos a dos, dificultan este control. La fórmula para el cálculo del índice kappa es

$$K = \frac{\frac{\sum_{i=1}^{k} n_{ii}}{N} - \frac{\sum_{i=1}^{k} n_{i+} \cdot n_{+j}}{N^2}}{1 - \frac{\sum_{i=1}^{k} n_{i+} + n_{+j}}{N^2}}$$

Ecuación 1

siendo n_{ii} las casillas de la diagonal principal de la matriz de confusión,
n_{i+} marginales de fila de la matriz de confusión,
n_{+j} marginales de columna de la matriz de confusión.

Fleiss (1981) caracteriza como regulares los valores de kappa que se hallan entre 0,40 y 0,60, buenos de 0,60 a 0,75, y excelentes por encima de 0,75.

La presente comunicación plantea el estudio de la fiabilidad entre observadores mediante del índice Kappa con el procedimiento del ANOVA. Si se verifica que no existe sesgo, la aplicación de un ANOVA unidimensional es suficiente para la estimación del coeficiente. Si, por el contrario, existiese sesgo entre los observadores, se deberán considerar como alternativas el ANOVA bidimensional de efectos aleatorios, o el modelo mixto de dos dimensiones.

Modelos para el estudio de la fiabilidad

Modelo de efectos aleatorios unidimensional

Una cuestión relevante en Metodología Observacional es sin duda el entrenamiento y competencia de los jueces u observadores que registran y el comportamiento de los sujetos. Supongamos a título de ejemplo, se ha solicitado a cuatro observadores que registren una situación utilizando el mismo sistema de categorías. Para este estudio de fiabilidad inter-observadores, seleccionamos una categoría que reviste cierta dificultad o complejidad para su registro. La codificación utilizada para este caso es la binaria, ocurrencias de la categoría (1) y no ocurrencias de la categoría (0). Además, la sesión se ha dividido en 20 intervalos, para facilitar el registro. Los datos se presentan en la tabla 1.

Un elemento cualquiera de esta tabla y_{ij} denota el registro del i-ésimo intervalo dado por el j-ésimo observador (i=1,2,… n; j=1,2,….. k). Por lo tanto se puede asumir que el modelo para la observación y_{ij} es

$$y_{ij} = \mu + g_i + e_{ij}$$

Ecuación 2

donde μ es la población global de las medidas, g_i es el i-ésimo intervalo; y e_{ij} es el error residual que se asume con una distribución normal de media cero y variancia σ_e^2. La variancia de y_{ij} viene dada por $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$.

Consecuentemente

$$Cov(y_{ij}, y_{i1}) = \sigma_g^2 \quad i=1,2,\ldots..n;\ j \neq 1,2,\ldots k$$

la correlación entre cualquier par de medidas en el mismo intervalo es

Tabla 1
Registro de una categoría por cuatro observadores en 20 intervalos

Intervalo	OB1	OB2	OB3	OB4	TOTAL
1	0	0	0	0	0
2	0	0	1	0	1
3	1	1	1	1	4
4	1	1	1	1	4
5	1	1	0	1	3
6	0	0	0	0	0
7	1	0	0	0	1
8	0	0	0	0	0
9	1	1	1	1	4
10	1	0	1	1	3
11	1	1	1	1	4
12	1	1	0	1	3
13	1	1	0	0	2
14	1	0	1	0	2
15	1	0	0	0	1
16	0	0	1	0	1
17	1	1	1	1	4
18	1	0	0	0	1
19	1	1	1	1	4
20	1	1	1	1	4
Proporción	0.75	0.50	0.55	0.50	46
Total	15	10	11	10	

Tabla 2
Modelo de efectos aleatorios unidimensional

F.V.	g.l.	SS	MS	Error (MS)
Entre	n – 1	$SS_e = \sum_{i=1}^{n} k_i (\bar{y}_i - \bar{y})^2$	$MS_e = \frac{SS_e}{(n-1)}$	$(1+(n_0-1)\ell)\sigma_y^2$
Intra	N – n	$SS_i = \sum_{i=1}^{n} \sum_{j=1}^{k} (y_{ij} - \bar{y}_i)^2$	$MS_i = \frac{SS_i}{(N-n)}$	$(1-\rho)\sigma_y^2$
Total	N – 1	$SS_{Total} = \sum_{i=1}^{n} \sum_{j=1}^{k} (y_{ij} - \bar{y})^2$		

$$p = \frac{cov(y_{ij}, y_{il})}{\sqrt{var(y_{ij})\,var(y_{tl})}} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

Este es el modelo de los componentes de variancia y el ANOVA que corresponde a la ecuación 2 se muestra en la tabla 2.

En la tabla 2, se tiene que

$$N = \sum_{i=1}^{n} k,$$

$$\bar{y}_i = \frac{1}{k_i}\sum_{j=1}^{k} y_{ij},$$

$$\bar{y} = \frac{1}{N}\sum_{i}^{n}\sum_{j}^{k} y_{ij},$$

$$n_0 = \frac{1}{k-1}\left[N - \frac{\sum_{i=1}^{k} n_i^2}{N}\right]$$

La estimación de σ_e^2 y σ_g^2 viene dada respectivamente por $\hat{\sigma}_e^2 = MS_i$ y $\hat{\sigma}_g^2 = (MS_e - MS_i)/n_0$. Por lo tanto se puede definir el estimador del ANOVA de ρ por

$$r_i = \frac{\hat{\sigma}_g^2}{\left(\hat{\sigma}_g^2 + \hat{\sigma}_e^2\right)} = \frac{MS_e - MS_i}{MS_e + (k-1)MS_i}$$

Ecuación 3

Para conocer si existe sesgo o no entre los observadores, y en el caso de datos dicotómicos, resulta adecuado calcular el χ^2 de Cochran, teniendo en cuenta que bajo el supuesto de hipótesis nula de homogeneidad marginal, Q_A es equivalente y se distribuye como un χ^2, con los mismos grados de libertad.

$$Q_A = \frac{k(k-1)\sum_{j=1}^{k}\left(y_j - \frac{y}{k}\right)^2}{ky - \sum_{i=1}^{n} y_i^2}$$

Ecuación 4

Aceptando la hipótesis nula, es decir, que los cuatro observadores tienen registros similares, asumiendo un error del 5% (α=0,05), el cálculo del índice Kappa se realiza como si los registros fuesen datos cuantitativos en lugar de categóricos (valores dicotómicos), donde los cuadrados medios proporcionan una buena estimación utilizando la expresión 3.

Se trata de desarrollar el procedimiento ANOVA que tiene como finalidad detectar hasta que punto los cuatro observadores son fiables. Para ello se ha utilizado el módulo de escalas –análisis de fiabilidad- del paquete estadístico SPSS versión 7.5, y los resultados obtenidos son los siguientes:

Relación e índices descriptivos de los observadores

****** Method 2 (covariance matrix) will be used for this analysis ******

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

		Mean	Std Dev	Cases
1.	OB1	,7500	,4443	20,0
2.	OB2	,5000	,5130	20,0
3.	OB3	,5500	,5104	20,0
4.	OB4	,5000	,5130	20,0

El valor medio más alto corresponde al observado uno (0,7500), en cambio la mayor variabilidad en las observaciones corresponde a los observadores dos y cuatro con 0,5130 en el total de los veinte intervalos.

A continuación se presentan las matrices de covariancia y correlación entre observadores. Evidentemente, cuanto mayor es el coeficiente de correlación y menor en grado de significación entre los observadores, son más fiables.

Matriz de Covariancia

	OB1	OB2	OB3	OB4
OB1	,1974			
OB2	,1316	,2632		
OB3	,0395	,0789	,2605	
OB4	,1316	,2105	,1316	,2632

Matriz de Correlación

	OB1	OB2	OB3	OB4
OB1	1,0000			
OB2	,5774	1,0000		
OB3	,1741	,3015	1,0000	
OB4	,5774	,8000	,5025	1,0000

Estadísticos de la escala total de los observadores

Se presentan los estadísticos de la media de los cuatro observadores. A continuación la media de los valores medios de los observadores, el valor mínimo de estas medias, el máximo, el rango, el cociente entre máximo y mínimo, y la variación de la distribución de medias. También se obtiene la misma información para las variancias de los observadores, para las covariancias y correlaciones entre observadores.

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

N of Cases = 20,0

Statistics for Scale	Mean	N of Variance	Std Dev	Variables
	2,3000	2,4316	1,5594	4

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,5750	,5000	,7500	,2500	1,5000	,0142

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
,2461	,1974	,2632	,0658	1,3333	,0011	

Inter-item Covariances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,1206	,0395	,2105	,1711	5,3333	,0030

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,4888	,1741	,8000	,6259	4,5957	,0449

Resumen de estadisticos observador-total

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
OB1	1,5500	1,6289	,5337	,3801	,7754
OB2	1,8000	1,3263	,7127	,6689	,6845
OB3	1,7500	1,6711	,3789	,2917	,8504
OB4	1,8000	1,2211	,8356	,7351	,6142

La primera columna presenta la media de las puntuaciones totales de los observadores donde en la suma de estas puntuaciones eliminamos el observador correspondiente. Es decir, 1,55 es la media de la variable suma del observador 2 más el observador 3 más el observador 4. La segunda columna son las variancias de esta variable *suma* así obtenida. La tercera columna presenta el coeficiente de correlación de Pearson entre cada observador y el total de observadores, restada de este total la puntuación del observador al que hace referencia el coeficiente. La cuarta columna son los cuadrados de los coeficientes de correlación múltiple entre cada observador y el resto, obtenidos a través de la regresión múltiple y que informa de la capacidad de predicción de la puntuación en un intervalo a partir del resto de observadores, por ejemplo, el 73,51% de la variabilidad de los registros del observador 4 puede ser explicada por el resto de observadores. Finalmente en la quinta y última columna tenemos un índice que informa de lo fiables que son los observadores. Se trata del coeficiente – de Cronbach, que es uno de los más utilizados para establecer la fiabilidad de una escala y está basado en la consistencia interna de la misma. Más concretamente, se obtiene como promedio de los coeficientes de correlación de Pearson entre todos los observadores si las puntuaciones de los mismos están estandarizadas, o como promedio de las covariancias si no lo están. Los valores de este coeficiente oscilan entre 0 y 1 y únicamente obtenemos valores negativos si la relación entre los observadores es negativa, en cuyo caso no procedería plantear la posibilidad de calcular un índice de fiabilidad entre observadores.

Análisis de la variancia

Source of Variation	Sum of Sq.	DF	Mean Square	Q	Prob.
Between People	11,5500	19	,6079		
Within People	8,0000	60	,1333		
Between Measures	,8500	3	,2833	6,3750	,0947
Residual	7,1500	57	,1254		
Nonadditivity	,1864	1	,1864	1,4991	,2259
Balance	6,9636	56	,1243		
Total	19,5500	79	,2475		
Grand Mean	,5750				

Hotelling's T-Squared= ,3124 F= 2,1809 Prob.= ,1277
Degrees of Freedom: Numerator = 3 Denominator = 17

Reliability Coefficients 4 items

Alpha = ,7937 Standardized item alpha = ,7927

En nuestro ejemplo, el α de Cronbach tanto no estandarizada como estandarizada da valores muy parecidos, esto es debido a que los observadores tienen variancias similares.

El test de Hotelling contrasta la hipótesis de si las medias de los observadores son estadísticamente iguales, como así sucede en este caso. La prueba de Tukey comprueba la hipótesis de la existencia o no de interacción multiplicativa entre los observadores.

El cálculo de la fiabilidad finaliza aplicando los valores proporcionados por el ANOVA a la ecuación 6 de tal forma que

$$r_1 = \frac{\hat{\sigma}_g^2}{\left(\hat{\sigma}_g^2 + \hat{\sigma}_e^2\right)} = \frac{MS_e - MS_i}{MS_e + (k-1)MS_i} = \frac{0,6079}{0,6079 + (3)0,1333} = 0,47084421$$

Este resultado es el valor del índice kappa y siguiendo los criterios establecidos por Fleiss (1981), se considera a este valor como una fiabilidad regular entre estos observadores.

Los resultados de este ejemplo sólo se aplican a estimadores de fiabilidad obtenidos para un modelo de efectos aleatorios ONE-WAY.

Modelos alternativos

Modelo de efectos aleatorios bidimensionales

Un elemento cualquiera y_{ij} denota el registro del i-ésimo intervalo dado por el j-ésimo observador (i=1,2,… n; j=1,2,….. k). Se puede asumir que el modelo para la observación y_{ij} en este caso es

$$y_{ij} = \mu + g_i + o_j + e_{ij}$$

Ecuación 5

donde μ es la población global de las medidas, g_i es el i-ésimo intervalo, o_j j-ésimo el observador y e_{ij} es el error residual que se asume con una distribución normal de media cero y variación σ_e^2.

En este modelo se asume que el o_j recoge el efecto aditivo de los observadores seleccionados normalmente con media cero y variancia σ_e^2. Las tres variables g, o, y e son mutuamente independientes, y la variancia de y_{ij} viene definida por

$$var(y_{ij}) = \sigma_g^2 + \sigma_o^2 + \sigma_e^2$$

La covariancia entre dos medidas en el mismo intervalo, tomado el intervalo i -ésima y el observador j -ésimo es

$$Cov(y_{ij}, y_{i1}) = \sigma_g^2$$

La correlación intraclase para calcular la fiabilidad es

$$R = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_o^2 + \sigma_e^2}$$

Ecuación 6

Las estimaciones de los componentes variantes imparciales de σ_g^2, σ_o^2, y σ_e^2, se calculan

$$\hat{\sigma}_g^2 = \frac{MS_{Intervalos} - MS_i}{k}$$

$$\hat{\sigma}_o^2 = \frac{MS_{Observadores} - MS_i}{n}$$

$$\hat{\sigma}_e^2 = MS_i$$

Un estimador de la fiabilidad se formula de la siguiente forma

$$r_2 = \frac{n(MS_{Intervalos} - MS_i)}{n(MS_{Intervalos}) + k(MS_{Observadores}) + (kn - k - n)\,MS}$$

Ecuación 7

fórmula que fue propuesta por Bartko (1966).

Modelo de efectos mixtos bidimensionales

A diferencia del modelo anterior donde se pretendía generalizar los resultados de los observadores de la muestra a un grupo más amplio de observadores, en este modelo sólo nos interesa el grupo de observadores de la muestra.

Siguiente de Fleiss (1986), el y_{ij} se calcula de la siguiente forma

$$y_{ij} = \mu + g_i + o_j + e_{ij}$$

Ecuación 8

Aquí, o_1, o_2, …….o_k, se asume que los efectos son constantes, y $\sum_{t=1}^{n} o_t = 0$.

Los supuestos respecto a g_i y e_j son idénticos a los modelos anteriores. El ANOVA para este caso se presenta en la tabla 3.

En este modelo el índice de fiabilidad de Fleiss (1986) es

Tabla 3
Modelo de efectos aleatorios y mixtos para dos variables ANOVA

F.V.	g.l.	SS	MS	Obs. mixto	Obs. aleatorio
Intervalo	$n-1$	$k\sum(\bar{y}_i - \bar{y})^2$	*MSint.*	$\sigma_e^2 + k\sigma_g^2$	$\sigma_e^2 + k\sigma_g^2$
Observador	$k-1$	$n\sum(\bar{y}_j - \bar{y})^2$	*MSObs.*	$\sigma_e^2 + \frac{n}{k-1}\sum_{i=1}^{k}\sigma_j^2$	$\sigma_e^2 + n\sigma_o^2$
Error	$(n-1)(k-1)$	$\sum\sum(y_j - \bar{y})^2$	MS_i	σ_e^2	σ_e^2
Total					

$$r_3 = \frac{n(MS_{Intervalos} - MS_i)}{n(MS_{Intervalos}) + (k-1)MS_{Observadores} + (k-1)(n-1)MS_i}$$

Ecuación 9

Fleiss (1986) describe el estimador r_3, con las siguientes matizaciones en el procedimiento

1. Probar la variancias de los observadores si difieren significativamente entre si. Para probar esta hipótesis (H_0: $o_1 = o_2 \ldots = o_n = 0$) se debe comparar la proporción $F = MS_e/MS_i$ en la tabla la distribución de la de F (n-1) y (n-1)(k-1) grados de libertad. Aceptar la hipótesis nula implica la ausencia de error entre los observadores, y se puede estimar la fiabilidad aplicando la ecuación 11. Si $F > F_{(n-1),(n-1)(k-1)}$ entonces la hipótesis nula se rechaza y se asume que existen diferencias entre los observadores.

2. Cuando se rechaza la hipótesis nula debe determinarse qué observador u observadores son los responsables de las diferencias en los registros. Si no se incluyen los registros de estos observadores la estimación de la fiabilidad aumentará.

Si por ejemplo, los registros del j -ésimo observador son posiblemente los causantes de las diferencias entre observadores, para comprobarlo se plantea el siguiente contraste

$$L = \hat{y}_j - \frac{1}{k-1}\left(\hat{y}_1 + \ldots \hat{y}_{j-1} + \hat{y}_{j-1} + \ldots \hat{y}_k\right)$$

con un error estándar

$$SE(L) = \left[\frac{kMS_i}{n(k-1)}\right]^{1/2}$$

No se consideran los registros del j-ésimo observador si el valor L/SE(L) es mayor que $|t_{(n-1)(k-1),\alpha/2}|$. En este caso se debería volver a calcular el ANOVA sin el j-ésimo observador y el nuevo coeficiente de fiabilidad utilizando la ecuación 9.

Referencias

Agresti, A. (1990). *Categorical Data Analysis*. New York, NY: Wiley Interscience.

Ato, M., y López, J. (1996). *Análisis estadístico para datos categóricos.* Madrid: Síntesis.

Cohen, J. (1960). *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 20, 37-46.

Demaris, A. (1992). *Logit Modeling: Practical Applications.* Newbury Park, CA: Sage

Fienberg, S.E. (1994). *The analysis of cross-classified categorical data (2nd Ed.)* Cambridge, Ma: MIT Press

Fleiss, J.L. , Cohen, J. & Everitt, B.S. (1969). *Large sample satandard errors of kappa and wighted kappa. Psychologycal Bulletin, 72, 323-327.*