EVALUACIÓN DEL CONSENSO ENTRE EXPERTOS: ÍNDICES DEL TIPO KAPPA, MODELOS Y PROPIEDADES

Carlos N. Bouza[1], Facultad de Matemática y Computación, Universidad de La Habana, Cuba

RESUMEN

La importancia de evaluar el consenso ha ganado en popularidad en muchas aplicaciones. Esto es de particular importancia en muchos estudios biomédicos en los que los expertos juegan un papel importante, por ejemplo, al valorar la eficacia de tratamientos, al clasificar enfermos, entornos de la biosfera o pruebas. En este trabajo se desarrolla una revisión bibliográfica del índice κ, los que han sido deducidos a partir de este y los que pueden ser deducidos como casas particulares bajo ciertos modelos. Se discuten modelos que se han utilizado para soportar inferencias paramétricas o para darle sentido estadístico a la valoración del consenso en la clasificación hecha por varios expertos. Con fines ilustrativos se desarrollan varios ejemplos.

ABSTRACT

In many applications the importance of evaluating the consensus is growing. This is particularly important in many biomedical studies where the role of the experts is crucial. That is the case when the efficacy of treatments, the classification of patients, the quality of the biosphere in certain places or tests is studied. This paper is devoted to the development of a review of the κ index, indexes deduced or deducible from it as particular cases under certain models. Properties of them for different models are discussed. The models support parametric inferences or provide a statistical frame for analyzing the consensus in the classification among experts. Different examples are developed with illustrating purposes.

Key words: Agreement, model based inference, consensus.

MSC: 91B12

1. INTRODUCCIÓN

Este trabajo tiene por objeto hacer una revisión de la literatura existente-referida al índice kappa y a índices relacionados. Esto ha sido motivado por el trabajo realizado alrededor de aplicaciones en el área biomédica en las cuales ha sido necesario trabajar con el consenso para modelar ciertos fenómenos. Esto es muy común en muchas otras áreas de aplicación de la estadística como los estudios de mercado. Hemos tratado de unificar las notaciones existentes y de agrupar los resultados de acuerdo a criterios y enfoques del tratamiento teórico y aplicado.

Primeramente se introducen las ideas básicas. Posteriormente se discute el caso más usual en las aplicaciones: el dicotómico. Este aparece en problemas como los de aplicar o no un tratamiento y en dos grupos similares. Esto es muy común en experimentos clínicos. Posteriormente se trata el modelo con múltiples respuestas: el multinomial. Aquí se presentan y analizan los modelos que sustentan los razonamientos que llevan a usar nuevamente índices relacionados con el kappa. Otras secciones replantean el problema del consenso a partir de usar modelos. Estos son generados por hipótesis como es el caso de la aleatoriedad, la homogeneidad y al utilizar un punto de vista tomado de la teoría de los modelos superpoblacionales muy usados en el Muestreo contemporáneo como una alternativa de tipo Bayes. El desarrollo de inferencias basadas en la normal es el tema de las últimas secciones. Dado que las hipótesis que permiten desarrollar una teoría basada en la normalidad no son de frecuente cumplimiento en las aplicaciones; los métodos inferenciales se asocian a la teoría de las muestras grandes.

Se discuten algunos ejemplos hipotéticos con fines ilustrativos.

Las referencias listadas cubren el tema discutido.

2. ALGUNAS IDEAS BÁSICAS

Consideremos el caso en que interesa la opinión de los clientes. Se hace una encuesta y se indaga sobre que opina sobre el producto. Por ejemplo este es el caso de una empresa de perfumería que lanza una nueva loción al mercado. Encuesta a posibles consumidores a los que se les pregunta:

E-mail: [1]bouza@matcom.uh.cu

P1. ¿Le agrada la loción?

P2. ¿Es su costo asequible para Ud.?

Si hacemos P1 a una parte de los muestreados y P2 a otra no podremos establecer si hay una concordancia entre las valoraciones. Si se evalúa la política que debemos tomar respecto a las acciones de un portafolio las preguntas:

P3. ¿Que acciones vendemos?

P4. ¿Que acciones compramos?

tienen un sentido parecido a P1 y P2 pero podremos utilizar varios expertos y nos puede interesar indagar si la propensión a la venta y la compra son razonablemente similares. Nuestro deseo sería saber si los expertos en bolsa consultados están de acuerdo: si hay un consenso suficientemente alto sobre la acción a tomar.

El problema de medir el acuerdo es sustancialmente diferente del de la asociación, dada su naturaleza subjetiva. Este se asocia al establecer en que medida un experto o criterio de clasificación concuerda con otro u otros. Esto es muchas veces conocido como "fiabilidad entre los clasificadores". Otro problema es el de establecer la concordancia entre la clasificación para los mismos objetos bajo diferentes condiciones. Entonces se busca medir la "fiabilidad de la clasificación-reclasificación".

El uso de una medida de consenso es necesario en estudios en los que es de interés establecer cuan reproducibles son ciertos resultados al utilizar un método alternativo de evaluación. Si uno de los clasificadores es "perfecto" podremos confiar en las valoraciones de otro que tiene un alto nivel de concordancia con él. Tal es el caso cuando en la encuesta sobre la loción utilizamos como testor del perfume una lámina de cartón impregnada en esta para evaluar P1. El que diseñó el experimento espera que haya un alto grado de concordancia entre el aroma percibido en el testor, por un individuo, y el que en realidad tiene. Por tanto, la valoración, dada por los encuestados que huelan la lámina debe ser similar a la que darían si utilizaran el líquido directamente.

La noción asociada al concepto de concordancia o consenso está dada por el que se obtenga un mismo resultado usando diferentes expertos, métodos o criterios. Note que la existencia de una asociación no implica que necesariamente exista un consenso. Indudablemente hay una correlación entre el aroma de la lámina impregnada con la loción y el del perfume. El problema del consenso es medir en que grado la valoración que hagan los encuestados al usar el testor o la loción directamente es similar.

Veamos un ejemplo ilustrativo.

Ejemplo 1. Se entrevistan 4 personas y se les solicita que califiquen en una escala del 1 al 10 los filmes épicos ganadores de Oscar:

1. *Gladiador*
2. *Quo Vadis*
3. *Ben Hur*
4. *Sinohué el egipcio*
5. *Cleopatra*

Las respuestas fueron:

Entrevistado	**Filme épico evaluado en la encuesta**				
	Gladiador	*Quo Vadis*	*Ben Hur*	*Sinohué*	*Cleopatra*
1	1	2	3	8	7
2	1	2	2	10	4
3	1	2	6	9	10
4	1	2	3	9	10

Podemos concluir del análisis de los resultados que:

- Hay un consenso perfecto al evaluar las 2 primeras producciones.
- Hay una alta asociación entre la primera y la tercera persona, pero no un consenso en la clasificación.
- Hay una asociación aceptable entre la primera y la cuarta y un nivel de consenso en la clasificación mayor que entre la primera y la tercera.
- La asociación y el consenso para la cuarta y la tercera es mejor que la existente entre la cuarta o la tercera con la primera.
- No hay ni asociación entre la primera y la segunda persona y tampoco un buen consenso en la clasificación.

La existencia de consenso esta dada en función de coincidencias en la clasificación.

Las discrepancias no son aceptadas aunque su causa sea un sesgo, una escala o un criterio de selección. Por ello solo se observó un consenso perfecto al analizar *Gladiador* y *Quo Vadis*, en las que todos Ie dan una calificación de 1. Sin embargo podemos considerar la existencia de un sesgo pues las técnicas filmográficas existentes al hacer los filmes son diferentes. Por ejemplo *Ben-Hur*, que trata un tema similar a *Gladiador* pudo ser afectada por esto.

La búsqueda del consenso en la clasificación se basa en el supuesto de que hay k clases disjuntas en las que el objeto i puede ser clasificado y J clasificadores (criterios, expertos, etc.). Cada i se asocia a m variables $Y_j^i,...,Y_J^i$, i = 1,...,n. Los objetos conforman una población y se supone que la distribución es invariante para cualquier permutación {1,...,m}. Si cada clasificador da respuesta a una misma pregunta las discrepancias, bajo la hipótesis de acuerdo, no es sino un error y esto lo modelamos al fijar como respuesta del individuo j a $Y_j^i = X^i + \varepsilon_{ij}$, j = 1,…,J, i = 1,…,n. ε es un error que puede ser considerado aleatorio o sistemático.

3. ALGUNOS ÍNDICES EN EL CASO DICOTÓMICO

En la medicina es frecuente que el médico, experto, trate de clasificar en un par de clases excluyentes a los pacientes sometidos a un tratamiento: mejora o no mejora. En problemas de mercado este hecho es también corriente: el agente económico analiza un producto, un paquete de acciones etc. y trata de calificarle como "positivo" o "negativo".

Asumiendo que la clasificación es realizada usando el método de "doble a ciegas", lo que implica independencia, se trabaja con una variable que solo toma dos posibles valores

$$Y_t = \begin{cases} + & \text{si} \quad t \in A \\ - & \text{si} \quad t \notin A \end{cases}$$

Se valora cada objeto t, usando dos criterios o dos expertos, y se obtienen dos valores de Y_t. Cada objeto se clasifica como "positivo" o "negativo". Al analizar los reportes de los dos expertos sobre el n objetos los resultados se pueden reflejar en una tabla como la Tabla 3.1.

Tabla 3.1. Clasificación de n objetos por dos expertos.

	Experto 1		
Positivo	n_{11}		
Negativo	n_{21}		
Total	n_{+1}		

Note que este sería el mismo enfoque cuando cruzamos dos preguntas dicotómicas. Si por ejemplo indagamos sobre el precio y la calidad de un producto identificamos positivo con bueno y negativo con malo. Por tanto n_{11} identificaría a aquellos entrevistados que dicen que el producto tiene un buen precio y una calidad adecuada.

Tenemos que no solo n_{11} es un valor importante pues también lo son los valores de las otras casillas y sobre todo funciones de ellas. Diferentes índices han sido propuestos para evaluar este tipo de problemas. Ellos hacen uso de las proporciones

$$p_{ij} = \frac{n_{ij}}{n}, \qquad q = \frac{1}{2}(p_{2+} + p_{+2}), \qquad q = \frac{1}{2}(p_{+1} + p_{1+}), \qquad \text{donde} \qquad q_{i+} = \frac{n_{i+}}{2},\ p_{+i} = \frac{n_{+i}}{n}$$

De una otra forma.

Es razonable que el interés se centre en un criterio asociado a las concordancias existentes entre aquellos considerados como positivos por al menos uno de los expertos.

Este enfoque es el usual en la Taxonomía Numérica en la que n_{22} no es tomado en cuenta. Uno de esos índices es el propuesto por Dice (1945)

$$I_D = \frac{2p_{11}}{p_{1+} + p_{+1}} = \frac{p_{11}}{p} \tag{3.1}$$

que es la probabilidad empírica condicionada de concordancia entre los expertos dada la positividad del criterio asociado a la casilla "positivo-positivo". El criterio opuesto establece el uso de

$$I'_D = \frac{2p_{22}}{p_{2+} + p_{+2}} = \frac{p_{22}}{q} \tag{3.2}$$

Por su parte Rogot-Goldberg (1966) propusieron usar la suma de estos dos índices

$$I_{RG} = \frac{p_{11}}{p} + \frac{p_{22}}{p} = I_D + I'_D \tag{3.3}$$

y

$$I'_{RG} = \frac{I_{RG}}{4} \tag{3.4}$$

Para estos índices el valor cero determina un perfecto desacuerdo de los expertos y el valor 1 un perfecto acuerdo. Cuando el acuerdo observado es equivalente al de un experimento aleatorio (3.4) es igual a 0,5.

En estos índices se ha tomado como principio el uso de medidas de posición- si se usaran criterios de dispersión se puede USaf el índice de Armitage

$$I^2_A = \frac{n}{n-1}\left[p_{11} + p_{12} - (p_{11} - p_{22})^2\right] \tag{3.5}$$

En el caso de desacuerdo perfecto este es cero y lo fuese el acuerdo toma el valor n/n - 1. Armitage-Blendys-Smyl (1966) reescalan este índice proponiendo

$$I'_A = \frac{p_{11} + p_{12} - (p_{11} - p_{22})^2}{1 - (p - q)^2} \tag{3.6}$$

Uno de los más populares índices es de Kruskal (1954)

$$I_K = \lambda_r = \frac{(p_{11} + p_{22}) - q}{p} \frac{2p_{11} - (p_{12} + p_{21})}{2p_{11} + (p_{12} + p_{21})} = 2I_b - 1 \tag{3.7}$$

Este se mueve en [-1, 1]. Si el acuerdo es perfecto $\lambda_r = 1$ si es perfecto. Si $n_{11} = 0$ el índice toma el valor -1 lo que es un severo problema al hacer ciertos análisis si el objetivo del evaluador está en los objetos calificados de positivos por al menos un experto.

El índice Scott (1955) es

$$I_S = \frac{4(p_{11}p_{22} - p_{12}p_{21}) - (p_{12}p_{21})^2}{(p_{1+} + p_{+1})(p_{2+} + p_{+2})} \tag{3.8}$$

4. MODELOS MULTINOMIALES DE RESPUESTA

A partir de tener dos expertos y k clases tomando

n_{ij} = número de individuos clasificado en C_i por el experto 1 y en C_j por el experto 2 y

π_{ij} = Prob (clasificar a un individuo en $C_i \cap C_j$)

podemos utilizar el modelo log-lineal

$$\log(n_{ij}) = \mu + \lambda_i^1 + \lambda_j^2 + \delta(i,j)$$

ver Agresti (1990), donde

$$\delta(i, j) = \begin{cases} \delta_i & \text{si } i = j \text{ y } i = 1,...,k \\ 0 & \text{si no} \end{cases}$$

Este es el modelo de quasi-independencia en el que se modela la independencia fijando $\delta_i = \delta$ para todo i. El acuerdo básico se expresa a través de μ, λ_i^1 y λ_j^2 miden el acuerdo y δ el acuerdo que se espera tener más allá de la independencia de los expertos.

Siguiendo los criterios usuales podemos estudiar también el ajuste dado por los modelos:

$$\log(n_{ij}) = \mu + \lambda_i^1 + \lambda_j^2 + \beta e_i e_j$$

(donde $e_i < e_j$ si i < j son los escores relacionados con C_i y C_j) y

$$\log(n_{ij}) = .\mu + \lambda_i^1 + \lambda_j^2 + \beta e_i e_j + \delta(i,j)$$

Ejemplo 4.1. Estos modelos fueron usados para analizar el acuerdo entre la calificación dada por 350 televidentes al calificar su opinión sobre la programación de los fines de semana de una tele emisora y la valoración que hacen de ella en su conjunto. Se definieron 7 clases

C_1 = Muy buena C_2 = Buena C_3 = Aceptable C_4 = Regular

C_5 = Mala C_6 = Muy Mala C_7 = No la veo.

Del estudio de los datos obtuvimos

Tabla 4.1. Resultados del análisis del ejemplo 4.1.

Modelo	Clasificación	G^2	Grados de libertad
$\log(n_{ij}) = \mu + \lambda_i^1 + \lambda_j^2 + \delta(i, j)$	Quasi-independencia	115,81	35
$\log(n_{ij}) = \mu + \lambda_i^1 + \lambda_j^2 + \delta$	Independencia diagonal	61,08	36
$\log(n_{ij}) = \mu + \lambda_i^1 + \lambda_j^2 + \beta e_i e_j + *\delta$	Independencia y asociación uniforme	4,05	34

Por tanto, el modelo adecuado para describir el modelo es el tercero y este es el que debe usarse para describir el acuerdo entre los encuestados.

La utilidad de tales modelos estriba en la posibilidad de valorar cuales son los efectos reflejados y planear campañas. En este caso el modelo que se ajusta a los resultados permite adgüir que la evaluación de la programación de fin de semana es independiente de la que se Ie da a la emisora como un todo. Por tanto, si los analistas opinan que es mala, pueden excluir de ello el estudio de esta programación. Sin embargo, la calificación dada a ambas preguntas si se asocia. En el modelo aceptado hay efectos de columna y aleatorios en el consenso de base (base-line) en la clasificación obtenida.

Está claro que la hipótesis de independencia al hacer las clasificaciones es poco realista al abordar problemas consensuales. En este ejemplo los televidentes no pueden aislar, como dice el modelo, su percepción de la calidad de la programación televisiva de la que tengan de la de fin de semana. Sin embargo, hacer el ajuste del modelo permitirá que se valore cual es el efecto que tiene en la calificación de la programación total.

Por ello, el estudio del consenso no puede ser restringido al ajuste de modelos.

5. EL MODELO ALEATORIO INDEPENDIENTE

Podemos considerar que los expertos hacen su valoración usando un mecanismo aleatorio independiente. Esta hipótesis fija que

$$\text{Prob}\{t \in C_i \text{ para experto 1 y } t \in C_j \text{ para experto2}\} = \pi_{ij} = \pi_{i+}\pi_{+j}$$

donde π_{i+} es la probabilidad marginal de que el experto 1 ubique un objeto en C_j y π_{+j} lo es de que lo haga experto 2 en C_j. Al analizar un indice I* sin usamos este modelo al calcular su esperanza condicional:

$$E(I^* \mid \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j = 1,\ldots,k) = I_0 \tag{5.1}$$

la que establece que

$$p_{ij} = p_{i+}p_{+j}$$

Entonces como 1 - I_0 es el máximo de las diferencias obtenibles respecto a una clasificación aleatoria una medida del acuerdo es el índice tipificado

$$M(I^*) = \frac{I^* - I_0}{1 - I_0} \tag{5.2}$$

Cuando el acuerdo observado es mayor que el achacable a un fenómeno aleatorio $M(I^*) > 0$, si es menor $M(I^*) < 0$ y será igual a cero si son equivalentes. Es conveniente acotar el hecho de que si $I_0 = 0{,}5$ entonces $\text{Min}\{M(I^*)\} = -1$. Por tanto, este es un índice que es deducido al aceptar la existencia de un modelo superpoblacional que fija que

$$E(n_{ij}) = E(n_{i+})E(n_{+j})$$

Esta esperanza no tiene nada que ver con el diseño muestral utilizado.

Un índice popular el llamado "kappa"

$$\kappa = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} p_{i+}p_{+i}}{1 - \sum_{i=1}^{k} p_{i+}p_{+i}} \tag{5.3}$$

debido a Cohen (1960). Este compara las frecuencias relativas observadas en la muestra.

Utilizando este marco podemos decir que el K-indice es un estimador ingenuo de

$$\kappa_{pob} = \frac{\sum_{i=1}^{k} \pi_{ii} - \sum_{i=1}^{k} \pi_{i+}\pi_{+i}}{1 - \sum_{i=1}^{k} \pi_{i+}\pi_{+i}} \tag{5.4}$$

En general el número de posibles calificaciones de un objeto pueden ser $k \geq 2$. En particular no es recomendable hacer preguntas con más de 5 alternativas. Pero esta es una recomendación asociada a la práctica pues para valores relativamente "grandes" de k los entrevistados tienden a verse desestimulados a ser "exactos".

Podemos analizar la esperanza de los índices más populares bajo este modelo de aleatoriedad.

Para (3.1)

$$E(I_D \mid \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j = 1,\ldots,k) = \frac{p_{1+}p_{+1}}{p}$$

y el índice tipificado es

$$M(I_D) = \frac{2(p_{11}p_{22} + p_{12}p_{21})}{p_{1+} + p_{+2} + p_{+1} + p_{+2}}$$

que es igual a (3.3).

Para I_{RG} tenemos

$$E(I_{RG} \mid \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j = 1,\ldots,k) = \frac{p_{1+}p_{+1}}{p} + \frac{p_{2+}p_{+2}}{q}$$

por lo que nuevamente obtenemos que

$$M(I_{RG}) = \kappa$$

Similarmente a partir de que

$$E(I_A \mid \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j = 1,\ldots,k) = \frac{p_{1+}p_{2+}}{1-(p-q)^2} + \frac{p_{1+}p_{+2}}{1-(p-q)^2}$$

el índice tipificado correspondiente

$$M(I_A) = \kappa$$

Fleiss (1975) destacó el hecho de que el índice κ caracteriza la clase de índices

$$T = \left\{ I \,\middle|\, I = \frac{I^* - I_\Omega}{Max(I^* - I_\Omega)} \right\}$$

donde $I_\Omega = E(I^* \mid \Omega)$.

Estos hechos soportan la popularidad del índice propuesto por Cohen aunque en la práctica se utilice muchas veces uno de los introducidos en la sección 2. Razones históricas, entre las que incluyen la costumbre, hacen que en ciertas áreas se utilicen uno de ellos en vez del κ-índice.

6. OTRA VISITA AL MODELO SUPERPOBLACIONAL

Al usar la consideración de que los expertos clasifican en forma aleatoria independiente los objetos permite usar como un modelo superpoblacional a

$$Y_{ij}^k = Y_{ij} + \varepsilon$$

$$Y_{ij}^k = \begin{cases} 1 & \text{si se clasifica en } C_i \cap C_j \\ 0 & \text{si no} \end{cases}$$

Esta es una variable con distribución $B(\pi_{ij} = \pi_{i+}\ \pi_{+j})$. Donde

$$\pi_{ij} = P(Y_{ij} = 1)$$

$$\pi_{i+} = P(Y_{ij} = 1 \mid j = 1)$$

$$\pi_{i+j} = P(Y_{ij} = 1 \mid i = 1)$$

Por tanto, los índices analizados serian predictores.

Usando κ este lo es del parámetro superpoblacional

$$\kappa_{pob} = \frac{2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{\pi_{1+}\pi_{+2}\pi_{+1}\pi_{2+}}$$

Kraemer (1979) por su parte propuso el modelo en el que un objeto es fijo y hay 2 expertos (criterios). Haciendo

$$E(Y_i) = P(Y_i = 1) = p_i$$

$$P(Y_i = 0) = 1 - p_i$$

Este es un modelo inherente al objeto. Modelando la superpoblación tomamos

$$p_i = \pi + \varepsilon_i$$

El parámetro condicionado tiene una esperanza

$$E(p_i) = \pi$$

$$V(p_i) = \sigma_\pi^2$$

bajo el supuesto de que $E(\varepsilon) = 0$.

Un índice del tipo κ inspirado por este modelo es:

$$\kappa_\rho = \frac{\sigma_\pi^2}{\pi(1-\pi)}$$

Cada experto o criterio es aplicado al objeto i y este es clasificado como positivo o negativo por cada uno. A partir de la independencia entre los criterios utilizados este modelo permite describir el comportamiento del proceso de calificación a partir de la Tabla 6.1.

Tabla 6.1. Modelo basado en la independencia.

Criterio o Experto 1	**Criterio o Experto 2**		**Marginal**
Positivo	$E(p_1^2)$	$E(p_1p_2)$	π
Negativo	$E(p_2p_1)$	$E(p_2^2)$	$1 - \pi$
Marginal	π	$1 - \pi$	

Ver Bloch-Kraemer (1989) y Hole-Fleiss (1993). Note que

$$\kappa_\rho = \frac{E(p_1^2 \mid M_\rho) - \pi^2}{\pi(1-\pi)}$$

Nos interesa estimar los parámetros que realmente están siendo generados por el proceso bajo el modelo superpoblacional. La distribución es una trinomial y

$$E(p_{11}\,|\,M\rho) = E(p_1^2\,|\,M\rho) = \pi^2 + \kappa\pi(1-\pi)$$

$$E(p_{22}\,|\,M\rho) = E(p_2^2\,|\,M\rho) = (1-\pi)^2 + \kappa\pi(1-\pi)$$

$$E(p_{12}\,|\,M\rho) = E(p_{21}\,|\,M\rho) = (1-\kappa\rho)\pi(1-\pi)$$

Un estimador de κ es obtenible al aplicar el método de Máxima Verosimilitud. En este caso

$$\ln(L,\ \kappa_\rho \mid n_{11},n_{12},n_{21},n_{22}) = n_{11}\ln(\pi^2 + \kappa_\rho\pi(1-\pi)) + n_{12} + n_{21})\ln((\pi(1-\pi)(1+\kappa_\rho)) + n_{22}\ln((1-\pi)^2 + \kappa_\rho\pi(1-\pi)$$

Hallando las derivadas respecto a ρ y π se obtiene que es estimador máximo verosímil de κ_ρ es

$$\hat{\kappa}_1 = \frac{4p_{11}p_{22} - (p_{12}+p_{21})}{(2p_{11}+p_{12}+p_{21})(2p_{22}+p_{12}+p_{21})}$$

y

$$\hat{\pi} = \frac{2p_{11}+p_{12}+p_{21}}{2}$$

lo es del parámetro de localización del modelo superpoblacional. Note que este índice tipo κ es el propuesto por Scott (1955) y es llamado coeficiente κ-intraclase y pertenece a .

7. EL CASO DE LAS PROBABILIDADES MARGINALES IGUALES

Krauth (1984) utilizó como hipótesis de base

$$H: \pi_{i+} = \pi_{+i} \qquad \forall i = 1,...,k$$

para trabajar con la marginalidad propuso el índice

$$\kappa_\pi = \frac{\sum_{i=1}^{k}\pi_{ii} - \frac{1}{2}\left(\sum_{i=1}^{k}\pi_{i+}^2 + \sum_{i=1}^{k}\pi_{+i}^2\right)}{1-\sum_{i=1}^{k}\pi_{i+}\pi_{+i}} = \frac{\sum_{i=1}^{k}\pi_{ii} - \left(\sum_{i=1}^{k}\pi_{i+}\pi_{+i}\right)}{1-\sum_{i=1}^{k}\pi_{i+}\pi_{+i}} - \frac{\frac{1}{2}\left(\sum_{i=1}^{k}\pi_{i+}-\pi_{+i}\right)^2}{1-\sum_{i=1}^{k}\pi_{i+}\pi_{+i}} = \kappa_{kr} - \lambda_\pi$$

Al utilizar los usuales estimadores de las probabilidades envueltas

$$\hat{\pi}_{ij} = p_{ij} = \frac{n_{ij}}{n}$$

$$\hat{\pi}_{i+} = p_{i+} = \frac{n_{i+}}{n}$$

$$\hat{\pi}_{i+} = p_{+i} = \frac{n_{i+}}{n}$$

tenemos el estimador ingenuo

$$\kappa_p = \frac{\sum_{i=1}^{k}p_{ii} - \left(\sum_{i=1}^{k}p_{i+}p_{+i}\right)}{1-\sum_{i=1}^{k}p_{i+}p_{+i}} - \frac{\frac{1}{2}\left(\sum_{i=1}^{k}p_{i+}-p_{+i}\right)^2}{1-\sum_{i=1}^{k}p_{i+}p_{+i}} = \frac{P_0 - P_e - P_h}{1-P_e} = \frac{P_1}{P_2}$$

Este no pertenece a la clase y puede tener un sesgo grande.

Bouza (1987) propuso un estimador corregido para el sesgo. Este es

$$\kappa_B = \frac{P_1 + \frac{1}{2n}\left(\sum_{i=1}^{k} P_{i+}(1-P_{i+}) + \sum_{i=1}^{k} P_{+i}(1-P_{+i})\right)}{P_2}$$

La preferencia de uno u otro va a depender de la preferencia del estadístico.

En todos los casos los errores son determinados asintóticamente.

8. EL MODELO DE HOMOGENEIDAD

Tomemos X_{ijh} como la variable que establece la evaluación de un individuo i por el experto j en el estudio h, i = 1,...,n; j = 1,...,J; h = 1,...,H. Un modelo polinomial es fijado asignándole $\pi_h = P(X_{ijh} = 1)$, $X_{ijh} = 1$ establece el "éxito" en la clasificación. Entonces usando el modelo usual de correlación bajo la aceptación de que no hay sesgo de experto π_h es constante en h. Fijando que si j = 1,2:

$$P_{1h}(\kappa_h) = P(X_{i1h} = X_{i2h} = 1) = P(S_1) = \pi_h^2 + \kappa_h\pi_h(1-\pi_h) \quad (8.1)$$

$$P_{2h}(\kappa_h) = P(X_{i1h} \neq X_{i2h}) = P(S_2) = 2\kappa_h\pi_h(1 - \pi_h) \quad (8.2)$$

$$P_{3h}(\kappa_h) = P(X_{i1h} = X_{i2h} = 0) = P(S_3) = (1- \pi_h)^2 + \kappa_h\pi_h(1 - \pi_h) \quad (8.3)$$

donde κ_h representa la correlación entre X_{i1h} y X_{i2h}.

Denotando por n_{1h}, n_{2h} y n_{3h} el número de individuos en que se observan los sucesos S_1, S_2 y S_3 en el estudio h y $n_{1h} + n_{2h} + n_{3h} = nh$, el número total de evaluaciones es

$$n = \sum_{h=1}^{H} n_h$$

y podemos estimar la probabilidad de interés mediante

$$\hat{\pi}_h = \frac{2n_{1h} + n_{2h}}{2n_h} \quad (8.4)$$

siendo la solución del sistema de ecuaciones

$$\hat{\kappa}_{hDE} = 1 - \frac{n_{2h}}{2n_h\hat{\pi}_h(1-\hat{\pi}_h)}$$

Si se utiliza la ponderación

$$W_h = \frac{n_h\hat{\pi}_h(1-\hat{\pi}_h)}{\sum_{h=1}^{H} n_h\hat{\pi}_h(1-\hat{\pi}_h)}$$

La medida del tipo K para los H estudios realizados por 2 expertos es

$$\hat{\kappa}_{DE} = 1 - \frac{n_{2+}}{2\sum_{h=1}^{H} n_h\hat{\pi}_h(1-\hat{\pi}_h)}$$

donde

$$n_{2+} = \sum_{h=1}^{H} n_{2h}$$

Las medidas de consenso de Fleiss (1971) y Landis-Koch (1977) son un caso particular cuando $n_1 = n_{2h} \ldots = n_h$.

Podemos ilustrar algunos de los problemas presentados al evaluar el diseño de la publicidad de una línea de perfumería.

Ejemplo 8.1. Una perfumería propuso tres combinaciones de envase-lema publicitario para la nueva fragancia. Los resultados fueron:

Tabla 8.1. Resultados de la encuesta sobre la nueva fragancia.

Lema 1	**Envase 1**		
Aceptación	Si	No	Total
Si	5	6	11
No	5	54	59
Total	10	60	70
Lema 2			
Aceptación	Si	No	Total
Si	6	4	10
No	8	40	48
Total	14	44	58
Lema 1			
Aceptación	Si	No	Total
SI	3	4	7
No	3	33	36
Total	6	37	43

Por lo que

	h = 1	**h = 2**	**h = 3**
$\hat{\kappa}_{hDE}$	0,38	0,50	0,36
$\hat{\pi}_h$	0,15	0,21	0,15
n_h	70	58	43

y $\hat{K}_{DE} = 0{,}375$.

Consideremos que un individuo puede ser clasificado en una de k categorías mutuamente excluyente y que las variables dicotómicas independientes $X_1 \ldots, X_k$, son tales que para todo i:

$$P(X_t = 1 \mid \pi) = n$$

$$P(Xt = 0 \mid \pi) = 1 - \pi$$

Al no variar π con t se establece que los clasificadores no prefieran dar una valoración de X_t por razones de la complejidad del análisis o dificultad de evaluar a t. Por su parte $P(X_t = 1 \mid \pi)$ modela que si un individuo es similar a otro no es necesario hacer una labor adicional al evaluarle.

La distribución conjunta de $(X_1, \ldots, X_k)$, ver de Finetti (1931), es

$$P(X_1 = x_1, \ldots, X_k = x_k) = \int_0^1 (1-\pi)^{k-H} \pi^H \partial F(\pi)$$

al considerar la distribución "a priori" de π y que

$$S_t = \sum_{i=1}^{k} X_i = H$$

Denotando por

$$\mu_t = \int_0^1 \pi^q \partial F(\pi),\ t = 1,\ldots$$

$$P(S_t = r) = \binom{k}{r} \int_0^1 (1-\pi)^{k-r} \pi^r \partial F(\pi) = \binom{k}{r} \Delta^{k-r} \mu_r \qquad r < k$$

$$P(S_t = k) = \int_0^1 \pi^k \partial F(\pi) = \mu_k$$

es la probabilidad de que todos los clasificadores asignen X_t =1 al individuo t.

Tomemos

$$\Delta\mu_r = \mu_r - \mu_{r+1},$$

y

$$N = \{F \mid \mu_t(F) = \mu \quad si \quad t < h\}$$

siendo

$$\mu_h^+ = Max\{\mu_t(F),\ F \in N\}$$

$$\mu_h^- = Min\{\mu_t(F),\ F \in N\}$$

entonces

$$p_h = \frac{\mu_h - \mu_h^-}{\mu_h^+ - \mu_h^-}$$

es llamado el momento canónico de orden h si $p_h \in]0,1[$, ver Lau (1991).

Los primeros tres momentos pueden ser calculados, al fijar q_h =1 - p_h, por

$$\mu_h = p_1$$

$$\mu_2 = p_1(p_1 + q_1 p_2)$$

$$\mu_3 = p_1(p_1 + q_1 p_2) + p_2 q_1(p_1 + q_1 p_2 + q_2 p_3)$$

y

$$p_1 = \mu_1$$

$$p_2 = \frac{\mu_2 - \mu_1^2}{\mu_1(1-\mu_1)}$$

$$p_3 = \frac{(\mu_3\mu_1 - \mu_2^2)(1-\mu_1)}{(\mu_2 - \mu_1^2)(\mu_1 - \mu_2)}$$

Usando la familia de índices del tipo κ

$$\kappa = \frac{V(\pi)}{E(\pi)(1-E(\pi))} = p_2$$

Note que en la muestra cada individuo t se asocia a un vector $X_t \in]\,0,\,1[^k$ y

$$n = \sum_{t=1}^{k} n_t$$

y $(n_0, n_1,...,n^k)$ sigue una instrucción polinomial $P(n, \Delta^k \mu_0, \Delta^{k-1}\mu_1,\ldots, \Delta^0\mu_k)$ por lo que el logaritmo de la función de verosimilitud es:

$$L = \text{constante} + \sum_{t=0}^{k} n_t \ln(\Delta^{k-1}\mu_t)$$

y los estimadores máximo verosímiles son obtenidos fácilmente como

$$\Delta^{k-t}\mu_t \hat{=} \frac{n_t}{n}$$

por lo que

$$\hat{\mu}_j = \sum_{t=j}^{k} \frac{\binom{t}{j}}{\binom{k}{j}} \frac{n_t}{n},\ j \le k$$

Este modelo permite obtener estimadores para κ para el caso en que los individuos puedan ser clasificados en varias clases.

Para k = 2

$$\hat{p}_1 = \frac{n_2 + \frac{n_1}{2}}{n}$$

$$\hat{p}_2 = \hat{\kappa}^{(2)} = \frac{\frac{n_2}{2} - \hat{p}_1^2}{\hat{p}_1(1-\hat{p}_1)}$$

Si k = 3

$$\hat{p}_1 = \frac{3n_3 + 2n_2 + n_1}{3n}$$

$$\hat{p}_2 = \hat{\kappa}^{(3)} = \frac{\frac{3n_3 + n_2}{3n} - \frac{3n_3 + 2n_2 + n_1}{(3n)^2}}{\frac{(3n_3 + 2n_2 + n_1)\,(3n - 3n_3 - 2n_2 - n_1)}{9n^2}}$$

Veamos como funciona esto en un ejemplo

Ejemplo. 8.2. Se toma una muestra de turistas recién llegados y se les entrevista. Estos establecen si visitarán la playa, si harán uso de las ofertas culturales y si alquilarán autos. Los resultados fueron

Tabla 8.2. Resultados de la encuesta del ejemplo 8.2.

1	2	3	Número
Playa	Cultura	Autos	Total
No	No	No	2381
No	No	Si	57
No	Si	No	50
No	Si	Si	42
Si	No	No	61
Si	No	Si	28
Si	Si	No	39
Si	Si	Si	342

Como se ve

$n_0 = 2381,$

$n_1 = 168,$ $\hat{p}_1 = 0{,}16$

$n_2 = 123$ $\hat{p}_2 = 0{,}78$

$n_3 = 342$ $\hat{p}_3 = 0{,}53$

El cálculo de los estimadores para los pares está dado por el correspondiente

$$\hat{\kappa}_{ji} = \frac{P(X_j = 1 | X_i = 1) - \text{Min}\, P(X_j = 1 | X_i = 1)}{\text{Max}\, P(X_j = 1 | X_i = 1) - \text{Min}\, P(X_j = 1 | X_i = 1)}$$

que en el ejemplo son

$$\hat{\kappa}_{12} = 0{,}80, \qquad \hat{\kappa}_{13} = 0{,}79, \qquad \text{y} \qquad \hat{\kappa}_{23} = 0{,}89.$$

Por tanto, los planes de hacer uso del auto e ir a la playa caracterizan bien en consenso entre las tres variables. Por su parte la selección de hacer turismo cultural y alquilar a un auto son los criterios con un consenso más fiable pues toma el valor máximo: 0,89.

Estos métodos permiten guiar la publicidad, así podemos inferir que ofertas con precios preferenciales para paquetes culturales que oferten un auto en alquiler deben ser bien acogidas.

En muchas ocasiones nuestro interés se conecta con lo que ocurre al condicionar respecto al resto de las variables. Esto se puede modelar a través de las sumas S_t, como

$$P(X_3 = 1 | X_1 = X_2 = 1) = P(X_3 = 1 | S_2 = 2)$$

se asocia al consenso en la tercera variable cuando el mismo resultado ocurrió en el resto de las variables evaluadas. A partir de fórmulas generales podemos deducir con simplicidad que

$$P(X_3 = 1 | S_2 = 2) = p_1 q_1 p_2 + \frac{q_1 q_2 p_2 p_3}{p_1 + q_1 p_2}$$

$$P(X_3 = 1 | S_2 = 1) = p_1 q_2 + p_2 q_3$$

$$P(X_3 = 1 | S_2 = 0) = \frac{p_1 q_2 (q_1 q_2 + p_2 p_3)}{q_1 + p_1 p_2}$$

las que son calculables s si conocemos los momentos

$$M_k = \frac{P(X_{k+1}|S_k = r) - \text{Min}P(X_{k+1}|S_k = h)}{\text{Max}P(X_{k+1}|S_k = r) - \text{Min}P(X_{k+1}|S_k = h)}$$

con

$$M_k = \begin{cases} p_k & \text{si} \quad k-h \quad \text{es par} \\ q_k & \text{si} \quad k-h \quad \text{is impar} \end{cases}$$

Estimándoles convenientemente usando estimadores ingenuos obtenidos al insertar los estimadores máximo verosímiles de los parámetros envueltos. Su consistencia es garantizada.

En el ejemplo anteriormente desarrollado tenemos

Tabla. 8.3. Estimación de las probabilidades condicionadas del ejemplo 8.2.

		$P(X_{k+1}\|S_k = h)$	
k		H	
	2	1	0
2	0,91	0,07	0,02
1	-	0,92	0,08
0	-	-	0,90

De estos resultados obtenemos que los turistas al hacer actividades culturales el consenso con ir a la playa y alquilar un auto es de 0,91 y con no hacer ninguna de las dos cosas sean un poco menor si no las hace: 0,90.

9. AJUSTE CON VARIABLES

Cada individuo puede estar acompañado de un vector de variables auxiliares

$$X_{-i} = (1, X_{i1} \ldots X_{ih})^T$$

Asumimos un modelo conocido

$$X_{i1} = X_{i2} = 1$$

$$P(i \mapsto + +) = P(C_1)\,\pi_i^2 + \rho\pi_i(1 - \pi_i)$$

$$X_{i1} = X_{i2} = 0$$

$$P(i \mapsto - -) = P(C_2) = (1 - \pi_i)^2 + \rho\pi_i(1 - \pi_i)$$

$$X_{i1} \neq X_{i2}$$

$$P(i \mapsto + \vee -) = P(C_3) = 2(1 - \rho)\pi_i(1 - \pi_i)$$

y que el modelo logit

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = (1, x_{i1}, \ldots, x_{ih})(\beta_0, \beta_1, \ldots, \beta_h)^T$$

linkea la probabilidad con la información auxiliar. Tomemos

$$z_i = \begin{cases} x_{-i} & \text{si} \quad t = 1 \\ 0 & \text{si} \quad t = 2 \\ -x_{-i} & \text{si} \quad t = 3 \end{cases}$$

$$s_i = \begin{cases} 1 & \text{si} \quad t = 1 \\ -2 & \text{si} \quad t = 2 \\ -1 & \text{si} \quad t = 3 \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{si} \quad i \text{ es clasificado en } C_j \\ 0 & \text{si} \quad \text{no} \end{cases}$$

La función de verosimilitud de este problema

$$L = \prod_{i=1}^{n} \frac{e^{x_i\beta}}{(1-e^{x_i\beta})^2} (e^{x_i\beta} + \rho)^{Y_{i1}} (2(1-\rho))^{Y_{12}} (e^{-x_i\beta})^{Y_{13}}$$

es resuelta usando software de la regresión logit como el de Barlow (1996) para ajustar un modelo, hacer predicciones y evaluar la importancia de las variables auxiliares.

10. ERRORES APROXIMADOS Y NORMALIDAD

Como hemos visto los estimadores del tipo κ son muy importantes. Estos son del tipo razón por lo que el cálculo de su esperanza y varianza requiere de fuertes hipótesis para que estas sean deducibles con cierta facilidad. Una de ellas es aceptar que

$$E\left(\frac{x}{y}\right) \approx \frac{E(x)}{E(y)}$$

Un método siempre utilizable es desarrollo de las Series de Taylor requeridas. Entre ellas está el método de Funatsu (1982) en el que solo se pide que $E(y) > 0$, $E(x) > 0$ y que $P(x) > 0) = 1$.

Estas son satisfechas en el caso de los índices del tipo K. El desarrollo es

$$E\left(\frac{x}{y}\right) \approx \frac{E(x)}{E(y)}\left[1 + \left(\frac{1}{t}\right)\sum_{i\geq 1}\left(-\frac{1}{tE(y)}\right)^i \left(\frac{\sigma_{i1}(t)}{E(x)}\right) - \frac{\sigma_{0i+1}(t)}{E(y)}\right]$$

donde

$$\sigma_{hk}(t) = E[(x + E(x))^k (y - E(y)^h)]$$

Donde t es un acelerador de la convergencia. Tomar t = 1 simplifica esta fórmula.

Aceptar que los términos de orden mayores que 2 son despreciables es lo usual en estos estudios.

Por ejemplo del modelo trinomial donde dedujimos el estimador

$$\hat{\kappa}_1 = \frac{4p_{11}p_{22} - (p_{12} + p_{21})}{(2p_{11} + p_{12} + p_{21})(2p_{22} + p_{12} + p_{21})}$$

y

$$\hat{V}(\hat{\kappa}_1) = \frac{1-\hat{\kappa}_1}{n}\left[(1-\hat{\kappa}_1)(1-2\hat{\kappa}_1) + \frac{\hat{\kappa}_1(2-\hat{\kappa}_1)}{2\hat{\kappa}_1(1-\hat{\kappa}_1)}\right]$$

es la estimación de su varianza al sustituir κ por su estimador en los términos considerados como significativos en el desarrollo de la Serie de Taylor.

Si le usamos para el caso de la homogeneidad a través del modelo de Donner **et al**. (1996)

$$\hat{\kappa}_{hDE} = 1 - \frac{n_{2h}}{2n_{2h}\hat{\pi}_h(1-\hat{\pi}_h)} \tag{10.1}$$

siendo

$$\hat{V}(\hat{\kappa}_{hDE}) = \frac{1-\hat{\kappa}_{hDE}}{n_h}\left[(1-\hat{\kappa}_{hDE})(1-2\hat{\kappa}_{hDE}) + \frac{\hat{\kappa}_{hDE}(2-\hat{\kappa}_{hDE})}{2\hat{\pi}_h(1-\hat{\pi}_h)}\right] \tag{10.2}$$

para h = 1,...,H. Entonces

$$\hat{V}(\hat{\kappa}_{DE}) = \sum_{1\le h\le H} W_h^2 \hat{V}(\hat{\kappa}_{hDE})$$

Por su parte para

$$\kappa_p = \frac{P_0 - P_e - P_h}{1-P_e}$$

$$\hat{V}(\kappa_p) = \frac{\sum_{i=1}^{k}\left(P_{ii}\left(1-P_e-(P_{+i}+P_{i+})(1-P_0+P_e)\right)\right)^2 + \sum_{i\neq j}\left(P_{ij}\left((P_{+i}+P_{+j})(P_0-P_e-P_h)-(P_{+i}+P_{j+})\right)\right)^2}{(1-P_e)^4 n} -$$

$$\frac{\sum_{i\neq j}\left(\left((P_0P_e - 2P_e - 2P_h + P_0)\right)\right)^2}{(1-P_e)^4 n}$$

Bajo las condiciones de regularidad usuales se garantiza la convergencia a la normal necesaria para hacer inferencias. En general el estudio de los mercados se centra en fijar una estimación del "consenso" entre los clientes para la homogeneidad. Por ello hacer inferencias no es lo más común sino describir pues el objetivo es detectar y no generalizar. Además al usar expertos se usan paneles de pequeño tamaño y con creciente popularidad se trabaja con "tanques pensantes".

11. ALGUNAS PRUEBAS DE HIPÓTESIS

Al estimar las probabilidades $P_{ih}(\kappa_h)$ definidas por (8.1)-(8,3) insertando las estimaciones de κ y $\hat{\pi}_h$ obtenemos el estimador ingenuo $\hat{P}_{ih}(\hat{\kappa})$. Las ideas presentes en las pruebas de la Bondad del Ajuste son utilizables. Tomando

$$\hat{e}_{ih} = n_h \hat{P}_{ih}(\hat{\kappa}_{ih})$$

el estadígrafo chi-cuadrado es

$$\chi^2_{DE} = \sum_{h=1}^{H}\sum_{i=1}^{3}\frac{(n_{th}-\hat{e}_{th})^2}{\hat{e}_{th}}$$

que sigue una distribución chi-cuadrado cuando los $n_h's$ son suficientemente grandes.

Eso permite hacer pruebas sobre si el κ-estimador describe convenientemente el proceso de acuerdos. Veamos que podemos decir del ejemplo 8.1.

Ejemplo 11.1. Usando los resultados del eiemplo 8,1 obtenemos que:

Tabla 11.1. Probabilidades deducidas de los datos del ejemplo 8.1.

		$\hat{P}_{ih}(\hat{\kappa})$	
	1	2	3
1	0,07	0,11	0,07
2	0,16	0,22	0,16
3	0,77	0,67	0,77

Como $\chi^2_{DE} \approx 1{,}1$ podemos aceptar.

Usando el desarrollo en Series de Taylor, Fleiss (1981) podemos obtener pruebas basadas en la normalidad. Esto lo podemos ejemplificar usando (10,1) y (10.2).

Una prueba se asocia a utilizar

$$U_{DE} = \sum_{i=1}^{H} \frac{(\hat{\kappa}_h - \hat{\kappa}_{DE})^2}{\hat{V}(\hat{\kappa}_{DE})}$$

que es un estadígrafo con distribución chi-cuadrado con H - 1 grados de libertad bajo las usuales hipótesis.

REFERENCIAS

AGRESTI, A. (1988): "A model for agreement between ratings on the ordinal scale", **Biometrics**, 44, 539-548.

AGRESTI, A. and J. LANG (1993): "Quasi-symmetric L class models with application to rater agreement", **Biometrics**, 49, 131-139.

AGRESTI, A. and M. YANG (1986): "An empirical investigation of some effects of sparness in contingency tables, **Computational Sta. and Data Anal**. 5, 9-21.

ARMITAGE, P.; L.M. BLENDYS and H.C SMYLLIE (1966): "The measurement of observer disagreementi n the recording of signs", **J. Royal Sta. Soc**. Series A, 129, 96-109.

BARLOW, W.; M.Y. LIU and S.P. AZEN (1996): "A comparison of methods for calculating a stratified kappa", **Statistics in Medicine**. 10, 1465-1472.

BAKER, S.G.; L.S. FREEDMAN and M.K.B. PARMAR (1991): "Using replicate observations in observer agreement studies with binary assessments", **Biometrics**, 47, 1327-1338.

BECKER, M. and A. AGRESTI (1992): "Loglinear modeling of pairwise interobserver agreement on a categorical scale", **Statistics in Medicine**, 11, 101-114.

BLECH, D.A. and H.C. KRAEMER (1989): "2x2 kappa coefficient: measures of agreement and assessment", **Biometrics**, 45, 269-287.

BOUZA, C.N. (1987): "A ratio estimator of the kappa index of agreement between two observers", **Biometrical J**., 29, 1011-1015.

CICHETTI, D. V. and A.R. FEINSEIN (1990): "High agreement but low kappa: II. Resolving he paradoxes", **J. of Clinical Epidemiology**. 43, 551-5558.

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

COHEN, J. (1968): “Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit”, **Psychological Bull**., 70, 213-220.

DARROCH, J.N. and P.I. McCLOUD (1986): “Category distinguishability and observer agreement”, **Australian J. of Sta**. 28, 29-42.

DAVIES, M. and J.L. FLEISS (1982): “Measuring agreement for multinomial data”, **Biometrics**, 38, 1047-1051.

DE FINETTI, B. (1931): “Funzione caratteritica de un fenomeno aleatorio”, **Atti delle R. Acad. Nazionale dei Lincii**. Ser. 6, 251-299.

DONNER, A.; M. ELIASZUO and N. KLAR (1996): “Testing the homogeneity of kappa stratified”, **Biometrics**, 52, 176-183.

_____ (1992): “A goodness of fit approach to inference procedures for the kappa stratified: confidence interval construction, significance tests and sample size estimates”, **Statistics in Medicine**, 11, 1511-1519.

FLEISS, J. (1975): “Measuring agreement between two judges on the presence in the absence of a trait”, **Biometrics**, 31, 651-659.

_____ (1971): “Measuring a nominal scale agreement among many raters”, **Psychological Bull**. 76, 378-382.

_____ (1965): “Estimating the accuracy of dichotomous judgements”, **Psychometrika**, 30, 469-479.

FLEISS, J.C.; J. COHEN and B.S. EVERITT (1969): “Large sample statistical errors of kappa and weighted kappa”, **Psychological Bull**. 72, 323-327.

FLEISS, J.C. and M. DAVIES (1982): “Jackknifing function of multinomial frequencies with an application to a measure of concordance”, **American J. of Epidemiology**, 115, 841-845.

FLEISS, J.; J.C. NEE and J.R. LANDIS (1979): “The large sample variance of the kappa in the case of different sets of raters”, **Psychological Bull**. 86, 974-977.

FORMAN, A. (1993): “Fixed latent class models for the analysis of sets of 2-way contingency tables”, **Biometrics**, 49, 511-521.

FUNATSU, Y.A. (1982): “A method for deriving valid approximate expressions for the bias in ratio”, **J. Stat. Planning and Inference**, 6, 210-225.

HOLE, C.A. and J.L. FLEISS (1993): “Interval estimates under 2 study designs for kappa with binary classification”; **Biometrics**, 49, 523-524.

KRAEMER, H.C. (1980): “Extension of kappa coefficient”, **Biometrics**, 36, 207-216.

_____ (1992): “Measurement of reliability for categorical data in medical research”, **Statistical Methods in Medical Res**. 1, 183-200.

_____ (1979): “Ramifications ofa population model for K as a coefficient of reliability”, **Psychometrika**, 44, 461-472.

KRAUTH, J. (1984): “Multivariate Behandlungstabilitat. Klinisher Skalen”, **Psycol. Beitrib**. 23, 438-457.

LAU, T.S. (1993): “Higher order kappa type statistics for the deichotomus attribute in multiple ratings”, **Biometrics**, 49, 535-542.

LEE, J.J. and Z.N. TU (1994): “A better confidence of interval for the kappa on measuring agreement between two raters with binary outcome”, **J. of Computational and Graphical Sta**t. 3, 301-321.

LIPSITZ, S.R.; N.M. LAIRD and T.A. BREMEN (1994): “Simple moments estimators of the kappa coefficent and its variance”, **Applied Stat**. 43, 309-323.

ROGOT, E. and I.D. GOLBERG (1966): “A proposed index for measuring agreement in test-retest studies”, **J. Chronic Dis**.19, 991-1006.

ROSNER, B. (1982): “Statistical methods in ophtalmology: an adjustment for the intraclass correlation between eyes”, **Biometrics**, 38, 899-905.

SCHOUTEN, H.J.A. (1986): “The kappa coefficient of agreement among observers”, **Psychometrika**, 51,453-466.

SCOTT, W.A. (1955): “Reliability of content analysis: the case of nominal scale coding”, **Public Opinion Quarterly**.

SHOUKRI, M.M.; S.W. MARTIN and I.U.H. MIAAN (1995): “Maximum likelihood estimation of the kappa coefficient from models of matched binary responses”, **Statistics in Medicine**, 124, 83-99.

TANNER, M.A. and M.A. YOUNG (1985): “Modeling agreement among raters”, **J. Amer. Stat. Ass**. 80, 175-180.

UBERSAX, J.C. (1993): “Statistical modeling of expert ratings in medical treatment appropiateness”, **J. Amer. Stat. Ass**. 88, 321-427.

UBERSAX, J.C. and W.M. GROVE (1993): “A latent trait finite mixture model for the analysis of rating agreement”, **Biometrics**, 823-835:

VON EYE, A. and S. SORENSEN (1991): “Model chance when measuring interrater agreement with kappa”, **Biometrical J**., 33,781-787.

ZWICK, R. (1988): “Another look at the interrater agreement”, **Psychological Bulletin**, 103, 374-378.