16

Bondad de ajuste

En el capítulo anterior se mostró la primera parte del análisis de frecuencias y se mencionó el empleo de la prueba de Chi-cuadrado y la prueba de G, para determinar si los valores observados se ajustaban a los esperados. En este tipo de ensayos se busca establecer si las diferencias observadas entre los datos reales y los teóricos se deben al azar, o si por el contrario la teoría no es buena para explicar la realidad. A eso se lo denomina ensayos de *bondad de ajuste*. Los problemas vistos cuando las frecuencias esperadas se calculaban con la teoría mendeliana, son en esencia casos de bondad de ajuste. En cambio, en los problemas presentados como tablas de contingencia, se buscaba probar la *independencia* de los factores involucrados entre sí. En este capítulo se trataran con más detalle los casos de ajuste. El caso clásico y más difundido es la prueba de Chi-cuadrado. Desde 1900 y a través de su revista Biometrika, Pearson difundió su prueba ampliamente. El ejemplo más famoso fue el informe de Student sobre el error de recuento en hemocitómetro, donde proponía ajustar con una Poisson los valores observados, en lugar de usar la acostumbrada curva de Gauss. En las última décadas la prueba de la G se está imponiendo por su mayor sensibilidad y robustez. Se constituye en la opción más recomendable para el caso paramétrico. Cuando los supuestos no se cumplen, o sencillamente por su facilidad de ejecución, la prueba más recomendable en el campo no paramétrico es la de Kolgomorov-Smirnov, para una y para comparar dos muestras entre sí. Todos estos modelos se presentan a continuación.

16.1 El método clásico de Pearson

Este método aplica el test de la distribución de Pearson: la Chi cuadrado. Consiste en comparar las frecuencias observadas en un experimento, con sus respectivas frecuencias observadas, obtenidas con la distribución teórica que explica el fenómeno estudiado. Hasta la primera década del siglo XX, la distribución Gaussiana era el enfoque aceptado por todos para explicar la Teoría de errores en mediciones. El punto de inflexión histórico a esa tendencia se produce en 1907, con el trabajo de un ingeniero químico W. Gosset, quien en su tarea diaria efectuaba conteos de células de levadura de cerveza en una cámara de recuento de Neubauer denominada hemocitómetro. Gosset sospechó que sus datos se ajustaban mejor a una distribución Poisson que a la de Gauss y para probarlo usó el método de Pearson. Publicó sus trabajos en la revista Biometrika dirigida por el mismo Pearson, su mentor en cuestiones matemáticas. Estas publicaciones llegaron a manos de un estudiante avanzado escocés: R.A. Fisher, quien usó un comentario de

Gosset respecto a la posible existencia de una distribución teórica, mejor que la de Gauss en muestras pequeñas, para desarrollar un nuevo modelo: el de Student. Allí comenzó la moderna estadística, con aplicaciones innovadoras en el trabajo experimental. Parece adecuado comenzar este tema con ese caso. En la Tabla 16.1 se presentan los datos de Student (1907) para mostrar el uso del ajuste a una función de Poisson:

Tabla 16.1 El método de la Chi cuadrado.

N	Oi	Ei	O*i	E*i	$\chi 2=(O*i-E*i)^2/E*i$	$\chi 2_{corr}$
0	75	66,12	75	66,12	1,193	1,062
1	103	119,02	103	119,02	2,156	2,024
2	121	107,11	121	107,11	1,801	1,674
3	54	64,27	54	64,27	1,641	1,485
4	30	28,92	30	28,92	0,040	0,012
5	13	10,41	17	14,56	0,409	0,258
6	2	3,12				
7	1	0,80				
8	0	0,18				
9 y más	1	0,05				
Total	400	400	400	400	7,240	6,515

N : número de células dentro de cada cuadrícula.
Oi frecuencia observada (cantidad de cuadrículas con N células adentro).
Ei : frecuencia esperada obtenida con la fórmula de Poisson.
E*i : frecuencia agrupada.

Si la frecuencia observada es muy chica (Oi < 5), los valores de frecuencia esperada se distorsionan mucho y los valores de χ^2 respectivos contribuyen demasiado al total, pudiendo mostrar significación donde en realidad no la hay. Una regla práctica establecida hace muchos años es agrupar las frecuencias anexas (cuando Oi < 5) en un valor manejable. En la Tabla 16.1 como las cuatro últimas frecuencias observadas son chicas, se agrupan con la más cercana, lo mismo que con las esperadas. Entonces, los datos agrupados se muestran en las columnas cuarta y quinta de la tabla (O*6 = 13+2+1+0+1 = 17 y la esperada es: E*6 = 10,41+3,12+0,8+0,18+0,05= 14,56) Luego de efectuar este arreglo, se calculan los valores de χ^2 respectivos que se colocan en la sexta columna y se obtiene un total de $\chi^2 = 7{,}24$. Si se aplica la corrección de Yates el valor disminuye mucho $\chi^2_{corrr} = 6{,}515$.

Los grados de libertad se calculan de la manera siguiente: Hay 10 clases de frecuencias observadas, pero con el agrupamiento quedan n* = n-4 = 6. En una distribución de Poisson se usa un solo dato poblacional μ, luego resulta r = 1. Entonces, los grados de libertad son:

$$\upsilon = n^* - r - 1 = (10 - 4) - 1 - 1 = 4$$

De tablas es $\chi^2_{0,95\ ;\ 4} = 9{,}488 > \chi^2_{corrr} = 6{,}515$ por lo tanto no se puede rechazar la hipótesis nula. En cambio, si se hubieran estimado las probabilidades con la función de Gauss en vez de la Poisson, no habría sido lo mismo: con Poisson se ajusta mejor que la de Gauss

Como los grados de libertad dependen del número de parámetros poblacionales que se usan con cada modelo teórico, y con el número de clases luego de ser agrupadas, las fórmulas para calcularlos más habituales son:

Tabla 16.2 Grados de libertad

Distribución	Parámetros pob.	r	$\upsilon = n^* - r - 1$	Donde n* se calcula con el número de clases, del problema analizado, luego de hacerse la agrupación para que no ocurra que n < 5
Poisson	μ	1	$n^* - 2$	
Gauss	μ ; σ	2	$n^* - 3$	
Binomial	p	1	$n^* - 2$	

16.2 El método moderno con G-test

Como se vio en el capítulo anterior, el G-test es más sensible y robusto que la prueba de la Chi cuadrado. Por lo tanto, se torna el método más recomendable para analizar el problema de la bondad de ajuste. Su uso es totalmente análogo al visto más arriba, con la salvedad de emplear otro estadígrafo G, para compararlo con los valores críticos de la Chi cuadrado y poder decidir acerca de la hipótesis nula. Aplicando este modelo a los datos de la Tabla 16.1 anterior se tiene:

Tabla 16.3 La prueba de G con Poisson.

N	Oi	Ei	O*i	E*i	Gi
0	75	66,12	75	66,12	18,903
1	103	119,02	103	119,02	-29,780
2	121	107,11	121	107,11	29,508
3	54	64,27	54	64,27	-18,804
4	30	28,92	30	28,92	2,200
5	13	10,41	13	10,41	5,777
6	2	3,12	4	4,15	-0,295
7	1	0,80			
8	0	0,18			
9 y más	1	0,05			
Total	**400**	**400,00**	**400**	**400**	**7,509**

Los grados de libertad en este caso son:

$\upsilon = n^* - r - 1 = (10\text{-}3) - 1 - 1 = 5$

Y la corrección de Williams es:

$$Q = 1 + [(k^2 - 1)] / 6N\upsilon$$

Con k = 7 : número de clases
N = 400 : número de casillas

$Q = 1 + [(49\text{-}1) / (6 \cdot 400 \cdot 5) = 1{,}004$

Si la frecuencia esperada es muy chica (Ei < 3), los valores de G no se pueden calcular. Una regla práctica es agrupar las frecuencias anexas tal que E*i sea mayor o igual a 3. En la Tabla 16.2 como las tres últimas frecuencias esperadas son chicas, se agrupan con la más cercana (la sexta) y lo mismo se hace con las observadas: O*i. Entonces, los datos agrupados se muestran en las columnas cuarta y quinta de la tabla (O*7 = 3,12+0,8+0,18+0,05 = 4,15 y la observada es: O*5 = 2+1+0+1 = 4). Luego de efectuar este arreglo, se calculan los valores de G respectivos, que se colocan en la sexta columna y se obtiene un total de G = 7,509. Si se aplica la corrección de Williams resulta Gcorr = G / Q = 7,50825 / 1,00825 = 7,45. Como los grados de libertad son $\upsilon = 4$ de es $\chi^2_{0,95\,;\,4} = 9{,}488 >$ Gcorr = 7,45 por lo tanto no se puede rechazar la hipótesis nula.

En el tema 7 se desarrolló un ejemplo de aplicación usando la probabilidad binomial, para estudiar el caso de 6.115 familias con 12 hijos. En el cuadro 7.1 se volcaron los datos de las frecuencias observadas y calculadas. Tomando esa información se prepara la Tabla 16.4 para ilustrar el uso de la prueba G con una población supuesta como binomial:

Tabla 16.4 La prueba de G con la Binomial.

Hijas mujeres	Observadas Oi	Esperadas Ei	O*i	E*i	Gi
0	7	2,346			
1	45	26,082	52	28,428	62,802
2	181	132,836	181	132,836	111,996
3	478	410,016	478	410,016	146,664
4	829	854,247	829	854,247	-49,740
5	1112	1265,628	1112	1265,628	-287,804
6	1343	1367,279	1343	1367,279	-48,124
7	1033	1085,211	1033	1085,211	-101,869
8	670	628,055	670	628,055	86,631
9	286	258,475	286	258,475	57,882
10	104	71,803	104	71,803	77,057
11	24	12,089	27	13,022	39,377
12	3	0,933			
Total	6115	6115	6115	6115	94,872

Los grados de libertad son:

$\upsilon = n^* - r - 1 = 9$

La corrección de Williams es:

$Q = 1 + [(k^2 - 1)] / 6N\upsilon$

Con k = 11

N = 6.115 resulta

$Q = 1+[(121-1)/(6 \cdot 6115 \cdot 9)]$

$Q = 1,0003634$

Gcorr = 94,83***

$\chi^2_{0,999\ ;\ 9} = 27,877 <<<$ Gcorr

Se rechaza la hipótesis nula que suponía una distribución binomial en la población. Se encontró evidencia altamente significativa que demuestra que el ajuste no es bueno.

En el tema 9 se presentó un ejemplo sobre las distribuciones en peso de 195 varones. Los datos mostrados en el Cuadro 9.4 se vuelcan a continuación en las dos primeras columnas de la Tabla 16.5, para poder usar el G-test en el caso de una distribución poblacional gaussiana:

Tabla 16.5 La prueba de G con la Normal.

Intervalos del peso	Observadas Oi	Esperadas Ei	O*i	E*i	Gi
hasta 63,5	3	3,3			
63,5-65,5	14	12,2	17	15,5	3,140693
65,5-67,5	30	31,7	30	31,7	-3,307158
67,5-69,5	48	51,1	48	51,1	-6,008015
69,5-71,5	48	50,7	48	50,7	-5,25359
71,5-73,5	39	31,1	39	31,1	17,6556
73,5-75,5	11	11,7	13	14,9	-3,546708
75,5 y más	2	3,2			
Total	195	195	6115	195	2,68082

Los grados de libertad son:

$\upsilon = n^* - r - 1 = 6 - 2 - 1 = 3$

La corrección de Williams es:

$Q = 1 + [(k^2 - 1)] / 6N\upsilon$

Con k = 6

N = 195 resulta

$Q = 1+[(36-1)/(6 \cdot 195 \cdot 3)]$

$Q = 1,00598$

Gcorr = 2.664 $< \chi^2_{0,95;5} = 11,07$

Para este caso, al número total de clases útiles k=6 hay que restarle r = 2 grados de libertad pues se usan los datos muestrales para estimar a los dos parámetros poblacionales (μ ; σ) necesarios para obtener las probabilidades gaussianas. Efectuada la corrección de Williams, el valor Gcorr es menor que el crítico de tablas para el 95% de confianza. Por lo tanto, no se puede rechazar la hipótesis de que los pesos tienen una distribución normal en la población.

16.3 La prueba de Kolgomorov-Smirnov

Se trata de un modelo *no-paramétrico* para comprobar bondad de ajuste. La clasificación de este modelo se debe al hecho que no depende de una distribución original especifica; no se sabe a priori la forma de la distribución de donde se sacan los datos. Por ejemplo el modelo de Student requiere una población original de tipo gaussiana, como la de Chi-cuadrado y la de Fisher. Aquí se trata de *distribuciones libres,* sin parámetros de tipo prefijado como μ o σ. Es cuando se está más interesado en comparar distribuciones antes que parámetros, ni tampoco cuando se trata de estimar parámetros (*modelos paramétricos*). El modelo de Kolgomorov-Smirnov se aplica a distribuciones de tipo continuas y es considerada *conservadora*. También se la usa para probar hipótesis acerca de distribuciones discretas. Es un modelo fácil de realizar. Se basa en calcular las diferencias, en valor absoluto, entre las frecuencias acumuladas relativas observadas y las esperadas, en cada clase. Luego se busca la mayor de las diferencias en valor absoluto, y el estadígrafo Dmáx así obtenido se compara con el valor critico de tablas. En este punto de trata del modelo aplicado al caso de una sola muestra y la Tabla 13 del anexo se presentan los valores críticos para validar hipótesis. Este modelo presenta especiales ventajas cuando se lo aplica a muestras pequeñas, y en tales casos no se necesita agrupar en clases. Otra ventaja es que permite construir limites de confianza para la distribución acumulada completa.

$$Dmáx = \left| OAi - EAi \right| máx$$

Donde OAi: frecuencia acumulada relativa de la clase **i** observada
EAi: frecuencia acumulada relativa de la clase **i** esperada

Para comparar este estadígrafo con un valor crítico de tablas, se busca ese valor en tablas entrando con el tamaño muestral N. Hay dos tablas para los valores críticos de esta distribución. La Tabla 13 que se emplea para los casos más comunes donde se usan los datos muestrales para estimar los valores poblacionales, o sea, para cuando se usan *hipótesis intrínseca*s. A su vez, la Tabla 14 es también para Kolgomorov-Smirnov aplicado a una sola muestra, pero para los raros casos donde se usan *hipótesis extrínsecas*.

La Tabla 13 se extiende correlativamente hasta N = 30, y luego a partir de allí se usa una aproximación dada por las relaciones siguientes:

Cuadro 16.1 Valores críticos para la prueba de Kolgomorov-Smirnov en una sola muestra. Hipótesis intrínsecas

N > 30

α	$K\alpha$	$D\alpha = K\alpha / \sqrt{N}$	
0,10	0,805	$0,805 / \sqrt{N}$	
0,05	0,886	$0,886 / \sqrt{N}$	Significativo
0,01	1,031	$1,031 / \sqrt{N}$	Muy significativo

La Tabla 14 del Anexo es hasta una muestra de tamaño 100. Por lo tanto se debe usar la aproximación asintótica dada por:

$$D\alpha = \sqrt{\frac{-\ln(\alpha/2)}{2N}} = K\alpha / \sqrt{N}$$

Y para los valores críticos usuales se puede preparar otro cuadro resumen:

Cuadro 16.2 Valores críticos para la prueba de Kolgomorov-Smirnov en una sola muestra. Hipótesis extrínsecas

N > 100

α	**Kα**	**Dα = Kα / $\sqrt{N}$**	
0,05	1,358	1,358 / $\sqrt{N}$	Significativo
0,01	1,628	1,628 / $\sqrt{N}$	Muy significativo
0,001	1,949	1,949 / $\sqrt{N}$	Altamente significativo

Este modelo en muchos casos posee mayor potencia que la prueba de la Chi cuadrado para bondad de ajuste. Su uso es más sencillo y es la recomendada para realizar este tipo de estudios en el área de los análisis clínicos y farmacológicos. Solo es aventajada en sensibilidad por el G-test, pero se puede aplicar a casi todos los casos. Para mostrar el uso de esta prueba se desarrollan los ejemplos siguientes, usando los datos de los casos vistos en el punto anterior:

Ejemplo 1) Aplicar el test al caso visto en la Tabla 16.3 de informe de Student en 1907.

n	**Oi**	**Ei**	**Oacum**	**Eacum**	**OAi**	**EAi**	**|OAi - EAi |**	
0	**75**	**66,12**	75	66,12	0,1875	0,1653	**0,0222**	→ **Dmáx**
1	**103**	**119,02**	178	185,14	0,4450	0,4629	0,0179	
2	**121**	**107,11**	299	292,25	0,7475	0,7306	0,0169	
3	**54**	**64,27**	353	356,52	0,8825	0,8913	0,0088	
4	**30**	**28,92**	383	385,44	0,9575	0,9636	0,0061	
5	**13**	**10,41**	396	395,85	0,9900	0,9896	0,0004	
6	**2**	**3,12**	398	398,97	0,9950	0,9974	0,0024	
7	**1**	**0,80**	399	399,77	0,9975	0,9994	0,0019	
8	**0**	**0,18**	399	399,95	0,9975	0,9999	0,0024	
9 y más	**1**	**0,05**	400	400	1,0000	1,0000	0,0000	
Total	**400**	**400,00**						

Para este problema como se usan los datos muestrales para estimar μ con la media y aplicar Poisson se trata de un caso de hipótesis intrínsecas. Luego será:

$$D\alpha = K\alpha / \sqrt{N} = 0,886 / 20 = 0,0443 > Dmáx = 0,0222 \qquad (95\%)$$

Y no se puede rechazar la hipótesis nula. La misma conclusión que la obtenida con el G-test. Aunque el resultado está muy cerca del límite para una confianza del 95%, alcanza para ver la aplicación de este modelo menos poderoso que el G-test.

Ejemplo 2) Aplicar el test al caso visto en la Tabla 16.4, del estudio de 6115 familias de Sajonia con 12 hijos, y ver la distribución de sexos entre ellos; n es la cantidad de hijas mujeres. Se postula una distribución binomial en esta población

n	Oi	Ei	Oacum	Eacum	OAi	EAi	| OAi- EAi |
0	**7**	**2,346**	7	2,346	0,0011	0,0004	0,0007
1	**45**	**26,082**	52	28,428	0,0085	0,0046	0,0039
2	**181**	**132,84**	233	161,26	0,0381	0,0264	0,0117
3	**478**	**410,02**	711	571,28	0,1163	0,0934	**0,0228** → **Dmáx**
4	**829**	**854,25**	1540	1425,5	0,2518	0,2331	0,0187
5	**1112**	**1265,6**	2652	2691,2	0,4337	0,4401	0,0064
6	**1343**	**1367,3**	3995	4058,4	0,6533	0,6637	0,0104
7	**1033**	**1085,2**	5028	5143,6	0,8222	0,8412	0,0189
8	**670**	**628,06**	5698	5771,7	0,9318	0,9439	0,0121
9	**286**	**258,48**	5984	6030,2	0,9786	0,9861	0,0076
10	**104**	**71,803**	6088	6102	0,9956	0,9979	0,0023
11	**24**	**12,089**	6112	6114,1	0,9995	0,9998	0,0003
12	**3**	**0,933**	6115	6115	1	1,0000	0,0000
Total	**6115**	**6115**					

El valor crítico para hipótesis intrínsecas es:

$$D\alpha = K\alpha / \sqrt{N} = 0,886 / 78,2 = 0,0113 < Dmáx = 0,0228 \ (95\%)$$

Y

$$D\alpha = K\alpha / \sqrt{N} = 1,031 / 78,2 = 0,0132 < Dmáx = 0,0228 \ (99\%)$$

Por lo tanto se tiene evidencia muy significativa para rechazar la hipótesis de una distribución binomial en la población.

Ejemplo 2) Aplicar el test al caso visto en la Tabla 16.5, del estudio del peso de 195 varones suponiendo una distribución Normal del peso en la población.

Peso	Oi	Ei	Oacum	Eacum	OAi	EAi	|OAi-EAi |
hasta 63,5	3	3,3	3	3,3	0,0154	0,0169	0,0015
63,5-65,5	14	12,2	17	15,5	0,0872	0,0795	0,0077
65,5-67,5	30	31,7	47	47,2	0,2410	0,2421	0,0010
67,5-69,5	48	51,1	95	98,3	0,4872	0,5041	0,0169
69,5-71,5	48	50,7	143	149,0	0,7333	0,7641	**0,0308** → **Dmáx**
71,5-73,5	39	31,1	182	180,1	0,9333	0,9236	0,0097
73,5-75,5	11	11,7	193	191,8	0,9897	0,9836	0,0062
75,5 y más	2	3,2	195	195,0	1,0000	1,0000	0,0000
Total	195	195					

El valor crítico para hipótesis intrínsecas es:

$$D\alpha = K\alpha / \sqrt{N} = 0,886 / 13,96 = 0,0635 > Dmáx = 0,0308 \ (95\%)$$

Y por lo tanto no se puede rechazar la hipótesis nula, de una distribución gaussiana de los pesos.

16.4 Test de Kolgomorov-Smirnov para 2 muestras

Es un modelo estadístico para verificar si *dos muestras independientes* han sido extraídas de la misma población o de poblacionales con igual función distribución. La prueba de dos colas es sensible a cualquier clase de diferencia entre las distribuciones de donde fueron extraídas ambas muestras. La prueba de una cola se usa para decidir si los valores de la población, de donde se extrajo a una de las muestras, son estocásticamente mayores (o menores) que las de la otra población. Por ejemplo, para probar la predicción de que los valores de un grupo experimental serán mejores que los del grupo control.

Como en el caso de una sola muestra, el modelo compara las dos frecuencias acumulativas relativas de ambas muestras y busca la mayor diferencia encontrada. Este valor multiplicado por ambos tamaños muestrales, se contrasta contra un valor de tablas. Sean las muestras A y B, con un tamaño muestral de N_A y N_B respectivamente, y con Dmáx como la mayor diferencia encontrada en valor absoluto, entonces el valor de K se calcula con:

$$K = N_A . N_B . Dmáx$$

Si este valor K es mayor que el de tablas $K\alpha$ para un nivel de significación dado, se rechaza la hipótesis nula de que ambas muestras provienen de la misma población. Cuando el tamaño de muestras es menor a 25 se pueden usar las Tablas 14 del Anexo. Cuando las muestras son grandes se puede emplear una aproximación, como se muestra al final de las Tablas 14, donde se da el valor crítico de $D\alpha$ para comparar con Dmáx. Finalmente, en la Tabla 15 y 16 del Anexo se presenta una versión simplificada de las tablas, para el caso de igual tamaño muestral, o sea cuando $N_A = N_B = n$; en esta última también se expresan los valores de Dmáx a comparar con $K\alpha$. Para ilustrar estas ideas se presenta el caso siguiente:

Ejemplo 1) De un estudio publicado por la *J. Amer. Med. Ass.* De Friedman, M. Et al., 217, pág, 929-932 (1971) se sacaron los datos de la tabla siguiente:

N	1	2	3	4	5	6	7	8	9	10	11
Grupo A	3,6	2,6	4,7	8,0	3,1	8,8	4,6	5,8	4,0	4,6	
Grupo B	16,2	17,4	8,5	15,6	5,9	9,8	14,9	16,6	15,9	5,5	7,5

A dos grupos de personas, descansados y luego de ingerir una infusión de Hipoclorito de argenina, se le midió los niveles pico de la hormona de crecimiento en plasma. Los sujetos se clasificaron en dos grupos. Grupo A: se trata de individuos relativamente predispuestos a enfermedades coronarias, cuyo temperamento se caracteriza por un excesivo sentido de competitividad, lide-

razgo y urgencia de tiempo. Grupo B: estos son relativamente resistentes a tener problemas coronarios y cuyo temperamento es lo inverso del grupo anterior. Estudios anteriores del mismo equipo (1950) indicaban que el Grupo A tenía mayor tendencia a sufrir enfermedades coronarias que los del Grupo B. En este trabajo, se investiga la incidencia del temperamento en la cantidad de hormona de crecimiento estimulada con la argenina.

La hipótesis de trabajo es que ambos grupos deben diferir. O sea, ambos grupos no provienen de la misma población. Se plantea entonces como

Ho : P (X≤ a) ≥ P (Y ≤ a) No hay diferencia entre ambas muestras.
H_1 : P (X≤ a) < P (Y ≤ a) El nivel hormonal es menor en el Grupo A que en el B.

Para desarrollar este modelo se deben seguir los pasos siguientes:

Paso 1) Se ordena cada grupo en orden ascendente. Con el rango se obtiene el número de clases, como en los histogramas, pero tratando de tener el mayor número de clases posible, para ambas muestras. Los resultados se muestran en la primer columna del cuadro siguiente.

Paso 2) Se calculan las frecuencias de cada grupo en las mismas clases, luego sus frecuencias acumuladas y finalmente, dividiendo por el respectivo tamaño muestral, se calculan las frecuencias acumuladas relativas, que se muestran en las columnas 2 y 3.

Paso 3) Para cada clase, se calculan las diferencias en valor absoluto de las frecuencias acumuladas relativas de cada grupo. Y se busca el valor máximo de esas diferencias Dmáx = 0,7

Paso 4) Se determina cual tabla de valores críticos usar. En este caso, se trata de muestras pequeñas con diferente tamaño muestral (Tabla 15) Con $N_A = 10$ y $N_B = 11$ se obtiene de tablas los valores críticos: $K_{0,05} = 60$; $K_{0,01} = 77$ y $K_{0,001} = 89$

Paso 5) Se comparan estos valores críticos con el estadígrafo K de Kolgomorov-Smirnov para dos muestras: $N_A \cdot N_B \cdot Dmáx = K = 10 \cdot 11 \cdot 0,7 = 77$

	Grupo A	Grupo B	Diferencias
2,5 - 3,4	0,2	0	0,200
3,5 - 4,4	0,4	0	0,400
4,5 - 5,4	0,7	0	**0,700 (Dmáx)**
5,5 - 6,4	0,8	0,182	0,618
6,5 - 7,4	0,8	0,182	0,618
7,5 - 8,4	0,9	0,273	0,627
8,5 - 9,4	1	0,364	0,636
9,5 - 10,4	1	0,455	0,545
10,5 - 11,4	1	0,455	0,545
11,5 - 12,4	1	0,455	0,545
12,5 - 13,4	1	0,455	0,545
13,5 - 14,4	1	0,455	0,545
14,5 - 15,4	1	0,545	0,455
15,5 - 16,4	1	0,818	0,182
16,5 - 17,4	1	1	0,000

Se concluye que se debe rechazar la hipótesis nula y aceptar la alternativa. O sea, se tiene evidencia muy significativa K = 77** de que ambas muestras no provienen de la misma población, los valores hormonales del Grupo A son menores que los del Grupo B.

16.5 Test de Bondad de ajuste con repetición

Es un modelo estadístico para los casos en que los datos para un test se hacen de manera repetida. La manera más simple es agrupar los datos repetidos en una especie de "pool" de datos y usar los totales para hacer la prueba. Si bien se respeta la regla de oro de maximizar los datos, el problema es que se puede perder información a causa del agrupamiento y así no tener un panorama mejor. Para mostrar un caso de este tipo se presenta el ejemplo siguiente:

Ejemplo 1) Una industria farmacéutica encara un estudio sobre el efecto de un nuevo analgésico. Para ello, efectúa relevamiento en 8 ciudades diferentes de un país elegidas al azar, donde se efectúa la misma prueba. Todos los casos analizados con respuesta positiva, se dividen según el sexo. La hipótesis de trabajo es que este factor no tiene influencia en los resultados obtenidos. Los datos obtenidos se muestran en la tabla siguiente. Con los cuales se realiza un G-test como se vio en el punto 15.2 del capítulo anterior:

Sexo	Observadas	Esperadas	G	Como $\chi^2_{0,999\ ;\ 1}$ = 10,828 los resultados fueron altamente significativos y se tiene una muy fuerte evidencia para rechazar Ho
Femenino	616	530	185,26	
Masculino	444	530	-157,22	
Total	1060	1060	28,04***	

El farmacéutico a cargo del estudio estadístico deber rechazar la hipótesis del 50% para cada sexo, parecería que este factor tiene algo que ver. Sin embargo, antes de tomar una decisión al respecto, decide aprovechar al máximo los datos recogidos, con el desglose en las 8 localidades y enfoca el estudio desde otro ángulo:

Caso 1) Test de independencia (tabla de contingencia)

Frecuencias Observadas

Loc.	f	m	Total
1	83	47	130
2	77	43	120
3	110	96	206
4	92	58	150
5	51	31	82
6	48	61	109
7	70	42	112
8	85	66	151
Total	616	444	1060

Frecuencias Esperadas

Loc.	f	m	Total
1	75,547	54,453	130
2	69,736	50,264	120
3	119,71	86,287	206
4	87,17	62,83	150
5	47,653	34,347	82
6	63,343	45,657	109
7	65,087	46,913	112
8	87,751	63,249	151
Total	616	444	1060

Coloca los datos observado en una tabla como la de arriba y para el cálculo de las frecuencias observadas, usa el concepto de independencia. Es decir, multiplicando los totales marginales de cada casilla y dividendo por N = 1060. En la primer casilla será E_{11} = (130 . 616) / 1060 = 75,547 y así sucesivamente completa la tabla de frecuencias esperadas de más arriba. Luego hace el test:

Loc.	Gf	Gm	Gf+Gm
1	15,61784	-13,8356	1,782242
2	15,26	-13,4239	1,836087
3	-18,6161	20,48096	1,864869
4	9,923213	-9,27915	0,644065
5	6,924128	-6,35701	0,567118
6	-26,6275	35,34652	8,719038
7	10,1883	-9,29289	0,895414
8	-5,41473	5,619838	0,205107
Tot	7,25518	9,25876	16,51394
G =	16,51394		

Para cada casilla se calcula el valor:

$Gij = 2\ Oij\ .\ \ln (Oij / Eij)$

La suma total para filas y columnas es:

G = 16,514 *

Los grados de libertad son: $\nu = (8-1)(2-1) = 7$

De tablas $\chi^2_{0,95\ ;\ 7} = 14,067$

Se tiene prueba (95%) de *heterogeneidad* en los datos. Esto significa, que el factor Sexo, no es independiente del factor Localidad. En la tercer columna se coloca el valor Gf + Gm para poder realizar otro tipo de análisis. Comparando el peso que tiene cada localidad en el total, se nota que más del 50% del valor total de G se debe a la Localidad 6, mientras que las demás no parecen ser significativas, porque tienen una contribución parecida en el total. El investigador ahora sospecha que los resultados inesperados del primer análisis que efectuó, se originan en la gran desproporción encontrada en esta localidad. Cosa que deberá ser investigada con mucho cuidado.

En realidad, con cada localidad se puede armar una tabla 2x2, usando los datos observados en ella, y para obtener los esperados se usa una estimación del porcentaje poblacional. Esto es, la probabilidad de que sea de sexo femenino es P(f) = 616 / 1060 = 0,581132 y para el masculino es: P(m) = 444 / 1060 = 0,418868. Entonces, en una localidad cualquiera, por ejemplo la 6, se tendrá una frecuencia esperada femenina Ef =109 . 0,581132 = 63,34 y masculina Em = 109 . 0,418868 = 45,66 lo que se obtiene multiplicando la probabilidad por el tamaño muestral de la localidad. Se puede entonces armar una tabla para esa localidad como la siguiente :

Sexo	Observadas	Esperadas	G	Como $\chi^2_{0,95;\ 1} = 3,841$ los resultados fueron a
Femenino	48	63,34	-26,628	significativos y se tiene evidencia para re-
Masculino	61	45,66	35,347	chazar Ho
Total	109	109	8,719	

Realizado el test de G, resultó significativo pues $G = 8,719\ * > \chi^2_{0,95;\ 1} = 3,841$. Análogamente se pueden plantear los G-test por localidad y comparar siempre contra el mismo valor crítico. Eso se muestra en la última columna de la tabla al principio de esta página. Se concluye que la única localidad con diferencia significativa es la Localidad 6. Como esto no es la tendencia general, algo debe haber ocurrido en tal localidad para tener esos resultados. Algo que merece ser investigado con cuidado en un próximo experimento, que deberá ser planeado cuidadosamente.

Se puede hacer un resumen de estos resultados calculando G agrupado = 1,5004 que se calcula usando Oi = 616 ; Oj = 444 y Ei = 1060 (9/16) ; Ej = 1060 (7/16)

Tests realizados	Grados de libertad	Estadígrafo G
Total Datos Agrupados	1	1,500
Total por localidad	7	16,514*
TOTAL	8	18,014*

16.6 Análisis de concordancia

En el análisis de frecuencias el problema general es ver si los datos experimentales se ajustan a una teoría dada, la cual trata de interpretar la realidad en términos matemáticos. Esto se vio en varios capítulos anteriores y en el presente, con varios de los modelos presentados. Cuando el concepto de independencia matemática, visto en el capítulo 6, es la base de cálculo para obtener las frecuencias esperadas teóricamente se habla de tablas de contingencia. En cambio, cuando se usa otra teoría como la Mendeliana, la de Gauss, etc. para el cálculo de las frecuencias esperadas se habla de bondad de ajuste o de tablas de contingencia para el ajuste. En el fondo ambos tipos de tablas son dos formas de abordar el mismo problema: los datos muestrales versus la realidad. Lo que ocurre es que a veces la realidad se la toma de una teoría cualquiera y otras de la teoría de la independencia estadística. El problema básico es que no siempre la *asociación estadística significa asociación clínica*. Para ilustrar este concepto se desarrollará a continuación el caso del Análisis de Concordancia (*Agreement*) donde se mostrará que el significado estadístico de la asociación, a veces no tiene el sentido clínico del problema. Un problema habitual en Medicina es ver cuando un método "nuevo", puede ser usado para reemplazar al habitual que se usa en el laboratorio (el "viejo"). El método más común de hacer esto es medir a cada individuo con los dos métodos a la vez, como se explicara en el Modelo de Student para muestras apareadas y en los modelos no paramétricos equivalentes.

Cuando a cada individuo se lo mide con más de un método a la vez, a las muestras obtenidas se las considera apareadas. En el caso general habrá k métodos, aplicado a n individuos y así el total de datos será N = n . k. Se trata de casos especiales de Tablas de contingencia donde se puede aplicar el modelo de Cochran para analizar estos casos. Para el caso particular de tener k = 2 métodos clínicos, hay varios modelos que se pueden usar, se trata del caso de individuos medidos con ambas técnicas a la vez, y entonces resulta un caso particular de las tablas de contingencia de 2 x 2. En estos casos la idea es armar la tabla de una manera tal que evite el apareo de los datos. El requisito básico para aplicar los modelos de tablas de contingencia es exigir la independencia de los datos, por lo tanto se debe evitar que un mismo paciente aparezca en dos lugares de la misma tabla, con un diseño como se mostró en la Tabla 14.3 anterior:

Cuadro 16.3 Tabla de Concordancia

	Test 1		
Test 2	(+)	(-)	Total
(+)	r_{11}	r_{12}	$r_{11}+ r_{12}$
(-)	r_{21}	r_{22}	$r_{21}+ r_{22}$
Total	$r_{11}+ r_{21}$	$r_{12} + r_{22}$	N

El modelo más usado para este análisis es el Test de McNemar, corregido con Yates. Pero además se pueden usar el G-test de McNemar corregido con Williams, el test no paramétrico de Cochran y otros dos menos conocidos como el del Log-Odds Ratio y el test Cohen-Kappa. Como se mostrará a continuación ninguno de estos cinco modelos puede ser recomendado para el análisis clínico del problema, por lo que la Cátedra ha desarrollado otro enfoque usando los conceptos clínicos primero y la Estadística después.

16.6.1 Modelo de McNemar

El test de McNemar para analizar la significación entre dos métodos clínicos, se basa en el supuesto siguiente: Asumiendo que hay una población en la cual se usan dos métodos en cada individuo de la misma, para medir una magnitud dicotómica cuyo resultado se expresa como positivo (+) o negativo (-) con cada una de ellas, entonces habrá cuatro conjuntos mutuamente excluyentes y colectivamente exhaustivos, si se ubica a cada uno de los individuos en una de las cuatro celdas de la tabla de concordancia, y así se tiene una partición de la población. Luego si se eligen al azar un total de N individuos los datos obtenidos se pueden presentar como en la Tabla 16.3 anterior. El interés principal se centra en los valores discordantes, los cuales seguirán una distribución Binomial con un parámetro igual a 0,5. El test de *dos proporciones correlacionadas* usa el cuadrado del estadígrafo normal tipificado y puede ser analizado con la distribución de Chi-cuadrado. Aplicando la corrección de Yates por continuidad resulta:

$$z^2 = (| r_{12} - r_{21} | - 1)^2 / (r_{12} + r_{21})$$

La Ho es que no hay asociación entre ambos métodos y se la testea con el valor de tablas para 1 grado de libertad y un nivel de significación α: $\chi^2_{(\alpha;1)}$ Entonces cuando $z^2 > 3.841$ se puede rechazar la Ho con un 95% de confianza. Esto es considerado por algunos clínicos como una prueba de la existencia de concordancia entre ambos métodos, cosa que es bastante discutible como se verá más adelante.

Ejemplo 1) Un Bioquímico decide testear dos técnicas de laboratorio para diagnosticar cierta enfermedad, para elegir entre ambas. Los resultados se clasifican como positivo cuando se la detecta y negativo cuando el paciente aparenta estar sano. Se quiere descubrir si hay asociación entre el carácter positivo-negativo y el procedimiento de diagnóstico. Para ello escoge 100 pacientes al azar y les aplica ambas técnicas. Los resultados obtenidos se pueden presentar como:

	Método de medición		
Resultado	Técnica A	Técnica B	Total
(+)	70	52	122
(-)	30	48	78
Total	100	100	200

Lo primero a notar es que hay 100 pacientes en el experimento, pero el total de la tabla es 200. Lo que indica que habrá pérdida de independencia, pues los pacientes se ubicaron en más de una celda de la tabla y no se pueden aplicar los modelos vistos. Por eso McNemar propuso cambiar la unidad muestral.

	Técnica A		
Técnica B	(+)	(-)	Total
(+)	38	14	52
(-)	32	16	48
Total	70	30	100

Ahora los datos son los mismos pero ordenados de otra forma: ++ ; +- ; -+ y - -. De esa manera, cada paciente aparece una sola vez en la tabla y se mantiene la independencia. McNemar propuso una fórmula con distribución Chi-cuadrado.

Se toman en consideración el par de valores donde los métodos difieren ($r_{12} = 32$ y $r_{21} = 14$) pero no se toman en cuenta los casos donde coinciden. El estadígrafo de McNemar sin la corrección de Yates es:

$$\chi^2 = (r_{12} - r_{21})^2 / (r_{12} + r_{21})$$

O sea, $\chi^2 = (14 - 32)^2 / (14 + 32) = 7.04^{**}$

Y con la corrección de Yates es $\chi^2_{corr} = (| r_{12} - r_{21} | - 1)^2 / (r_{12} + r_{21}) = 6{,}28^{*} > \chi^2_{0{,}95\ ;\ 1} = 3{,}841$

Donde se puede apreciar que la corrección vuelve más conservador a este test, es decir le va a costar más rechazar la Ho. La conclusión estadística es que: hay evidencia significativa de asociación entre ambas técnicas, estas no son independientes. Desde el punto de vista clínico, a esto se lo interpreta como que existe *concordancia* entre ambas técnicas y entonces, la Técnica 1 puede reemplazar a la Técnica 2 y viceversa. Esta es la manera habitual de hacer el análisis de concordancia. Sin embargo, hay dos objeciones básicas que se le pueden hacer:

Objeción 1: No se toma en cuenta, ni el tamaño muestral total N, ni el número de concordancias halladas.

Objeción 2: Cuando la cantidad de discordancias de un tipo sea muy similar a la del otro tipo, el resultado siempre será no significativo, sin importar ni el tamaño de las mismas.

Por ejemplo, suponiendo ahora que los resultados obtenidos en 400 muestras sean:

	Método viejo		
Método nuevo	(+)	(-)	Total
(+)	180	22	202
(-)	10	188	198
Total	190	210	400

$\chi^2 = (22 - 10)^2 / (10 + 22) = 4{,}5^{*} > \chi^2_{0{,}95\ ;\ 1} = 3{,}841$
Sin la corrección de Yates hay concordancia

$\chi^2 = (| 22 - 10| - 1)^2 / (10 + 22) = 3{,}78 < 3{,}841$
Con la corrección de Yates no hay concordancia

Luego, la conclusión es que no se puede reemplazar al método viejo con el nuevo si se toma en cuenta la corrección de Yates. Esto, desde el punto de vista estadístico. Sin embargo se puede ver que si se mantiene el número de discordancias en 32 y se aumenta el número de muestras a 400 millones (en vez de 400) las conclusiones serían las mismas. Pero desde el punto de vista clínico, no es lo mismo tener 32 discordancias en 400 muestras que en 400 millones. El sentido común indica que obtener 32 discordancias en 400 millones de casos, es prueba más que suficiente para poder reemplazar a un método con el otro. Y esta es la base de la primera objeción.

Por otra parte, con respecto a la segunda objeción, suponiendo que en el ejemplo de más arriba se hubiesen obtenido $r_{12} = r_{21} = 1$, el valor del estadígrafo sería prácticamente nulo y no se podría rechazar Ho. El problema es que lo mismo ocurre si hubiese sido $r_{12} = r_{21} = 199$. Pero, no es lo mismo tener 2 discordancias que 398 discordancias en 400 casos desde el punto de vista clínico. El sentido común nos dice que en el primer caso la concordancia es casi perfecta, mientras que en el segundo casi no existe.

Ambas objeciones muestran que el Test de McNemar puede ser útil desde el punto de vista estadístico, pero *no puede ser recomendado* para ser usado en Clínica. Lo que ocurre es que el concepto de independencia estadística es en realidad lo que estudia McNemar. Y no tiene nada que ver con el concepto clínico de concordancia, porque que exista asociación estadística, no implica necesariamente que haya asociación clínica.

16.6.2 Modelo de Cochran

Este test puede ser aplicado para el caso de que haya k = 2 muestras. Notar que si se calcula el estadígrafo de Cochran $Q = [2.CS - T^2] / [2\ T - RS]$ para el último ejemplo es:

$T = (r_{12} + r_{21}) + 2\ r_{11} = (22 + 10) + 2 . 180 = 392$

$RS = (r_{12} + r_{21}) + 4\ r_{11} = (22 + 10) + 4 . 180 = 752$

$CS = (r_{11} + r_{21})^2 + (r_{11} + r_{12})^2 = (180 + 10)^2 + (180 + 22)^2 = 76.904$

$Q = [2 . 76.904 - 153.664] / [2 . 392 – 752] = 144 / 32 = 4,5^* > \chi^2_{0,95\ ;\ 1} = 3,841$

Notar que el valor del estadígrafo Q es exactamente igual al valor de χ^2 de McNemar sin la corrección de Yates. Este resultado muestra que hay evidencia muy significativa para rechazar la hipótesis nula. La población *no tiene* proporciones iguales de positivos respecto a los dos procedimientos de diagnóstico. El bioquímico concluye que los métodos analizados, el nuevo y el viejo son concordantes. Es decir, que decide usar ambos métodos indistintamente.

Sin embargo las dos objeciones planteadas en el punto anterior, se pueden plantear nuevamente para este caso porque el valor de Q es igual al de McNemar, como puede demostrarse fácilmente haciendo pasaje de términos. Es decir Q es igual al McNemar sin corrección de Yates:

$Q = [2.CS - T^2] / [2\ T - RS] = \chi^2 = (r_{12} - r_{21})^2 / (r_{12} + r_{21})$

Nuevamente entonces se puede concluir que este test no puede ser recomendado para Clínica.

16.6.3 G-test de McNemar

El G-test para el caso de McNemar consiste en calcular el Likelihood para el estadígrafo basándose en la distribución Multinomial, este método también se llama *individuos doblemente testeados*. Hay dos probabilidades básicas: la observada y la esperada, entonces el logaritmo natural del cociente entre ambas es G/2, donde el estadígrafo G se distribuye aproximadamente como una Chi-cuadrado con un grado de libertad. Se calcula con:

$$G = 2\ \{r_{12} \ln[2r_{12} / (r_{12} + r_{21})] + r_{21} \ln[2r_{21} / (r_{12} + r_{21})]\}$$

Y la corrección de Williams por continuidad es q = 1 + (1 / 2 N), o sea G' = G / q. Para ilustrar este procedimiento se puede usar el ejemplo anterior:

	Método A		
Método B	(+)	(-)	Total
(+)	180	22	202
(-)	10	188	198
Total	190	210	400

$G = 2\ \{22.\ln[2.22 / (22 + 10)] + 2.10 \ln[2.10 / (10 + 22)]\}$

$G = 4,551$ y $q = 1 + [1 / (180+188)] = 1,00125$ luego

$G' = 4,54^* > \chi^2_{0,95\ ;\ 1} = 3,841$

En síntesis, analizando los datos con los tres tests presentados hasta ahora, las conclusiones estadísticas serían:

- Si se aplica el test de McNemar corregido con Yates: No hay asociación entre ambos métodos.
- Si se aplica el test de Cochran o el McNemar sin corregir: Hay asociación significativa.
- Si se aplica el G-test: Hay asociación significativa.

Se puede ver entonces que el G-test tiene mayor poder de discriminación que los restantes y es el más recomendable desde el punto de vista estadístico. El de Cochran es bastante similar y la diferencia entre ambos es pequeña y como es un test no paramétrico su campo de aplicación es más grande. El método clásico de McNemar pierde su poder al ser corregido con Yates. La conclusión estadística sería la siguiente: Se tiene evidencia significativa de la concordancia entre ambos métodos. Pero también al G-test se le pueden efectuar las dos objeciones básicas, pues el tamaño muestral N solo aparece en la corrección de Williams, cuyo peso relativo es muy pequeño en el resultado final. Notar que si N tiende a infinito, q tiene a uno y la corrección no existe.

Por lo tanto *la utilidad de estos tres tests desde el punto de vista clínico no es aceptable*. A ninguno de estos tres estadígrafos se lo puede considerar como un índice de asociación clínico, sino solamente de asociación estadística.

16.6.4 El test del log-Odds Ratio

Este test se aplica en realidad para las Tablas de la verdad vistas en el Gráfico 4.1 cuando se explicaron los índices clínicos. Pero si uno de los dos métodos puede ser considerado como el de referencia (por ejemplo la inmuno fluorescencia para la toxoplasmosis, etc.), entonces la Tabla de la verdad se convierte en una de contingencia para muestras apareadas. Así la Tabla 16.3 puede ser considerada como una Tabla de la Verdad y se puede aplicar este modelo estadístico. El logaritmo natural del OR llamado L, es simétrico alrededor del cero, cosa que no ocurre con OR. El cuadrado de L dividido por su desvío estándar SD (L) se distribuye como una Chi-cuadrado con un grado de libertad. La Ho : El valor esperado: Ξ (L) es nulo, es lo mismo que decir: no hay significación estadística entre los métodos. Y el estadígrafo se calcula con:

$L = \ln \{[(r_{11} + 0.5) (r_{22} + 0.5)] / [(r_{12}+ 0.5) (r_{21} +0.5)] \}$ y

$SE^2 (L) = [1/ (r_{11} + 0.5)]+[1/ (r_{22} + 0.5)] + [1/ (r_{12}+ 0.5)] + [1/(r_{21} + 0.5)]$ luego será

$\chi^2 = [L - \Xi (L)]^2 / SE^2 (L) = L^2 / SE^2 (L)$ para ser contrastado con $\chi^2_{(0.05; 1)} = 3.841$

Aplicando este método al ejemplo del punto anterior resulta $\chi^2 = 164{,}1 >>> \chi^2_{(0.05; 1)}$ y hay una fuerte evidencia como para rechazar la Ho. En este modelo ninguna de las dos objeciones anteriores puede ser aplicada, pero desgraciadamente hay una nueva:

Objeción 3: Si los valores de ambas diagonales en la tabla son intercambiados, χ^2 no varía.

Esto significa que si ahora fuese: $r_{11} = 10$, $r_{22} = 22$, $r_{12} = 188$ y $r_{21} = 180$, el resultado sería de nuevo $\chi^2 = 164{,}1$. Estadísticamente significa lo mismo, pero clínicamente no es lo mismo encontrar 32 que 368 discordancia en 400 muestras. *Tampoco se puede recomendar este test*.

16.6.5 El test de Cohen-Kappa

En este test visto en el capítulo 14 no se puede hacer ninguna de las tres objeciones presentadas más arriba. Sería el único que resiste al sentido común. Sin embargo existen muchas objeciones respecto a su uso como:

- Kappa no es realmente una medida de concordancia corregida por el azar.
- Kappa es un índice de concordancia del tipo "ómnibus", no distingue entre varios tipos de fuentes de desacuerdo. Para la tala de 2 x 2, no distingue entre un caso (+ -) de otro que sea (- +).
- Kappa resulta influenciado por la prevalencia y las categorías analizadas. Como resultado este índice es raramente comparable con procedimientos, poblaciones y estudios de corte. (Thompson y Walter, 1988, Feinstein y Cicchetti, 1990)
- Kappa puede resultar bajo aunque haya un alto nivel de concordancia, incluso cuando los valores individuales sean exactos. (Uebersax, 1988)
- Kappa requiere que dos métodos usen el mismo tipo de categorías, y a veces el interés está basado en analizar la consistencia entre diferentes categorías (por ejemplo, un método usa las categorías A, B y C, mientras que el otro usa A, B, C, D y E)
- Hay tablas que pretenden categorizar los rangos de Kappa como: "buenos", "justos", "pobres", "malos" etc., las que no deben ser usadas por su arbitrariedad

El estadístico kappa para el análisis de concordancia se calcula con:

$$\text{Kappa} = \frac{\text{concordancia observada} - \text{concordancia esperada}}{1 - \text{concordancia esperada}}$$

Donde la concordancia observada es el nivel de concordancia = $(r_{11} + r_{22}) / N$

Y la concordancia esperada con: $[(r_{11} + r_{12})(r_{11} + r_{21}) / N] + [(r_{22} + r_{12})(r_{22} + r_{21}) / N]$

Aplicando este método al ejemplo anterior resulta kappa = 0,84, resultado que puede ser calificado como una concordancia casi perfecta, de acuerdo a la tabla de calificaciones especificada por Guyatt, G. en el *JAMA* (2002):

Calificación de la concordancia de acuerdo al valor de kappa

Valor de kappa	**Calificación de la concordancia**
0	Muy pobre
Entre 0 y 0,02	Pobre
Entre 0,2 y 0,4	Ligera
Entre 0,4 y 0,6	Moderada
Entre 0,6 y 0,8	Substancial
Entre 0,8 y 1,0	Casi perfecta

Feinstein y Cicchetti en 1990, presentaron dos paradojas básicas de kappa: 1) Alta concordancia pero bajo kappa. 2) Totales marginales desbalanceados producen valores más altos de kappa, que si estuvieran balanceados (como era de esperar). Por ello, la recomendación oficial a los clínicos es que utilicen el procedimiento de phi en lugar de kappa.

16.6.6 El procedimiento phi

Este método se conoce también como: concordancia chance-independiente, porque es independiente del nivel de concordancia, solo depende del Odds Ratio. Lo cual supone una ventaja al no depender tanto de la prevalencia en los totales marginales como kappa. Su valor se calcula con la relación siguiente:

$$\text{Phi} = \frac{(ad) - (bc)}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

Si se aplica al ejemplo anterior resulta phi = 0,84, el mismo valor que el obtenido por kappa. Sin embargo a pesar de las recomendaciones para los médicos, phi presenta las mismas desventajas de kappa y si se aplica a las paradojas de Feinstein tampoco las resuelve. Hay dos problemas:

Problema 1: Puede ocurrir que kappa y phi sean nulos, (o sea una gran discordancia), no importa en nivel de concordancia hallado en el estudio. Cuando (a b) = (c b) entonces siempre resultará kappa = phi = 0. Por ejemplo, sean los dos casos siguientes:

Caso 1

	Método A		
Método B	(+)	(-)	Total
(+)	3600	60	3660
(-)	60	1	61
Total	3660	61	3721

Caso 2

	Método A		
Método B	(+)	(-)	Total
(+)	20	400	420
(-)	1	20	21
Total	21	420	441

En el Caso 1 se puede ver a simple vista la gran concordancia observada en el estudio, pues en 3720 casos se encontraron 120 discordancias. Esto implica un nivel de concordancia del 97%, y sin embargo el valor de kappa es nulo, lo mismo que el valor de phi.

En el Caso 2 se puede ver de un vistazo que no hay concordancia, pues se encontraron solo 40 concordancias en 441 casos, lo que implica un nivel de concordancia del 9%. Acá también los valores de kappa y phi son nulos, lo que estaría de acuerdo con los valores observados del nivel de concordancia. Pero el punto es que cuando se cumple la condición vista más arriba, entonces no importa la concordancia observada, kappa y phi se anularan y serán ciegos ante estos hechos clínicos. Esto muestra, que ninguno de ambos procedimientos se basa en concepto clínico de la concordancia sino en el estadístico.

Problema 2: Kappa y phi pueden ser negativos, no importa el tipo de concordancia hallada en el estudio. Cuando a = 0, o bien d = 0, los valores de kappa y phi serán negativos, no importa como son los demás valores hallados.

En las dos tablas siguientes se muestran dos situaciones bien opuestas, y sin embargo los valores de kappa y phi se parecen mucho. En el Caso 3 la concordancia es muy clara, y sin embargo el valor kappa = phi = - 0,01. Mientras que en el Caso 4 la discordancia es muy clara y los valores calculados son: kappa = -0,04 y phi = - 0,13, indicando la pobre concordancia hallada.

Caso 3

	Método A		
Método B	(+)	(-)	Total
(+)	390	5	395
(-)	5	0	5
Total	395	5	400

Caso 4

	Método A		
Método B	(+)	(-)	Total
(+)	0	260	260
(-)	10	250	260
Total	10	510	520

16.6.7 El método de visión dual

La concordancia sirve para ver si un método clínico puede ser reemplazado por otro. La idea no es ver cual de los dos es el mejor, sino simplemente si se pueden intercambiar. Para analizar el problema de la concordancia se propone una nueva idea: "Visión dual de la concordancia". Se basa en un supuesto clínico principal:

Dos métodos clínicos concuerdan cuando tienen la misma sensibilidad y especificidad.

Esto se verifica, si y solo si $r_{11} = r_{22}$ En efecto, si se supone que uno de los dos métodos es el ideal, entonces la Tabla de concordancia se transforma en una Tabla diagnóstica y se pueden calcular la sensibilidad y especificidad del otro método. Ahora, haciendo la suposición inversa se pueden calcular el otro par de sensibilidad y especificidad. La única forma de que se hagan iguales es cuando $r_{11} = r_{22}$. Los tests estadísticos para verificar ese supuesto son el McNemar G-test con la corrección de Williams, el Q-test de Cochran, o el clásico Chi-cuadrado con la corrección de Yates. Pero esto solo tendría las objeciones vistas más arriba, por lo tanto se necesita estudiar además el punto de vista clínico usando el nivel de concordancia, o los Odds de discordancia.

El procedimiento de visión dual consiste de dos etapas:

1) *Visión estadística*: Lo primero es verificar si hay concordancia estadísticamente hablando. Para eso se puede usar cualquiera de los tests estadísticos mencionados. Sin embargo, la sugerencia es usar el G-test porque es el más potente de los tres como se vio antes. O bien, el Q-test cuando alguna de las frecuencias observadas de discordancia sea nula. La idea es obtener el valor de G corregido y compararlo con el valor crítico de la Chi-cuadrado para 95% de confianza. Cuando sea $G > 3{,}841$ entonces hay prueba estadística que muestra que los métodos no tienen la misma sensibilidad y especificidad. Si $G < 3{,}841$ no hay evidencia como para pensar que no se puedan intercambiar los métodos.

Pero esto como estos no es suficiente porque estos tests no tienen el cuenta ni el tamaño muestral, ni el nivel de concordancia, pueden aparecer paradojas clínicas como las vistas, y por lo eso se necesita de una segunda etapa.

2) *Visión clínica*: Lo segundo es ver si la concordancia encontrada es suficiente desde el punto de vista clínico. Para ello se necesita definir un criterio clínico de aceptación. Por ejemplo:

La concordancia es aceptable cuando la observada sea mayor o igual que un nivel de concordancia crítico ($\lambda_{crítico}$) definido por los médicos para cada enfermedad

La idea es encontrar el intervalo de confianza del 95% para el nivel de concordancia observado en el estudio, y ver si el valor crítico cumple con la condición de ser mayor o igual que el límite inferior del intervalo. Por ejemplo, si el criterio médico es: *La concordancia es aceptable cuando no hay más de 10 discordancias en 100 casos.* Esto significa que el nivel crítico de concordancia se establece en un 90%. Pero también se puede usar el otro índice: Odds de Discordancias, donde DO = 10 /90, o sea DO es 1 : 9 Clínicamente, la concordancia será aceptable cuando no haya más de 1 discordancia, y 9 concordancias en 10 casos.

Casos posibles:

a) Si hay concordancia estadística en el primer paso se decide directamente con el segundo paso. Esto significa que todo el peso de la responsabilidad por el cambio recae en el médico, y no en estadística como es ahora.

b) Si no hay concordancia estadística en el primer paso significa que algo pasará con la sensibilidad y especificidad si se hace el intercambio de métodos. Y eso hay que estudiarlo con más cuidado (análisis más incisivo).

b-1) Ahora se hace el segundo paso: Si no se acepta por el criterio clínico adoptado, entonces significa que: estadísticamente y clínicamente hay prueba como para rechazar el intercambio.

b-2) Si en cambio, hay aceptación clínica uno se siente inclinado a aceptar el cambio, porque este criterio es el decisivo. El problema es que no hay en la literatura médica criterios de concordancia que sirvan de guía para todos los casos y para todas las enfermedades. Entonces el médico antes de aceptar tiene que hacer un estudio más cuidadoso:

Análisis más incisivo: Suponiendo que el Método 1 es el de referencia (o la verdad) entonces se pueden obtener la sensibilidad, especificidad e Índice de Youden del Método 2 (digamos S_2, E_2 y Y_2). Viceversa, si suponemos al Método 2 como el de referencia se pude obtener lo mismo para el Método 1 (digamos S_1, E_1 y Y_1). Potencialmente hablando, hacer el intercambio de métodos puede significar un cambio en estos índices, que se puede estimar con: $\Delta S = S_1 - S_2$ y análogamente: ΔE y ΔY.

Entonces la idea es ver si un cambio de digamos $\Delta S = 20\%$ es aceptable para el tipo de enfermedad que se está estudiando: Tipo I (el error más grave es un falso negativo, y hay que estudiar la variabilidad potencial de la sensibilidad). Tipo II (lo más grave es un falso positivo, y hay que estudiar la variabilidad potencial de la especificidad) y el Tipo III (los dos errores son igualmente graves, y hay que estudiar la variabilidad potencial del índice de Youden).

Solo el criterio clínico puede decidir en estos casos. Por ejemplo, hay enfermedades donde no sería aceptable un DO de 1 : 9 (90% de nivel de concordancia) como HIV, ciertos cánceres, etc.

Conclusión: Son necesarias las dos etapas y además el análisis más incisivo cuando el G-test te avisa de que hay un problema con los índices. Pero, siempre el criterio decisivo es el del médico. Estadística es solo para ayudar, pero no para decidir.

Todas las cuentas necesarias para hacer este nuevo procedimiento (y los anteriores), se pueden hacer con un algoritmo médico desarrollado por la cátedra, y que se puede bajar libremente de su página web - www.bioestadistica.com.ar - pulsando el botón: Algoritmos.

El factor común entre todos los modelos anteriores presentados es que, en alguna parte de su deducción aparece el concepto de independencia matemática, y luego del estadígrafo obtenido se deriva la conclusión clínica. La idea básica es que: *asociación estadística, no tiene por que ser asociación clínica*. Por todo esto esta cátedra propone un camino alternativo para resolver la cuestión: Como se muestra a continuación:

Problema: Analizar la concordancia entre tres nuevos métodos clínicos y uno viejo

Caso 1

	Método 1		
Método 2	+	-	
+	180	22	202
-	10	188	198
	190	210	400

G = 4,54 * λ 95% (89 ; 95)%
λ = 92% Kappa = 0,84
Phi = 0,84 ΔS = 5,6 %
ΔE = 5,4 % ΔY = 0,1 %

Caso 2

	Método 1		
Método 3	+	-	
+	180	14	194
-	10	196	206
	190	210	400

G = 0,6 λ 95% (92 ; 96)%
λ = 94% Kappa = 0,88
Phi = 0,88 ΔS = 2 %
ΔE = 1,8 % ΔY = 0,1 %

Caso 3

	Método 1		
Método 4	+	-	
+	110	84	194
-	80	126	206
	190	210	400

G = 0,1 λ 95% (54 ; 64)%
λ = 59 % Kappa = 0,18
Phi = 0,18 ΔS = 1,2 %
ΔE = 1,2 % ΔY = 0 %

En el Caso 1 hay diferencia significativa entre la sensibilidad y especificidad de acuerdo al G-test que es significativo. Cosa que no ocurre en los otros dos casos. La diferencia potencial de sensibilidad es del 5,6%, de especificidad del 5,4% y de Youden 0,1%. La concordancia observada es muy alta (92%). Por lo tanto, de acuerdo a la visión estadística se rechaza la concordancia y con el análisis más incisivo, se ve que la variabilidad potencial de los índices no es tan alta, por lo que se podría aceptar cambiar el Método 2 por el viejo Método 1 para ciertas enfermedades.

En el Caso 2 como G-test no es significativo, se usa directamente la visión clínica y se nota que el intervalo de confianza permite aceptar un valor crítico de hasta el 96%, porque el valor observado fue del 94%. Esto es suficiente para la mayoría de las enfermedades y no hay dudas que se puede reemplazar el Método 1 con el Método 3.

En el Caso 3, G no es significativo, pero la concordancia observada es muy baja (59%). Con la visión clínica se deduce que se puede aceptar un valor crítico de concordancia de hasta el 64%; lo cual no es aceptable para muchas enfermedades. Por lo tanto el Método 4 no es aceptable desde un punto de vista clínico.

La conclusión final de tipo cualitativo, es que la mejor opción para efectuar el reemplazo del método viejo es el Método 3. Para efectuar una conclusión de tipo cuantitativo, habría que comparar estadísticamente los cuatro casos, usando un modelo que todavía no se descubre.

16.7 Resolución de las paradojas de Feinstein

En 1990 fueron publicadas dos paradojas básicas para mostrar las falencias del procedimiento de kappa, y proponer una solución para las mismas, la cual no fue aceptada con el paso del tiempo. En este punto se usará el nuevo procedimiento de visión dual para hacerlo, y de paso mostrar que el procedimiento phi tampoco las resuelve. Por lo que no se puede aceptar a este procedimiento como la solución, por el contrario hay más fallas para el mismo:

Paradoja 1: Nivel de concordancia alto, con bajo kappa y phi

Caso A

	Método 1			
Método 2	+	-		
+	40	9	49	G = 0,58
-	6	45	51	
	46	54	100	λ 95% CI (78 ; 92)

λ = 85%
kappa = 0,7
phi = 0,7

Caso B

	Método 1			
Método 2	+	-		
+	80	10	90	G = 1,64
-	5	5	10	
	85	15	100	λ 95% CI (78 ; 92)

λ = 85%
kappa = 0,32
phi = 0,33

En los dos casos anteriores A y B, el nivel de concordancia es el mismo: 85%. Sin embargo, comparando ambos casos resultan ser los valores de kappa y phi son muy diferentes entre sí. Esta disparidad inesperada ilustra las fallas de ambos procedimientos. Feinstein atribuye esta disparidad a la diferencia entre los totales marginales, porque mientras en el caso A están balanceados, no ocurre lo mismo en el Caso B. Ahora bien, con este argumento se muestra que no pueden ser aceptables desde un punto de vista clínico, porque no puede ser que la decisión del reemplazo de un método por otro, dependa de la prevalencia de la enfermedad que se encuentre.

Ambas paradojas se pueden resolver usando el procedimiento de visión dual. En ambos casos el G-test no es significativo, lo que indica que no habrá grandes diferencias de sensibilidades ni especificidades. O sea, desde el punto de vista estadístico no hay razones para rechazar la concordancia. Usando la visión clínica, se deduce que en ambos casos un nivel crítico del 92% sería aceptable. Y como para la mayoría de las enfermedades esto es suficiente, no hay dudas que se pueden intercambiar los métodos en ambos casos.

Paradoja 2: Totales marginales muy desbalanceados, producen un valor de kappa y phi más alto, que si estuvieran balanceados (como era de esperar de acuerdo a lo visto más arriba).

Los Casos C y D presentan el mismo nivel de concordancia observado: 60%. Los totales marginales del Método 2 son los mismos, pero en el Método 1 están invertidos. De acuerdo a la visión de kappa, uno esperaría encontrar mejor valor en el Caso C que en el D, pero esto no es así, sino todo lo contrario, lo que constituye una paradoja clínica para Feinstein.

Aplicando el procedimiento dual, la visión estadística muestra una diferencia entre ambos casos, mientras en el Caso C el G-test no es significativo, en el otro sí lo es. Por lo tanto, se puede aplicar la visión clínica directamente en el Caso C para observar que el nivel crítico máximo admisible es del 70%, lo cual no es suficiente para muchas enfermedades. La conclusión para este caso sería: no hacer el intercambio de métodos de acuerdo al criterio clínico.

Caso C

	Método 1			
Método 2	+	-		
+	45	15	60	G = 2,5
-	25	15	40	
	70	30	100	λ 95% CI (50 ; 70)

λ = 60%
kappa = 0,13
phi = 0,13

Caso D

	Método 1			
Método 2	+	-		
+	25	35	60	G = 25***
-	5	35	40	
	30	70	100	λ 95% CI (50 ; 70)

λ = 60%
kappa = 0,26
phi = 0,31

En el caso D hay que hacer un análisis más incisivo para estudiar las variabilidades potenciales de los tres índices principales de acuerdo a cada tipo de enfermedad. Se encuentra que $\Delta S = 42\%$,

ΔE = 38%, ΔY = 4% lo que sería inaceptable para las enfermedades del Tipo I y II, para las del Tipo III podría ser aceptable en algunos casos, sin embargo usando la visión clínica en tal caso, se concluye lo mismo que anteriormente y el intercambio de métodos no puede ser aceptado. O sea, la paradoja deja de ser tal para este procedimiento, en cambio comparando kappa con phi, sigue existiendo. Esta segunda paradoja se refleja de nuevo en los dos casos siguientes:

Caso E

	Método 1		
Método 2	+	-	
+	85	5	90
-	5	5	10
	90	10	100

λ = 90%
kappa = 0,44
phi = 0,44
G = 0
λ 95% CI (84 ; 96)

Caso F

	Método 1		
Método 2	+	-	
+	70	10	80
-	0	20	20
	70	30	100

λ = 90%
kappa = 0,74
phi = 1
Q = 10***
λ 95% CI (84 ; 96)

Aquí la paradoja es que a pesar de que los totales marginales están "simétricamente" desbalanceados en la misma dirección (similar prevalencia), los valores de kappa y phi son diferentes y en el Caso E no reflejan la alta concordancia observada (90%). Esto se resuelve usando el procedimiento de visión dual. En junio del 2003 se presentaron 3 nuevas paradojas para los demás métodos, pero que se pueden solucionar con el método de visión dual [Ver 10-1 en bibliografía].

En el Caso E, se puede ver la concordancia a simple vista. En efecto, al ser iguales las frecuencias de discordancia se deduce que G = 0 y por lo tanto desde un punto de vista estadístico, no hay diferencias. Usando la visión clínica, se detecta un valor muy alto de nivel de concordancia máximo admisible (96%). Por lo tanto, no hay dudas que el intercambio de métodos puede realizarse. Pero esto no ocurre en el Caso F, a pesar de la alta concordancia observada, porque ahora la visión estadística muestra una variabilidad significativa entre la sensibilidad y especificidad de ambos métodos. Acá no se puede aplicar el G-test porque una frecuencia es nula, y hay que usar el Q-test que resulta altamente significativo. Entonces, realizando el análisis más incisivo se ve que las variabilidades potenciales son: ΔS = 13%, ΔE = 33%, ΔY = 21% lo que sería inaceptable para mayoría de las enfermedades.

16.8 Problemas propuestos

1) Marcar la respuesta correcta a cada una de las afirmaciones siguientes, o completar la frase:

1) Bondad de ajuste significa que tan bien se parecen los datos reales a los teóricos.	**V**	**F**
2) El método clásico para efectuar una prueba de bondad de ajuste es	.....	
3) Se acostumbra a agrupar las clases con sus anexas cuando la frecuencia observada es < 3.	**V**	**F**
4) Para el cálculo de los grados de libertad se usa el número de clases agrupadas.	**V**	**F**
5) El G-test comparado con el clásico es mejor porque es	.....	
6) Cuando la frecuencia esperada en un G-test es menor que 5 hay que agrupar las clases.	**V**	**F**
7) El método de Kolgomorov-Smirnov (KS-test) busca la diferencia máxima de frecuencias.	**V**	**F**
8) El KS-test se puede usar para una sola muestra o para comparar dos entre sí.	**V**	**F**
9) Los pasos a seguir en un KS-test para una sola muestra son	.....	
10) Da lo mismo usar hipótesis extrínsecas que intrínseca en un KS-test.	**V**	**F**
11) El KS-test para dos muestras se puede usar si estas son independientes.	**V**	**F**
12) Las tablas estadísticas para el KS-test de una y dos muestras son iguales.	**V**	**F**
13) Los pasos a seguir para efectuar un KS-test en dos muestras son……	……	
14) El estadígrafo de comparación en el KS-test es el mismo.	**V**	**F**
15) Conviene ordenar los datos en forma creciente antes de usar el KS-test para 2 muestras.	**V**	**F**

16) Un test de bondad de ajuste con repetición solo permite comparar los datos agrupados. **V** **F**
17) Las pruebas que se pueden plantear en un test con repetición son ....
18) La variabilidad total en un test de bondad de ajuste se cuantifica con..... .....
19) Heterogeneidad de los datos repetidos implica encontrar significación en el G-test . **V** **F**
20) El G de datos agrupados, más el G de datos desagrupados es igual al G total. **V** **F**
21) Explicar las fallas que presentan los tests de McNemar, Log(OR), kappa y phi: …… ……
22) Explicar el método de visión dual. …… ……
23) La concordancia es un concepto estadístico. **V** **F**
24) Cuando una de las frecuencias de discordancia es nula, hay que usar el Q-test . **V** **F**
25) Ídem anterior, pero ocurre que OR es infinito y phi tiende a uno. **V** **F**
26) En la concordancia, la visión clínica es más importante que la visión estadística. **V** **F**
27) Explicar y resolver las paradojas de Feinstein. . . . . .
28) Las mismas fallas de kappa, ocurren en phi pero más ligeramente. **V** **F**

2) Aplicar el KS-test a los tres problemas vistos en el punto 16.1 y comparar las conclusiones obtenidas con las del texto usando el G-test.

3) Ídem anterior para los problemas vistos en capítulos anteriores.

4) Se efectuaron 1000 pruebas de efectividad de un nuevo medicamento, en 5 grupos de 200 individuos enfermos cada uno, elegidos al azar con el método de doble ciego. Los resultados se clasificaron en SI cuando fueron efectivos, y NO en caso contrario. Con esta información decidir si el medicamento es efectivo.

Caso	SI	NO	Total
1	120	80	200
2	110	90	200
3	98	102	200
4	130	70	200
5	125	75	200
Total	583	417	1000

5) Para los problemas siguientes se pide analizar la concordancia entre los métodos clínicos para ver si el Método 1 (el viejo) puede ser reemplazado por alguno de los nuevos:

Caso 1

	Método 1		
Método 2	+	-	
+	285	25	310
-	15	275	290
	300	300	600

Caso 2

	Método 1		
Método 3	+	-	
+	166	14	180
-	10	410	420
	176	424	600

Caso 3

	Método 1		
Método 4	+	-	
+	390	10	400
-	30	170	200
	420	180	600

Realizar el análisis con los cinco tests estadísticos tradicionales y discutir los resultados desde un punto de vista estadístico y desde un punto de vista clínico.

6) Aplicar el método de visión dual en el problema anterior.

Apéndice 1: Método de Quintin McNemar para analizar la concordancia

Cuando los clínicos deben evaluar una magnitud usando un criterio propio, a menudo no concuerdan en sus observaciones. Esto es palpable cuando se trata de medir magnitudes organolépticas tales como olor, turbidez, etc. Entre 1947 y 1955 McNemar introdujo una manera estocástica de analizar la concordancia entre dos observadores en estudios psicológicos, y a partir de allí se fue incorporando a Medicina y otras ciencias. Siguiendo su nomenclatura original, el razonamiento seguido fue:

Tabla de valores observados

	Observador 1		
Observador 2	+	-	
+	a	b	a + b
-	c	d	c + d
	a + c	b + d	N

a y d son las concordancias entre ambos
c y b son las discordancias entre los dos
N es el número total de casos analizados

La hipótesis nula planteada es que el número de los dos tipos de discordancias debe ser el mismo, para que haya una concordancia aceptable entre ambos observadores. Brevemente:

Ho) Ξ (b) = Ξ (c)

Tomando como mejor estimación del total de discordancias: Ξ (b + c) = b + c, entonces los valores esperados en las celdas respectivas serán:

Ξ (b) = Ξ (c) = (b + c) /2

Para examinar estadísticamente los resultados obtenidos, se puede usar la distribución Chi-cuadrado entre valores observados y esperados de la manera siguiente:

$$\chi 2 = \frac{\{b - [(b + c)/2]\}^2 + \{c - [(b + c)/2]\}^2}{[(b + c) /2]}$$

Resolviendo esta ecuación resulta:

$\chi 2 = (b - c)^2 / (b + c)$

La cual puede ser interpretada como una distribución Chi-cuadrado con un grado de libertad. A su vez si se efectúa la corrección de Yates por continuidad resulta:

$\chi^2 = (| b - c | - 1)^2 / (b + c)$ la cual se puede expresar como se vio más arriba con

$z^2 = (| r_{12} - r_{21} | - 1)^2 / (r_{12} + r_{21})$

La recomendación es usar la corrección por continuidad cuando (b + c) < 10. En los demás casos puede ser obviada porque no contribuye grandemente a la aproximación.

Apéndice 2: Propuesta de Feinstein y Cicchetti para resolver las paradojas de kappa

En 1990 aparecieron dos trabajos de estos autores, en el primero presentan dos paradojas que muestran fallas en el método de Kappa, y en el segundo una forma de solucionarlas. Todo esto para los casos binarios.

La idea es obtener un índice general de concordancia para los casos positivos y otro para los casos de negativos. Por ejemplo, si el total de acuerdo en positivos para el primer observador resulta ser $r_{11} + r_{12} = n_1$ y para el segundo observador es $r_{11} + r_{21} = f_1$, entonces los valores esperados para la concordancia en positivos será:

$\Xi\ (+) = (n_1 + f_1) / 2$ Análogamente para el caso de los negativos.

$\Xi\ (-) = (n_2 + f_2) / 2$ Donde $r_{22} + r_{12} = f_2$ $r_{22} + r_{21} = n_2$

Tabla de Concordancia

	Test 1		
Test 2	(+)	(-)	Total
(+)	r_{11}	r_{12}	$r_{11} + r_{12} = n_1$
(-)	r_{21}	r_{22}	$r_{21} + r_{22} = n_2$
Total	$r_{11} + r_{21} = f_1$	$r_{12} + r_{22} = f_2$	N

La proporción de positivos se puede calcular teniendo en cuenta que r_{11} es el valor observado:

$P_{pos} = r_{11} / \Xi\ (+) = 2\ r_{11} / [(n_1 + f_1)]$

$P_{neg} = r_{22} / \Xi\ (-) = 2\ r_{22} / [(n_2 + f_2)]$

El índice kappa debería ser acompañado siempre por los valores de estas dos proporciones en los informes, para tener una indicación más clara de lo que está pasando. Notar que los valores esperados se calculan dentro de las fórmulas de las proporciones.