ARCHIVOS LATINOAMERICANOS DE NUTRICION
Organo Oficial de la Sociedad Latinoamericana de Nutrición
Vol. 42 Nº 2, 1992

Estandarización de personal en diagnóstico clínico de bocio: ¿Cómo evaluar la concordancia entre examinadores de la tiroides?

Jorge Matute[1] *y Erick Boy*[2]

Instituto de Nutrición de Centro América y Panamá (INCAP), Guatemala

RESUMEN. El uso de las pruebas t de Student y Cochran para evaluar medidas de tendencia central, así como la prueba de F para evaluar la variabilidad, y el análisis de correlación, han sido usados en forma equivocada para evaluar concordancia entre examinadores del tiroides. En el presente artículo se presenta la prueba de Kappa intraclase (Bloch y Kraemer, 1989), así como la experiencia de su uso en Centroamérica para estandarizar a los examinadores del tiroides que participaron en las encuestas nacionales para determinar prevalencia de bocio en 1990.

SUMMARY. Personnel standarization in clinical diagnosis of goiter: ¿How to evaluate concordance between thyroid examiners? The use of the t Student and Cochran tests to evaluate central tendency measures, as well as the F test to evaluate variability and correlation analysis, have been incorrectly used for the evaluation of concordance between thyroid examiners. This paper presents the intraclass Kappa test (Bloch & Kraemer, 1989), as well as the experience of its use in Central America to standarize thyroid examiner personnel who participated in national surveys carried out during 1990 to determine goiter prevalence.

INTRODUCCION

Cuando se llega el momento de evaluar si las personas entrenadas para realizar el diagnóstico de bocio por inspección-palpación lo hacen en igual forma que la persona que los entrenó, los investigadores y capacitadores se enfrentan con el problema de seleccionar la herramienta estadística más conveniente. La evaluación de concordancia (acuerdo, o reproducibilidad) es un tema que surgió en los años 50, pero no ha sido, sino hasta este último quinquenio que la metodología estadística ha desarrollado nuevas formas de evaluación más exactas; entre ellas tenemos los trabajos de Bloch y Kraemer (1), I-Kuci Lin L. y Aickin M.(2, 3).

Por lo regular el procedimiento estadístico para evaluar concordancia no se encuentra en los textos de estadística corrientes y, en algunos casos se encuentra en libros más avanzados, pero sin correcciones que la literatura moderna (como las mencionadas arriba) han incorporado.

Tradicionalmente la concordancia se ha evaluado utilizando medidas de tendencia central (i.e.: promedio y mediana), variabilidad y correlación, así como pruebas tales como la t de Student, Wilcoxon, y la de Cochran. Por ejemplo, Maclennan y colaboradores(4) utilizan el análisis de varianza para evaluar las diferencias de diagnóstico de bocio entre examinadores. Así mismo, hace algunos años en el Instituto de Centro América y Panamá (INCAP), para evaluar la concordancia de la palpación de tiroides entre el experto y los aprendices, se usaba la prueba de Cochran (con la cual se evaluaba la similitud entre prevalencias de bocio obtenidas por un observador y el estándar).

DIFERENCIAS ENTRE CONCORDANCIA Y ASOCIACION

La prueba de Cochran no mide el grado de acuerdo entre el experto y el personal entrenado, por lo que sus resultados no pueden aplicarse para evaluar si el nuevo personal está o no estandarizado. Veamos un caso extremo: un examinador novato con una precisión pobre (poco acuerdo consigo mismo) puede producir prevalencias similares a las de su maestro, pero clasificando como

1 Biostadístico de INCAP
2 Investigador de la División de Nutrición y Salud del INCAP

bociosos a niños que para el experto son normales y viceversa. Esta prueba no evalúa la coincidencia entre experto y novato, en cuanto a lo que interesa en un ejercicio de estandarización: verdaderos positivos, verdaderos negativos, falsos positivos, y falsos negativos. En un ejemplo como el citado, la prueba es sesgada para medir concordancia.

Concordancia y correlación son dos cosas muy diferentes que generalmente tienden a confundirse. Bloch y Kraemer (1) presentan un ejemplo sencillo que permite aclarar lo anterior: Cuatro diferentes jueces (estándar, juez 1, juez 2 y juez 3) califican a cinco sujetos (A,B,C,D, y E). Se supone que los jueces contestan la misma pregunta y evalúan con la misma escala.

Los resultados de la Tabla 1 indican que únicamente el Juez 1 posee acuerdo con el Juez Estándar. Si se procediera a realizar las pruebas de t de Student (pareada) para evaluar si los promedios son semejantes entre el Juez Estándar y cada uno de los otros jueces, el resultado sería que con el Juez 3 los promedios son diferentes (con los Jueces 1 y 2 no se puede realizar la prueba 't' porque la varianza es cero, por lo mismo se puede inferir que el promedio del Juez 1 es igual al del Estándar, no así del Juez 2). Si se evalúa la igualdad de las varianzas, todos los jueces poseen la misma variabilidad que el Juez Estándar (ver Tabla 1). Si se evalúa la correlación entre el Juez Estándar y los jueces 2 y 3, el coeficiente de Pearson, r, es igual uno. No es posible omitir el sesgo dado por el Juez 2, o el cambio en la escala, que presenta el Juez 3. Al hacer caso omiso de dicho sesgo o cambio en la escala, se confunde asociación con concordancia. En consecuencia, la concordancia no se puede, ni se debe medir a través de la estadística inferencial comúnmente usada para comparar medidas de tendencia central, variabilidad o correlación.

¿COMO MEDIR CONCORDANCIA?

Bloch y Kraemer (1), son claros en presentar los principios para evaluar concordancia:

Que los evaluadores contesten la misma pregunta. Por ejemplo, si un médico pregunta "¿Este niño tiene bocio? y otro médico pregunta ¿Este niño va a la escuela?, un "si" o un "no a ambas preguntas no significa acuerdo, como tampoco un "si" a la primera y un "no" a la segunda significa desacuerdo. No puede

TABLA 1
RESULTADO DEL EXAMEN

Sujeto	Juez Estandar	Juez 1	Juez 2	Juez 3
A	1	1	2	2
B	2	2	3	4
C	3	3	4	6
D	4	4	5	8
E	5	5	6	10
Promedio	3	3	4	6
Desviación Estándar	1,58	1,58	1,58	3,16
Prueba F para igualdad de varianzas		F= 1,P>0,05	F= 1,P>0,05	F= 4,P>0,05
Promedio de la diferencia *		0	1	3
D.E de la diferencia		0	0	1.58
t		-	-	t= 4,243
Valor P		-	-	P= 0,0132

* Promedio de la diferencia con respecto al Juez Estándar.

existir acuerdo o desacuerdo entre respuestas a preguntas diferentes.

Que los evaluadores usen la misma escala para evaluar. Por ejemplo, se les puede hacer la misma pregunta a dos médicos: "¿Es el nivel de yoduria de este niño de 20?". Si un médico está usando microgramos de yodo por gramo de creatinina y otro usa microgramos de yodo por decilitro de orina, entonces un "sí" o un "no" de los dos puede significar desacuerdo, mientras que un "sí" de uno y un "no" del otro pueden darse cuando realmente hay acuerdo respecto a la cantidad de yoduria en el niño.

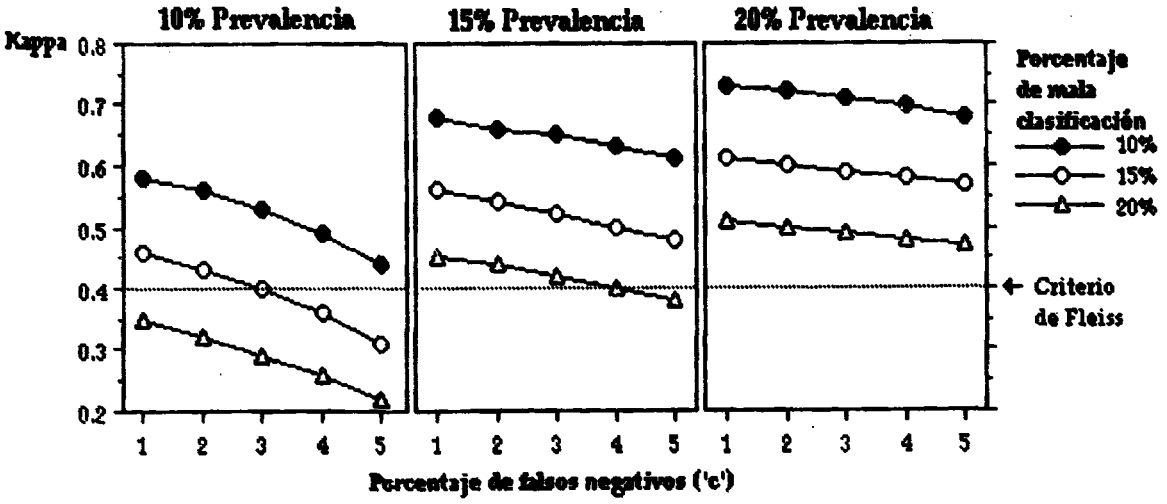
Los estadísticos que miden concordancia son Kappa Intraclase (1) para variables discretas, y el Coeficiente de Correlación de Concordancia (2) para variables continuas. Para evaluar la estandarización de los examinadores de tiroides en las últimas encuestas realizadas en Centroamérica (1990) hemos utilizado Kappa Intraclase, bajo el criterio de Fleiss (5) para establecer el grado de acuerdo, el cual es más estricto que el criterio presentado por Landis y Koch (6) (ver Tabla 2).

Más adelante recomendamos que los lugares donde se realicen los ejercicios de estandarización tengan como mínimo un 10% de prevalencia. Usando este criterio y deseando un porcentaje bajo de mala clasificación en las personas examinadas, el criterio de Fleiss es muy "permisivo" para tomar la decisión de que realmente existe concordancia entre dos jueces, además este criterio no toma en cuenta el hecho de que el valor Kappa cambia de acuerdo a la prevalencia y al porcentaje de mala clasificación. Esto se puede ver en la Gráfica 1, donde a mayor prevalencia sube el valor de Kappa (ver la tendencia de los tres paneles), así como a mayor porcentaje de mal clasificados se obtiene un Kappa menor. El criterio sugerido por Fleiss tolera una elevada proporción de mala clasificación cuando las prevalencias son altas.

Con base en los resultados presentados en la Gráfica 2, aceptando entre un 10% y un 15% de mala clasificación, y asumiendo una prevalencia entre 10% y 20%, sugerimos que se use el criterio presentado en la Tabla 3 para establecer el grado de acuerdo.

TABLA 2
CRITERIO DE FLEISS PARA ESTABLECER EL GRADO DE ACUERDO

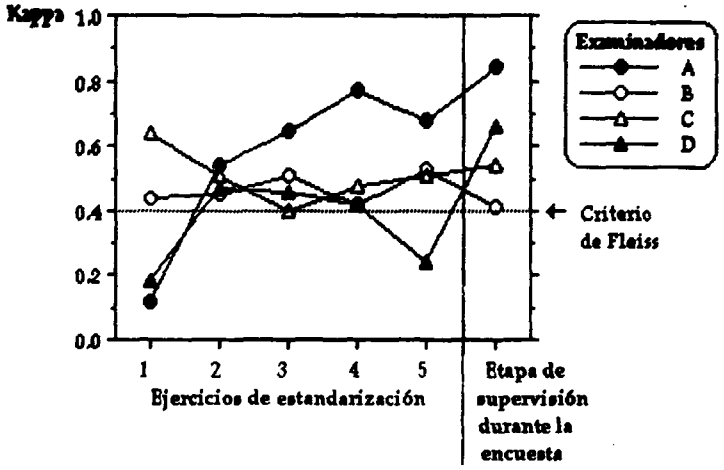
Valor de Kappa	Interpretación
<0,40	No hay acuerdo
0,40-0,75	Acuerdo intermedio (aceptable)
>0,75	Buen acuerdo (excelente)



GRAFICA 1
VALORES DE KAPPA CON DIFERENTES PREVALENCIAS Y PORCENTAJES DE MALA CLASIFICACIÓN. A mayor prevalencia, el valor de Kappa sube (ver la tendencia de los tres paneles). Valores de Kappa altos con una prevalencia alta, aceptan un mayor porcentaje de mal clasificados usando el criterio de Fleiss.

TABLA 3
CRITERIO PARA ESTABLECER EL GRADO DE ACUERDO DE LOS VALORES DE KAPPA

Prevalencia 10%-15%		Prevalencia de 15%-20%		
10% mal clas.	15% mal clas.	10% mal clas.	15% mal clas.	Interpretación
<0,50	<0,40	<0,55	<0,50	No hay acuerdo
0,50-0,60	0,40-0,50	0,55-0,65	0,50-0,60	Acuerdo inaceptable
>0,60	>0,50	>0,65	>0,60	Acuerdo aceptable



GRAFICA 2
RESULTADOS DE KAPPA EN EL PROCESO DE ESTANDARIZACION DE EXAMINADORES DE LA ENCUESTA PARA LA PREVALENCIA DE BOCIO A NIVEL NACIONAL EN EL SALVADOR, 1990.
De esta figura se deduce que el examinador 'D' no tenía capacidad para producir resultados confiables al finalizar la capacitación. Así mismo, se puede observar que los examinadores 'B' y 'C' requerían una mayor supervisión, dado que al finalizar la capacitación su grado de acuerdo con el estándar apenas cumplía con el requisito mínimo de aceptabilidad utilizado en ese entonces.

METODOLOGIA EMPLEADA EN EL INCAP

A continuación se presenta la metodología que utilizó el INCAP para evaluar la estandarización del nuevo personal en sus últimas encuestas nacionales (El Salvador 1990 y Panamá 1990): una vez impartida la parte teórica para el personal en entrenamiento (no necesariamente médico), se realizaron ejercicios consecutivos de estandarización, examinando únicamente entre 30 y 50 sujetos en cada ejercicio para evitar fatiga, y malos hábitos semilógicos. Los examinadores se ubicaron en posiciones que no permitían conocer diagnósticos hecho por sus compañeros vecinos o el estándar. Posteriormente los examinadores re-evalúan a todos los sujetos cuyos diagnósticos (bocio si, bocio no) no concuerdan con el diagnóstico del estándar. Los ejercicios se realizan en grupos con prevalencia de bocio igual a, o mayor del 10% y se continuaron hasta que el índice de Kappa Intraclase, de todos los examinadores, llegó a un valor aceptable en dos pruebas consecutivas (>0.40, según Fleiss). La metodología anteriormente expuesta, así como el estadístico usado (Kappa Intraclase) resultaron sumamente valiosos, para la estandarización del personal que realizó las encuestas en 1990, tanto para fines docentes como para motivar en el personal la autocrítica necesaria, que les permitió superarse en la mayoría de los casos y en algunos otros permitió a los coordinadores de la encuesta identificar al personal que no podía ser utilizado en la recolección de datos. Al personal en entrenamiento se le presentaba los resultados en forma gráfica (ver Gráfica 2). De la Gráfica 2 se deduce que el examinador 'D' no estaba en capacidad de producir

resultados confiables al finalizar la capacitación. Así mismo, en la Gráfica 2 se puede observar que los examinadores 'B' y 'C' requerían una mayor supervisión, dado que al finalizar la capacitación, su grado de acuerdo con el estándar apenas cumplía con el requisito mínimo de aceptabilidad utilizado en ese entonces. La experiencia derivada de estas encuesta ha modificada los criterios utilizados para predecir el grado de acuerdo aceptable entre Estándar y examinadores en entrenamiento. De la Gráfica 2 se puede derivar que mientras Kappa acusa un grado de acuerdo aceptables según Fleiss con prevalencias altas, resulta pragmáticamente útil elevar el punto de quiebre para establecer acuerdo.

Cabe resaltar que este indicador de concordancia (Kappa Intraclase), también se utilizó posteriormente durante la supervisión de la encuesta, para conocer si después de algún tiempo había algún cambio en la concordancia de los examinadores, con respecto al experto. Para el caso con el examinador 'D', Kappa Intraclase permitió que se tomara la decisión de hacer el diagnóstico tiroideo por acuerdo entre dos examinadores, evitando el sesgo que dicho examinador hubiera cometido de haber trabajado sólo. Cuando se trata de conocer la concordancia entre un evaluador recién entrenado y la persona estándar, utilizando Kappa Intraclase (Ki), esta se puede calcular fácilmente de la siguiente manera (1):

Partimos de una tabla 2 x 2:

		Diagnóstico del Examinador Estándar: bocio	Diagnóstico del Examinador Estándar: no bocio
Diagnóstico del Examinador en Entrenamiento	bocio	a	b
	no bocio	c	d

n = total de sujetos observados = a + b + c + d

$$P = \text{Probabilidad de acuerdos positivos} = \frac{2a + b + c}{2n}$$

$$Ki = \text{Kappa intraclase} = \frac{4(ad-bc)-(b-c)^2}{(2a + b + c)(2d + b + c)}$$

Var(Ki) = Varianza de Kappa intraclase =

$$= \frac{(1-Ki)\left[(1-Ki)(1-2Ki) + \frac{Ki(2-Ki)}{2P(1-P)}\right]}{n}$$

$IC_{95\%}$ = Intervalo de confianza al 95% para Ki=

$$= Ki \pm 1.96\sqrt{var(Ki)}$$

EJEMPLO

		Diagnóstico del Examinador Estándar: bocio	Diagnóstico del Examinador Estándar: no bocio
Diagnóstico del Examinador en Entrenamiento	bocio	30	8
	no bocio	11	81

N= 130

$$P = \frac{2*30 + 8 + 11}{2*130} = \frac{79}{2*130} = 0.303846$$

$$Ki = \frac{4(30*81-8*11) - (8-11)^2}{(2*30+8+11)(2*81+8+11)} = \frac{9359}{14299} = 0.6545213$$

Esto quiere decir que la concordancia del examinador recién entrenado con la del examinador estándar es aceptable. Pero es conveniente calcular el intervalo de confianza de este estimador, para observar si el limite inferior todavía entra entre la categoría de aceptable:

$$Var(Ki) = \frac{X + \frac{0.6545213(2-0.6545213)}{2*0.3038462(1-0.3038462)}}{130}$$

donde $X = (1-0.6545213)(1-0.6545213)(1-2*0.6545213)$

$$Var(Ki) = \frac{0.6822863}{130} = 0.0052284$$

Intervalo de confianza al 95% para Ki

$$0.6545213 \pm 1.96\sqrt{0.0052484} = (0.512528, 0.7965146)$$

Intervalo que se interpreta, como que el verdadero valor de Kappa Intraclase para concordancia puede ser desde 0.512 hasta 0.796. Observamos que el valor inferior todavía se encuentra en la categoría de aceptable, y que el valor superior apenas pasa a la categoría de buen acuerdo, por lo que podemos concluir que esta persona recién entrenada posee un acuerdo aceptable con la persona estándar.

TAMAÑO DE LA MUESTRA

El tamaño de la muestra es esencial para obtener valores de Kappa confiables y precisos. Flack V.(7) brinda las fórmulas necesarias para el cálculo de éste así como una tabla con tamaños de muestra para diferentes valores de precisión y poder. Si le es imposible conseguir el artículo de Flack, recomendamos que cada ejercicio se realice con no menos de 100 personas.

DISCUSION Y CONCLUSIONES

El uso de la prueba Kappa intraclase permitió tener una evaluación en el campo de estandarización del personal que participó del personal en las encuestas de prevalencia de bocio en El Salvador y Panamá (1990), en forma rápida, fácil y con un sesgo minimizado y posible de estimar en las evaluaciones de prevalencia. Para futuros ejercicios de estandarización es recomendable buscar lugares que posean prevalencia de bocio mayor del 10% (si está entre 10% y 20%, se pueden usar los criterios de grado de acuerdo indicados en la Tabla 3). El diagnóstico del tamaño de la tiroides por consenso entre dos evaluadores, cuando uno de ellos no logra alcanzar el grado aceptable de acuerdo durante la etapa de estandarización, disminuye el sesgo en la estimación de la prevalencia. Por otra parte, la prueba no se limita a medir acuerdo entre examinadores de la tiroides. Su uso potencial incluye todas aquellas instancias en las que es necesario asegurar un grado óptimo de concordancia entre un estándar conocido y aceptado con los resultados obtenidos mediante pruebas nuevas o examinadores en entrenamiento. Es necesario resaltar el uso de este estadístico cuando de la calidad de la información depende la toma de decisión para seleccionar individuos, áreas o grupos que recibirán intervenciones no siempre inócuas. Existe aún la necesidad de crear una guía metodólogica para llevar a cabo el proceso de estandarización: este documento representa el primer paso para la elaboración de dicho manual. La guía deberá contemplar el tamaño adecuado de muestra para cada ejercicio de estandarización, así como el número de ejercicios que se deben realizar para contemplar el entrenamiento del personal y para garantizar la calidad de los resultados producidos durante las encuesta Así mismo, dicha guía deberá contemplar la forma de ajustar las estimaciones de prevalencia cuando se conoce el sesgo introducido por los examinadores. En esta ocasión no hemos tocado el uso de Kappa para evaluar concordancia cuando la respuesta no es únicamente presencia de bocio si o no, sino que se desea evaluar la concordancia cuando se realiza clasificación de bocio (no hay bocio, 1a, 1b, 2 y 3). Para esto se remite al lector al artículo de Aickin (3).

REFERENCIAS

1. Bloch D. & Kraemer H. 2x2 Kappa Coefficients: measures of agreement or association. Biometrics, 45:269-287, 1989.
2. I-Kuci Lin, L. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 45:255-268, 1989.
3. Aickin, M. Maximun likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's Kappa. Biometrics, 46:293-302, 1990.
4. Maclennan R, Gaitán & Miller M.C. Observer variation in grading and meásuring the thyroid in epidemiological surveys. In: Endemic Goiter. John Stanbury (Ed). Ginebra, PAHO,1969, 76-77.
5. Fleiss, J. Statistical Methods for Rates and Proportions, 2nd ed. New York: John Wiley & Sons, 1981, 218.
6. Landis J. R, y Koch G. The Measurements of observer Agreement for Categorical Data. Biometrics, 33:159-174, 1977.
7. Flack V, Afifi A. & Lanchenbruch P. Sample size determinations for the two rater Kappa Statistic. Psycometrika, 53:321-325,1988.