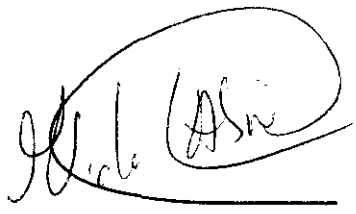
UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS

DEPARTAMENTO
DE
MATEMÁTICA APLICADA (BIOMATEMÁTICA)

PRECISIÓN DE LAS CLASIFICACIONES CLÍNICAS

Tesis Doctoral



Víctor Jesús Abraira Santos

Madrid 1997

La memoria titulada Precisión de las Clasificaciones Clínicas, de la que es autor Víctor Jesús Abraira Santos, ha sido redactada bajo mi dirección en el Departamento de Matemática Aplicada (Biomatemática) de la Universidad Complutense de Madrid. A mi juicio su contenido es científicamente valorable para constituir una tesis doctoral.

Madrid a 7 de Marzo de 1997

Fdo. Alberto Ignacio Pérez de Vargas Luque
Catedrático de Matemática Aplicada
Universidad Complutense de Madrid

A María, Vera y Celia
A mis padres

The practical application of general theorems is a different art from their establishement by matematical proof.

The science of statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data.

Ronald A. Fisher

Lo real y lo cierto no son necesariamente una misma cosa.

Salman Rushdie

ÍNDICE

Agradecimientos iii

0. Introducción 1

1. Objetivos 7

2. Marco conceptual 8
   Modelos de los procesos de medida 8
   Diseños experimentales para estimar la precisión de una medida 10

3. El problema de las clasificaciones 12
   Precisión de las clasificaciones multinomiales 20
   Índices de concordancia para múltiples observadores 28

4. Generalizaciones del índice kappa 35
   4.1 Inclusión de diseños incompletos en el índice kappa basado en definiciones explícitas de acuerdo 35
   4.2 Inclusión de pesos en el caso de múltiples observadores para variables ordinales y acuerdo por pares 36
   4.3 Otras definiciones de acuerdo 38

5. Distribución muestral del índice kappa 42

6. Descripción de los programas 49

7. Estudios observacionales 52
   7.1 Primer estudio 52
      Diseño 53
   7.2 Segundo estudio 55
      Diseño 56
   7.3 Tercer estudio 57

8. Resultados 60
   8.1 Resultados del estudio del SAT 60
      8.1.1 Contracturas articulares 60
      8.1.2 Neuropatía periférica 62
      8.1.3 Cambios esclerodermiformes 64
   8.2 Resultados del estudio del *nevus flammeus* 67
      8.2.1 Color 67
         Todos los dermatólogos 67
         Sólo los dermatólogos que tratan esta patología 69
      8.2.2 Aclaramiento 71
   8.3 Resultados del estudio sobre uso apropiado de la angioplastia 73
      8.3.1 Riesgo quirúrgico Bajo/Moderado 74

8.3.2 Riesgo quirúrgico Alto 77

9. Conclusiones 79

Apéndices 81
Apéndice I 81
Apéndice II 83
Apéndice III 102
Apéndice IV 105
Apéndice V 112

Referencias Bibliográficas 113

AGRADECIMIENTOS

Es de justicia, además de un placer, mostrar aquí mi agradecimiento a:

Alberto Pérez de Vargas, amigo, antes que director de tesis, compañero de fascinantes aventuras intelectuales y maestro en el arte de disfrutar con ellas.

Francisco Pozo, fuente inagotable de estímulos y ejemplo permanente de ciencia comprometida con la sociedad que la nutre. A él le debo, entre otras cosas, mi interés por el tema de las clasificaciones clínicas y mi admiración *weyliana* por los investigadores clínicos.

El Departamento de Matemática Aplicada (Biomatemática) de la Universidad Complutense y a todos sus integrantes, por haberme brindado, no sólo el ambiente científico propicio para la realización de este trabajo, sino, y sobre todo, un entorno amistoso, siempre dispuesto para solucionar las dudas que fueron surgiendo.

Agustín Gómez de la Cámara, Pablo Lázaro y Bibiana Pérez por haberme permitido explorar las propuestas que conforman este trabajo en el contexto de sus propias investigaciones. Ellos representan aquí a todos los médicos implicados en ellas.

Rosa Ruíz, secretaria de la Unidad de Bioestadística Clínica del Hospital Ramón y Cajal, por su amoroso cuidado de la Unidad y de las otras personas que la formamos.

Concha Muñoz, bibliotecaria del Hospital Ramón y Cajal, por su excelente trabajo como documentalista.

El Hospital Ramón y Cajal, por acoger a la Unidad de Bioestadística Clínica en cuyo seno se ha desarrollado este trabajo.

El Fondo de Investigación Sanitaria (FIS) con cuya financiación, a través de los proyectos FIS 95/1956 y FIS 96/0421, se ha realizado el trabajo aquí presentado.

0. INTRODUCCIÓN

Todo proceso de medición (proceso mediante el cual se cuantifica una magnitud) está amenazado por diversas fuentes de error, derivadas tanto de las limitaciones del instrumento de medida, como de la naturaleza de la magnitud a medir. Si se piensa, por ejemplo, en la medida de una masa mediante una balanza de precisión, el proceso consiste en colocar la masa problema en uno de los platillos de la balanza y unas masas conocidas en el otro, hasta conseguir que el fiel de la misma se equilibre. Incluso en este caso tan sencillo, existen siempre dos posibles fuentes de error:

1ª La masa es una magnitud continua[1], sin embargo el conjunto de masas conocidas que se usa como referencia (las pesas) es necesariamente discreto. En una medición, por tanto, no se obtendrá la verdadera masa sino un valor más o menos próximo a ella, formado por la suma de las masas del subconjunto de pesas que mejor equilibre el fiel. La aproximación que se consiga está limitada por la pesa más pequeña de que se disponga y, evidentemente, hay un límite para la masa más pequeña disponible, establecido sobre todo por la sensibilidad del mecanismo de la balanza para apreciar diferencias entre las mismas. A esta masa más pequeña se le denomina *precisión* y al error asociado a la misma *falta de precisión*. Es decir, si la pesa más pequeña es de 5 mg, la precisión de la balanza es de 5 mg y existirá un error máximo por falta de precisión de 5 mg, de modo que si se trata de medir la masa de un objeto de 107 mg, se obtendrá una masa de 105 ó 110 mg, dependiendo de la capacidad del observador para apreciar cuándo el fiel de la balanza está más próximo

[1] Realmente, según la teoría atómica, la masa no es una magnitud continua, sino que está formada por la suma de las masas de las partículas elementales que constituyen la materia, sin embargo al nivel macroscópico que se está considerando aquí se puede considerar como continua.

a la marca del equilibrio y del comportamiento inercial del mecanismo. Como consecuencia de esta falta de precisión, si se repite varias veces la medición de la misma masa, no siempre se obtiene el mismo resultado, es decir la precisión limitada del instrumento atenta contra la *reproducibilidad* de la medición, siendo la variabilidad de la misma un indicador de la precisión. Por lo tanto, a lo largo del texto se usarán como sinónimos los términos *precisión* y *reproducibilidad* de un instrumento de medida (lo que en la literatura en inglés se denomina *"reliability"* o *"reproducibility"*). Es de destacar que el error asociado a la falta de precisión es un *error aleatorio*, es decir, tendrá un valor distinto, e impredecible dentro de un rango, en cada una de las repeticiones de la medición.

2ª La propia construcción de la balanza: si los dos brazos no tienen exactamente el mismo peso o la misma longitud (nunca lo tendrán, debido a que también son magnitudes continuas), la medida también tendrá, por ello, un cierto error. A este error se le denomina *falta de validez* o *sesgo*, y a diferencia del anterior es un *error sistemático*, es decir tendrá siempre el mismo valor en cada una de las repeticiones de la medición, ya que depende exclusivamente del instrumento y no del proceso de medida. Si por ejemplo, el brazo de la balanza donde se coloca la masa problema tiene una masa 13 mg mayor que el otro, en todas las determinaciones que se hagan se obtendrá una masa 13 mg menor que la real.

Si bien, cuando se realiza una medición ambos errores están presentes (y de modo inseparable, si sólo se realiza una medición) conviene distinguir claramente entre ambos para entender el modo de controlarlos. De modo esquemático se puede decir que la validez depende exclusivamente del instrumento y tiene que ver con la cuestión de si el mismo mide lo que debe medir, mientras que la precisión depende

tanto del instrumento como del proceso de medición y tiene que ver con cuánto se aproxima la medida al verdadero valor de la magnitud. En ambos casos es siempre cuestión de grado, no existen instrumentos infinitamente precisos y válidos, hay sólo instrumentos más precisos y/o válidos que otros. Puede haber, por ejemplo, instrumentos muy precisos y poco válidos e instrumentos muy válidos y poco precisos: una regla de 1 m, graduada en mm pero cuya longitud real es de 1,15 m es muy precisa y poco válida, otra regla graduada en cm y cuya longitud real es 1,001 m es menos precisa pero más válida. Cualquiera de las dos resultará extraordinariamente imprecisa si se trata de medir con ellas el espesor de una hoja de papel.

Al modo habitual de controlar la validez de un instrumento de medida se le denomina *calibración*, y consiste en comparar las medidas obtenidas con él con unos patrones de referencia. Por ejemplo, para calibrar un termómetro se usan los puntos de fusión y ebullición del agua pura al nivel del mar, para calibrar un cronómetro se usan los relojes atómicos, etc. Evidentemente, como para calibrar un instrumento hay que realizar medidas y éstas están sujetas al error aleatorio producido por la precisión limitada del mismo, el proceso de calibrado es instrumentalmente dependiente, aunque conceptualmente distinto, del control de la precisión que se comenta a continuación.

El modo habitual de controlar la precisión de un instrumento es comparar entre sí medidas repetidas de un mismo objeto y evaluar el grado de acuerdo o concordancia entre ellas. Cuanto más parecidas son dichas medidas, más preciso es el instrumento; como se comentó más arriba, la variabilidad de las mismas es un indicador de la precisión.

En las mediciones biológicas, y en particular en la práctica clínica, el proceso de control de la precisión y validez de una medida es más complejo que el esbozado hasta aquí, debido a dos fenómenos inherentes a las mismas y que, hasta ahora, no se han considerado. De un lado, las magnitudes a medir son *aleatorias*, es decir presentan diversos grados de variabilidad impredecible propia. Si, por ejemplo, se trata de controlar la precisión de la medición de la presión arterial diastólica habría que repetir la medición en un mismo individuo varias veces y como la propia presión arterial es variable a lo largo del tiempo, el resultado se vería afectado simultáneamente por la variabilidad introducida por la precisión limitada del instrumento (manómetro y observador) y la variabilidad propia de la presión arterial. En el diseño de estudios para establecer la precisión y validez de las mediciones biológicas habrá que tener esto muy en cuenta.

Por otro lado, además de magnitudes tales como presión, temperatura, concentración de substancias químicas en el líquido plasmático, etc., se trabaja con magnitudes como dolor, mejoría en un proceso patológico, grado pronóstico de una afección, actitudes comportamentales etc., para las cuales no existe un patrón de referencia claro y objetivo ni escala métrica apropiada. Tales magnitudes suelen describirse en escalas ordinales o, incluso, nominales, cuya apreciación puede estar muy distorsionada por influencias subjetivas. Estas magnitudes suelen denominarse *variables blandas* (Feinstein [1]) y dan lugar a clasificaciones mejor que a mediciones en sentido estricto (que implican la existencia de una escala métrica), aunque en este texto, a veces, se usará el término de medición en un sentido amplio que incluye también las clasificaciones. Evidentemente, existen también variables objetivas (*"duras"* en la terminología clínico epidemiológica) que dan lugar a clasificaciones, por ejemplo muerte/vivo. Los procesos de clasificación sufren los mismos problemas

de validez y precisión que los de medición, pero con ciertas complicaciones añadidas en el caso de las variables blandas. Para controlar su validez, no suelen existir patrones de referencia, o no son tan objetivos o accesibles como en el caso de una magnitud física. Por ejemplo, en la calibración de las imágenes obtenidas por resonancia nuclear magnética (u otro procedimiento radiológico) para diagnosticar lesiones de menisco, ¿cuál es el patrón de referencia adecuado? ¿la visión directa mediante artroscopia? ¿es ésta suficientemente objetiva, o también está influida por factores subjetivos dependientes de la propia técnica, de la experiencia del médico que la realiza, de la diversidad de meniscos "normales"?. En este sentido se suele distinguir entre dos modos de controlar la validez de un instrumento de medida[2]: cuando se hace con patrones objetivos se habla de *exactitud* (*"conformity"*), mientras que cuando se controla comparando simplemente con una referencia considerada mejor ("patrón de oro", o *"gold standard"* en la literatura en inglés) se habla de *conformidad* (*"accuracy"*).

En cuanto a la reproducibilidad, sobre todo con métodos de clasificación, se distingue entre la reproducibilidad del mismo instrumento (típicamente un observador en este caso) en dos instantes de tiempo diferentes y se habla de *concordancia* (*"agreement"* en la literatura en inglés) *o consistencia interna o intraobservador*, por ejemplo un radiólogo o un servicio de radiología tomado como unidad, ¿clasifica igual la misma radiografía estudiada hoy y dos meses después?, y reproducibilidad del mismo instrumento usado en diferentes condiciones, por ejemplo dos radiólogos diferentes ¿clasifican del mismo modo la misma radiografía?,

[2] Nótese que se está usando el término *instrumento de medida* en un sentido muy amplio, en este ejemplo no es sólo el "aparato" usado para obtener la imagen, sino el conjunto formado por el aparato que produce la imagen y el observador que la interpreta, siendo, además, este último más crítico para los errores de medición-clasificación.

se habla entonces de *concordancia o consistencia externa o interobservador.*

Estos conceptos se resumen esquemáticamente en la Tabla 1.

Tabla 1

Esquema de conceptos asociados a errores de medición

Precisión o reproducibilidad o concordancia	concordancia intraobservador
	concordancia interobservador
Validez	exactitud
	conformidad

1. OBJETIVOS

Los objetivos de esta tesis son:

1) Generalizar el índice de concordancia κ, propuesto inicialmente por Cohen [2] para evaluar la precisión de las clasificaciones binomiales con dos repeticiones, al caso más general de clasificaciones multinomiales ordinales, con más de dos repeticiones, incluyendo diseños incompletos, distintos sistemas de "pesos" y distintas definiciones de acuerdo.

2) Desarrollar un programa de ordenador, integrado en el paquete estadístico PRESTA [3], que implemente dicho índice y su error estándar estimado por la técnica "jackknife", así como otros contrastes relevantes en el proceso de evaluación de la precisión de los sistemas de clasificación.

3) Aplicarlo en tres estudios para evaluar la precisión de clasificaciones clínicas sobre:

   3.1 neuropatía periférica, cambios esclerodermiformes de la piel y contracturas articulares.

   3.2 color del *nevus flammeus* y su aclaramiento por láser de colorante pulsado.

   3.3 uso apropiado de la angioplastia coronaria transluminal percutánea.

2. MARCO CONCEPTUAL

Modelos de los procesos de medida

El modelo más sencillo de un proceso de medida es el siguiente:

$$X = A + \varepsilon \qquad [2.1]$$

donde X es una variable aleatoria que representa el resultado de la medida, A una constante que representa el verdadero valor de la magnitud a medir y ε otra variable aleatoria que representa el error de la medida. Nótese que en este modelo, lo único observable es X.

La variabilidad de la medida está contenida en ε y la medida más habitual de esta variabilidad es su varianza, por lo tanto para estimar la precisión del instrumento se debe estimar la $Var(\varepsilon)$. Si el instrumento fuera perfectamente preciso $Var(\varepsilon)=0$.

La medida más común de centralización de una variable es su valor esperado o medio, por lo tanto para estimar la validez se debe comparar A con el valor esperado de X. Si el instrumento fuera perfectamente válido, se verificaría $E(X)=A$.

En consecuencia, una medida del sesgo es $E(X) - A$ y una medida de la precisión es $Var(\varepsilon)$ (cuanto mayores sean ambas, menor validez y precisión, respectivamente).

Cuando la magnitud que se pretende medir no es constante, sino que es también una variable aleatoria (el caso de la presión arterial o la imagen del menisco considerados en la introducción) el modelo toma la forma:

$$X = Y + \varepsilon \qquad [2.2]$$

donde, en lugar de la constante A para representar la magnitud de interés, figura la variable Y. Generalmente se asume que las variables Y y ε son independientes, es decir, el error de la medida es independiente de la variable a medir. Un modo alternativo de escribir [2.2] es:

$$X = \mu + \varepsilon_Y + \varepsilon \qquad [2.3]$$

donde se ha descompuesto la variable Y en la suma de su media μ (constante) y la variable ε_Y que contiene toda la variabilidad de Y alrededor de la media, por lo tanto $E(\varepsilon_Y)=0$. La asunción de independencia es, ahora, entre las variables ε_Y y ε.

Si se calculan valores esperados en [2.3], teniendo en cuenta que el valor esperado es lineal, se obtiene:

$$E(X) = E(\mu) + E(\varepsilon_Y) + E(\varepsilon) = \mu + E(\varepsilon) \qquad [2.4]$$

Una medida del sesgo será, ahora, la diferencia $E(X) - \mu$ y para su estimación será necesario tener estimadores de ambas medias, la de la medida y la de la variable.

Si en [2.4] se calculan varianzas, teniendo en cuenta que la asunción de independencia entre ε_Y y ε implica que $Cov(\varepsilon_Y,\varepsilon)=0$, resulta:

$$Var(X) = Var(\varepsilon_Y) + Var(\varepsilon) + 2Cov(\varepsilon_Y,\varepsilon) = Var(\varepsilon_Y) + Var(\varepsilon) \qquad [2.5]$$

es decir, la varianza de la medición observada tiene dos componentes: uno es la varianza de la propia variable y otro la del error aleatorio de la medida; por lo tanto, con este modelo (el más común en Biología, como se dijo en la introducción) la varianza de la medida observada no es un buen indicador de la precisión, en su lugar se usa el cociente entre la varianza de la variable y la varianza de la medida, propuesto por Fisher [4] y denominado *coeficiente de correlación intraclase* (*CCI* o ρ):

$$\rho = \frac{Var(\varepsilon_Y)}{Var(X)} = \frac{Var(\varepsilon_Y)}{Var(\varepsilon_Y) + Var(\varepsilon)} \qquad [2.6]$$

asumiendo que la variable *Y* es una variable aleatoria y que por lo tanto su varianza no es cero, este índice tomará valores, sin alcanzar nunca esos límites, entre 1, cuando $Var(\varepsilon)$ tiende a cero, es decir si la medida fuera absolutamente precisa, y 0 si $Var(\varepsilon)$ tiende a infinito, es decir si la medida fuera infinitamente imprecisa. Este coeficiente puede ser interpretado, en ciertas circunstancias (Bartko [5]), como el coeficiente de correlación entre dos mediciones de la misma magnitud.

Diseños experimentales para estimar la precisión de una medida

Un estudio para establecer la validez de un instrumento, según el modelo [2.1], consistirá en repetir la medida para un mismo objeto cuya magnitud *A* sea conocida (exactitud) o asumida como tal (conformidad), lo que produce una muestra aleatoria

de la variable X y a partir de ella estimar $E(X)$. Su mejor estimador es la media muestral del conjunto de las medidas ($\overline{X}$). La estimación del sesgo se hará, por lo tanto, a partir de la variable:

$$\overline{X} - A \qquad [2.7]$$

Un estimador de la precisión será un estimador de $Var(\varepsilon)$, ahora bien, ε no es directamente observable, sin embargo si en el modelo [2.1] se calcula la varianza de X, teniendo en cuenta que A es una constante y que, por lo tanto, su varianza es cero, resulta:

$$Var(X) = Var(\varepsilon) \qquad [2.8]$$

es decir, se puede estimar la precisión por la varianza de X, que sí es observable. El mejor estimador de la varianza de X es su varianza muestral (S^2).

El diseño experimental más sencillo[3] para establecer la precisión de un instrumento de medida de una variable aleatoria, según el modelo [2.2] o [2.3], consistirá en efectuar K medidas repetidas sobre una muestra aleatoria de N sujetos, o lo que en análisis de la varianza se denomina *análisis de la varianza de medidas repetidas* o *de bloques completos aleatorios*. Este análisis produce estimadores centrados de $Var(\varepsilon)$ y $Var(\varepsilon_Y)$ y por lo tanto permite estimar el *CCI*, así como realizar contrastes de hipótesis sobre él (Bartko [5]) en el caso de que la variable a medir cumpla las asunciones del análisis de la varianza (normalidad y homocedasticidad).

[3] Otros diseños más complejos, por ejemplo para separar la variabilidad introducida por el observador de la del instrumento propiamente dicho, pueden verse en Latour, Abraira y col. [6].

3. EL PROBLEMA DE LAS CLASIFICACIONES

En el caso de las clasificaciones, la distribución de la variable resultado de la clasificación es binomial (clasificación en dos categorías) o multinomial (varias categorías) y, por tanto no se cumplen las asunciones del análisis de la varianza. Aunque se han propuesto estadísticos tipo *CCI*, y estudiado modelos para su distribución muestral (Rosner [7], Donner y Donald [8]), para estimar la precisión de las clasificaciones, el estadístico más usado en la literatura clínica es el coeficiente κ propuesto inicialmente por Cohen [2] para el caso más sencillo de una clasificación con dos categorías (binomial) denominadas "positivo" y "negativo", por ejemplo: enfermo (positivo) y no enfermo (negativo), presencia (positivo) o ausencia (negativo) de un síntoma, etc., realizada por dos observadores. Para caracterizar su precisión será necesario establecer la concordancia tanto interna como externa. En ambos casos se procede del mismo modo: se observa un número de individuos y se clasifican por dos observadores independientemente, o por el mismo en dos instantes de tiempo diferentes y suficientemente separados para poder asumir su independencia. Los resultados se pueden resumir en una tabla de frecuencias de doble entrada como la Tabla 2.

Tabla 2

Clasificación binaria con dos observadores

	Observador A		
Observador B	Positivo	Negativo	Total
Positivo	X_{11}	X_{12}	$X_{1.}$
Negativo	X_{21}	X_{22}	$X_{2.}$
Total	$X_{.1}$	$X_{.2}$	N

El convenio de subíndices usado en esta tabla es el usual de la notación matricial, es decir, para identificar una celda se usan dos subíndices: el primero para la fila y el segundo para la columna, por lo tanto X_{ij} es la variable que representa el número de individuos que el observador A ha clasificado en la categoría j y el observador B en la i. Para indicar los totales marginales, se usa un punto en el lugar del subíndice con respecto al que se ha sumado.

Existen varios índices de concordancia, revisados por Fleiss [9], el más obvio es la proporción de acuerdos observados, es decir, para la Tabla 2:

$$P_o = \frac{X_{11} + X_{22}}{N} \qquad [3.1]$$

Este índice es muy intuitivo y fácilmente interpretable: tomará valores entre 0 (total desacuerdo) y 1 (máximo acuerdo). Sin embargo como indicador de reproducibilidad tiene el inconveniente de que aun en el caso de que los dos observadores

clasifiquen con criterios independientes se producirá un cierto grado de acuerdo por azar. Por ejemplo, si se tiran dos dados y se registra si sale un cierto número, p.e. el dos (resultado positivo) u otro cualquiera (resultado negativo), en un cierto número de veces (con una probabilidad de 26/36, para ser más preciso, siguiendo el razonamiento que se hace más abajo) ambos dados producirán el mismo resultado. Es deseable que un índice de concordancia tenga en cuenta este hecho y que, de algún modo, indique el grado de acuerdo que existe por encima del esperado por azar. En este sentido, la propuesta de Cohen es usar el *índice kappa* (κ) definido por:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad [3.2]$$

siendo P_o la proporción de acuerdos observados y P_e la proporción de acuerdos esperados bajo la hipótesis de independencia entre los observadores, o dicho de otro modo, de acuerdos por azar. Para calcular P_e debe recordarse que dos sucesos A_1, A_2 son independientes si $p(A_1 \cap A_2) = p(A_1)p(A_2)$, es decir si la probabilidad de que ocurran simultáneamente es el producto de las probabilidades de cada uno de ellos. En este caso, si llamamos A_1 al suceso: "el observador A clasifica un individuo como positivo" y A_2 al suceso: "el observador B clasifica un individuo como positivo" el suceso $A_1 \cap A_2$ será: "ambos observadores clasifican un individuo como positivo"; como el observador A ha clasificado $X_{.1}$ individuos como positivos y el B ha clasificado $X_{1.}$ el mejor estimador de $p(A_1)$ es $X_{.1}/N$ y el de $p(A_2)$ es $X_{1.}/N$; por lo tanto el mejor estimador de la probabilidad de que ambos clasifiquen como positivo a un individuo, bajo la hipótesis de que ambos son independientes es su producto, es decir $X_{.1}\ X_{1.}/N^2$; por la misma razón, la probabilidad de que ambos clasifiquen como negativo a un individuo por azar es $X_{2}\ X_{2.}/N^2$; en consecuencia:

$$P_e = \frac{X_{.1}\, X_{1.} + X_{.2}\, X_{2.}}{N^2} \qquad [3.3]$$

Cuando hay acuerdo total, $X_{12} = X_{21} = 0$; por lo tanto el valor de P_o es 1 y en consecuencia el índice κ también vale 1 para el máximo acuerdo; si el acuerdo observado es igual al esperado por azar, κ vale 0. Obsérvese que si el acuerdo observado es menor que el esperado por azar, el índice κ toma valores negativos. Un modo intuitivo de interpretar este índice puede hacerse despejando P_o de [3.2]:

$$P_o = \kappa + (1 - \kappa)\, P_e \qquad [3.4]$$

si se piensa que la proporción de acuerdos observados (P_o) es un valor intermedio entre el máximo acuerdo posible (1) y la proporción de acuerdos esperados por azar (P_e), en la fórmula [3.4], κ se puede interpretar como el peso que el máximo acuerdo posible tiene en los acuerdos observados.

Para facilitar su interpretación Landis y Koch [10] propusieron, y desde entonces ha sido ampliamente usada, la escala de valoración del índice κ que figura en la Tabla 3.

Posteriormente Fleiss [11] simplifica esta escala a sólo tres niveles: ≤0,40: pobre, entre 0,40 y 0,75: de moderado a bueno y ≥0,75: excelente. No obstante el alto grado de aceptación de ambas escalas en la literatura clínico-epidemiológica (p.e. Elmore et al. [12], Jelles et at. [13]) debe tenerse en cuenta que, como los propios autores resaltan, son arbitrarias y que, además, el valor del índice κ no sólo depende de los acuerdos observados, sino también de los esperados y, por lo tanto, pueden

Tabla 3

Valoración del índice κ propuesta por Landis y Koch

kappa	grado de acuerdo
< 0,00	sin acuerdo
0,00 - 0,20	insignificante
0,21 - 0,40	mediano
0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

darse diversos efectos poco intuitivos (Feinstein y Cicchetti [14], Cicchetti y Feinstein [15], Guggenmoss-Holzmann [16]). En primer lugar, el valor de κ depende de la prevalencia del carácter observado. Considérese, por ejemplo, la Tabla 4 correspondiente a un estudio hipotético de precisión entre dos radiólogos interpretando si una radiografía muestra evidencias de pulmonía (resultado positivo) o no (resultado negativo). Según [3.1] y [3.2], para esta tabla, la proporción de acuerdos observados es 0,84 y el índice κ es 0,245. En este ejemplo, la prevalencia de pulmonía es baja: el radiólogo A diagnostica un 14% de pulmonías y el B un 10%. Para los resultados de la Tabla 5, en la que las prevalencias respectivas son 40% para A y 36% para B, con la misma proporción de acuerdos observados (84%) el índice κ toma el valor 0,661.

Tabla 4

Clasificación de radiografías por dos radiólogos

	Radiólogo A		
Radiólogo B	Pulmonía	No pulmonía	Total
Pulmonía	4	6	10
No pulmonía	10	80	90
Total	14	86	100

Tabla 5

	Radiólogo A		
Rad. B	30	6	36
	10	54	
	40		

En general, cuanto más cercana a 0,5 sea la prevalencia (cuanto más balanceados estén los totales marginales en la tabla) mayor es el κ para igual proporción de acuerdos observados; dicho de otro modo, prevalencias muy bajas, o muy altas, penalizan el índice κ, debido a que en ese caso la proporción de acuerdos esperados por azar es mayor que cuando la prevalencia es cercana a 0,5.

En el supuesto, simplista, de que ambos observadores clasifiquen con la misma prevalencia, $X_{.1} = X_{1.} = X_1$, la expresión [3.3] se puede escribir:

$$P_e = \frac{X_1^2 + (N - X_1)^2}{N^2} = \frac{2X_1^2 + N^2 - 2NX_1}{N^2} = 2P^2 + 1 - 2P$$

representando por P a la prevalencia común. Si se sustituye en [3.2] se obtiene:

$$\kappa = \frac{P_o - (1 + 2P^2 - 2P)}{2P - 2P^2} = 1 + \frac{P_o - 1}{2P - 2P^2}$$

que hace explícita la dependencia del índice κ con respecto a la prevalencia. En la Figura 1 se presentan curvas para distintos valores de P_o, junto con valores de κ para distintas tablas simuladas (**Apéndice I**) en la situación más realista de distintas prevalencias para los dos observadores.

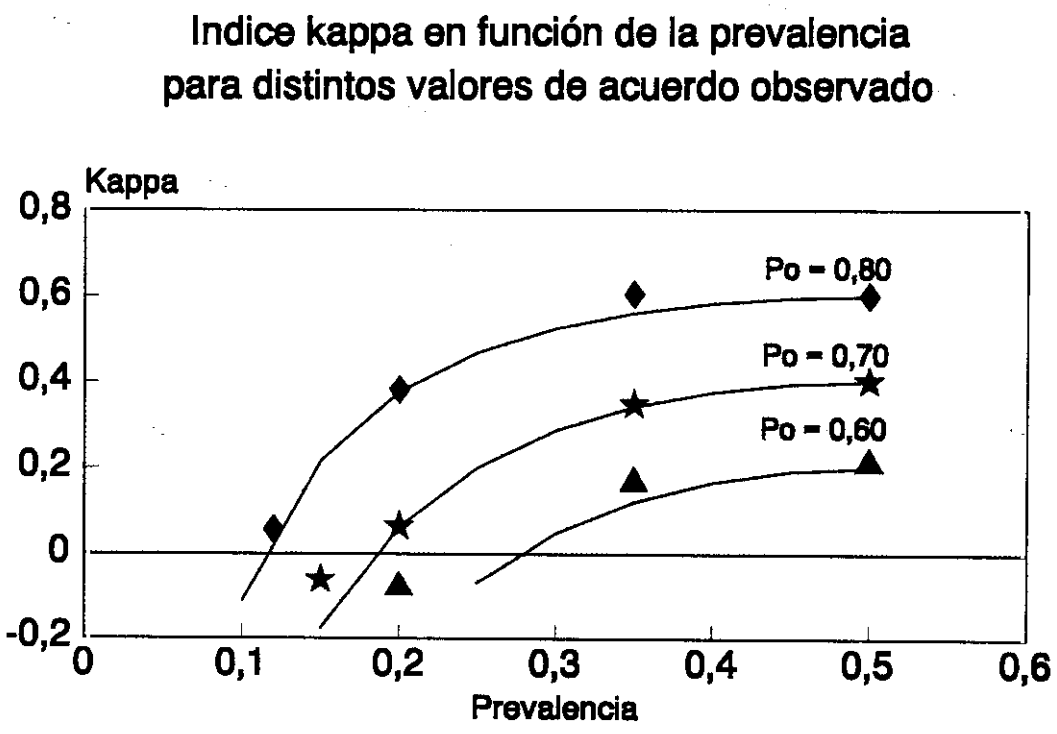


Figura 1

Por otro lado, el índice κ también se ve afectado por la simetría de los totales marginales. Considérese, por ejemplo, las Tablas 6.1 y 6.2:

Tabla 6.1

	Radiólogo A		
Rad. B	45	15	60
	25	15	
	70		

Tabla 6.2

	Radiólogo A		
	25	35	60
	5	35	
	30		

en ambas la proporción de acuerdos observados es la misma (0,60) y también es la misma la prevalencia observada por el radiólogo B (0,60), sin embargo la del radiólogo A es 0,70 en la Tabla 6.1 y 0,30 en la 6.2; por lo tanto hay mayor desacuerdo entre las prevalencias observadas en la Tabla 6.2, aunque en ambos casos están igualmente alejadas de 0,5, es decir, tienen la misma falta de balanceo en los totales marginales, aunque para la Tabla 6.1 de modo simétrico con respecto a ambos observadores (en ambos son mayores de 0,5) y asimétricamente para la Tabla 6.2 (para A es menor de 0,5 y para B mayor). El índice κ vale 0,13 en la tabla 6.1 y 0,26 en la tabla 6.2. En general, la simetría en la falta de balanceo en los totales marginales también penaliza el índice y tanto más, cuanto más "perfecta" (la misma diferencia con respecto a 0,5) sea la misma. Dicho de otro modo, en igualdad de acuerdos observados, cuanto mayor sea la diferencia entre las prevalencias observadas[4] por

[4] Algunos autores, p.e. Brennan y Silman [17], también denominan sesgo a la diferencia entre observadores. En este texto, el término sesgo se reserva para la diferencia entre una observación y la *verdadera* magnitud, por lo tanto en un estudio de concordancia no se hablará de sesgo, sino de diferencia o falta de homogeneidad de las frecuencias marginales.

cada observador mayor es el índice κ (esto es lo que Feinstein y Cicchetti [14,15] denominan segunda paradoja de kappa y que en nuestra opinión sí constituye una paradoja, mientras que la dependencia de kappa de la prevalencia no es inmediatamente intuitivo, pero no es paradójico, simplemente es consecuencia de que cuanto más alejada esté la probabilidad de un resultado de 0,5 mayor es la probabilidad de coincidencia por azar).

En consecuencia, para interpretar el índice κ es necesario contar, también con el valor de las frecuencias marginales de la tabla (prevalencias observadas por cada observador).

El pequeño valor de κ para los datos de la Tabla 4 (mediano en la escala de Landis y Koch y pobre en la de Fleiss) es "explicado" a la luz de los efectos anteriores por el hecho de que estamos en la peor de las situaciones posibles: baja prevalencia, y similar, en ambos observadores o, en la terminología anterior totales marginales "desbalanceados" con casi perfecta simetría.

Precisión de las clasificaciones multinomiales

Las clasificaciones con sólo dos categorías consideradas en el apartado anterior tienen la ventaja de su sencillez, sin embargo a menudo resultan insuficientes en la práctica clínica. Con frecuencia se realizan clasificaciones multinomiales, es decir con más de dos categorías. Por ejemplo, un psiquiatra clasifica los trastornos de los pacientes en psicóticos, neuróticos u orgánicos, o un reumatólogo clasifica las artritis en leves, moderadas o graves. Ambas clasificaciones son multinomiales (tres categorías), no obstante existe una diferencia entre ellas, las categorías en el caso de la artritis

pueden ordenarse de un modo relevante para el problema: una artritis grave es más que una moderada, y ésta más que una leve, mientras que para la clasificación psiquiátrica este orden no existe. A las variables multinomiales que tienen implícito un orden se les denomina *ordinales* y a las que no, *nominales*. Para estudiar la precisión de una clasificación multinomial, hay ciertas diferencias según que ésta sea ordinal o nominal.

Como en las clasificaciones binarias, los resultados de un estudio de concordancia se pueden resumir en una tabla de doble entrada como la Tabla 7, que supone una generalización de la Tabla 2, ahora con K filas y K columnas, siendo K el número de categorías de la clasificación.

Tabla 7

Clasificación multinomial con dos observadores

	Observador A				
Obser. B	Cat. 1	Cat. 2	...	Cat. K	Total
Cat. 1	X_{11}	X_{12}	...	X_{1K}	$X_{1.}$
Cat. 2	X_{21}	X_{22}	...	X_{2K}	$X_{2.}$
.	.	.	...	.	.
Cat. K	X_{K1}	X_{K2}	...	X_{KK}	$X_{K.}$
Total	$X_{.1}$	$X_{.2}$	...	$X_{.K}$	N

Para una clasificación multinomial se puede definir un índice kappa idéntico a [3.2], generalizando el cálculo de P_o y P_e por medio de:

$$P_o = \frac{\sum_{i=1}^{K} X_{ii}}{N} \qquad P_e = \frac{\sum_{i=1}^{K} X_{i.} X_{.i}}{N^2} \qquad [3.5]$$

Otra alternativa para estudiar concordancia entre clasificaciones multinomiales consistë en definir un índice kappa para cada una de las categorías, colapsando la tabla $K \times K$ original en K tablas 2×2 en las que se compara cada categoría con todas las demás. De este modo se puede estudiar la contribución de cada una de ellas a la concordancia de la clasificación.

Es fácil ver (Fleiss [11]) que el kappa global es una media ponderada de los kappas individuales: el kappa global se puede expresar como suma de los numeradores de los kappas individuales dividida por la suma de sus denominadores.

Una solución que puede verse como intermedia entre las anteriores (un único kappa global o K kappas individuales para cada categoría), pero que, en general, sólo tiene sentido para variables ordinales, es el denominado *kappa ponderado*, también propuesto por Cohen [18], en el cual se asignan unos pesos para cuantificar la importancia relativa entre los desacuerdos.

La idea de este índice ponderado es asignar a cada celda de la tabla un peso w_{ij} comprendido entre 0 y 1 que represente la importancia relativa del desacuerdo. Se atribuye el máximo peso al acuerdo perfecto, y pesos proporcionalmente menores según la importancia del desacuerdo. Se tendrá:

$$w_{ii} = 1 \qquad 0 \leq w_{ij} < 1 \quad \forall i \neq j$$

además, obviamente, se impone exigir simetría, es decir:

$$w_{ij} = w_{ji}$$

Las proporciones ponderadas de acuerdos observados y esperados se definen del modo siguiente:

$$P_{o(w)} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\, X_{ij}}{N} \qquad P_{e(w)} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\, X_{i.}\, X_{.j}}{N^2} \qquad [3.6]$$

y a partir de aquí, el kappa ponderado:

$$\kappa_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}} \qquad [3.7]$$

Nótese que, en el caso extremo, si se definen los pesos por:

$$w_{ij} = 0 \quad \forall i \neq j\ ; \qquad w_{ii} = 1$$

el kappa ponderado coincide con el kappa global.

En el otro extremo, si se definen:

$$w_{ii} = 1 \quad \forall i; \quad w_{ij} = 1 \quad \forall i \neq m,\ j \neq m; \qquad w_{mj} = 0 \quad \forall j \neq m$$

el kappa ponderado es el kappa individual de la categoría m. En este sentido es en

el que, como se dijo anteriormente, el kappa ponderado constituye una situación intermedia entre el kappa global y los kappas individuales.

Como ya se indicó, la principal ventaja del kappa ponderado no consiste en esta formulación unificadora de los kappa global e individuales, sino en la posibilidad de cuantificar diferentes grados de desacuerdo. Los valores de los pesos dependerán, en cada caso, de la importancia que se conceda a cada desacuerdo. Hay que tener presente, sin embargo, que ello añade cierta dificultad a su interpretación: si en dos estudios diferentes, se calcula el kappa ponderado con dos sistemas de pesos distintos, es difícil realizar comparaciones entre ellos. Por ello algunos autores, por ejemplo Maclure y Willet [19], sugieren limitarse a usar sistemas de pesos estándar. Los más usados en este sentido son, por su sencillez, los denominados *pesos lineales*, propuestos inicialmente por Cohen [18]:

$$w_{ij} = 1 - \frac{|i - j|}{k - 1} \qquad [3.8]$$

y, por la equivalencia entre el kappa ponderado resultante y el coeficiente de correlación intraclase, los denominados *pesos bicuadrados*, propuestos posteriormente por Fleiss y Cohen [20]:

$$w_{ij} = 1 - \frac{(i - j)^2}{(k - 1)^2} \qquad [3.9]$$

Sin embargo, Graham y Jackson [21] demuestran que el índice kappa con pesos bicuadrados es más un índice de asociación que de acuerdo, ya que puede ser insensible a diferencias en las proporciones de acuerdo observado.

Landis y Koch [10] sugieren usar un sistema de pesos en relación jerárquica para identificar la o las categorías más discrepantes. En el trabajo citado se analiza el acuerdo entre dos grupos de neurólogos en la clasificación de esclerosis múltiple en las 4 categorías siguientes:

1: cierta

2: probable

3: posible (relación 1:1)

4: dudosa, improbable o claramente sin esclerosis

Los resultados para un grupo de 69 pacientes estudiados se muestran en la Tabla 8 y los pesos jerárquicos propuestos figuran en la Tabla 9.

Se denominan pesos jerárquicos porque cada uno considera como concordantes todas las categorías que se consideran concordantes en el anterior. Nótese que el kappa ponderado con los pesos w_1 es el kappa global sin ponderar; el kappa ponderado con los pesos w_2 es el kappa global que resultaría para una tabla 3×3 en la que se hubieran colapsado en una sola las dos primeras categorías, por lo tanto la diferencia entre κ_{w_1} y κ_{w_2} informa sobre la contribución al desacuerdo total del desacuerdo entre las categorías 1 y 2; el kappa ponderado con los pesos w_3 es el kappa global que resultaría para una tabla 2×2 en la que se hubieran colapsado en una sola las dos primeras categorías y en otra las otras dos; finalmente los pesos w_4 dan lugar a un kappa en el que no se consideran desacuerdos las discrepancias entre las categorías contiguas.

Tabla 8

		Grupo A				
	Diag.	1	2	3	4	Total
Grupo B	1	5	3	0	0	8
	2	3	11	4	0	18
	3	2	13	3	4	22
	4	1	2	4	14	21
	Total	11	29	11	18	69

Tabla 9

Pesos jerárquicos propuestos por Landis y Koch

Peso		w_1				w_2				w_3				w_4			
Obs.		A				A				A				A			
	Diag	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
B	1	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0
	2	0	1	0	0	1	1	0	0	1	1	0	0	1	1	1	0
	3	0	0	1	0	0	0	1	0	0	0	1	1	0	1	1	1
	4	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	1

Los kappas ponderados para los datos de la Tabla 8, correspondientes a estos pesos son $\kappa_{w_1} = 0{,}297$ $\kappa_{w_2} = 0{,}332$ $\kappa_{w_3} = 0{,}386$ y $\kappa_{w_4} = 0{,}789$, que ilustran claramente que las discrepancias se producen, fundamentalmente, entre las categorías 2 y 3, ya que el kappa sólo aumenta sensiblemente para los pesos w_4, y estos son los únicos que no distinguen entre dichas categorías.

Usando la misma representación matricial, los pesos lineales [3.8] y bicuadrados [3.9] para una clasificación en 4 categorías se muestran en la Tabla 10.

Tabla 10

Pesos lineales y bicuadrados para cuatro categorías

Peso		lineal (w_l)				bicuadrado (w_b)			
Obs.		A				A			
	Diag	1	2	3	4	1	2	3	4
B	1	1	2/3	1/3	0	1	8/9	5/9	0
	2	2/3	1	2/3	1/3	8/9	1	8/9	5/9
	3	1/3	2/3	1	2/3	5/9	8/9	1	8/9
	4	0	1/3	2/3	1	0	5/9	8/9	1

Indices de concordancia para múltiples observadores

Todo sistema de clasificación aplicable en la práctica clínica, realmente es usado por múltiples observadores, por lo tanto conviene evaluar también su reproducibilidad para más de dos observadores. En la literatura han aparecido diversas aproximaciones a este problema: la más sencilla y obvia (Dunn [22]) y quizás por ello todavía utilizada (Elmore y col. [12]) consiste en calcular el kappa [3.2], o el kappa ponderado [3.7], para cada uno de todos los pares posibles de observadores y resumir todos ellos a través de su valor medio y cualquiera de los indicadores habituales de dispersión, por ejemplo su rango o su desviación típica. Si existen J observadores, se pueden formar $\binom{J}{2}$ pares distintos, en consecuencia se pueden calcular $\binom{J}{2}$ kappas, a partir de los cuales se obtendría su media y desviación.

Otra aproximación, debida a Landis y Koch [23] y basada en la identidad entre el índice kappa y el coeficiente de correlación intraclase (Fleiss y Cohen [20]) y, por lo tanto, válida en principio sólamente para clasificaciones binomiales consiste en el cálculo del *CCI* a partir del análisis de la varianza de una vía (los diferentes observadores). Sean N sujetos, J_i el número de observadores que clasifican al sujeto i y X_i el número de clasificaciones positivas para el sujeto i, evidentemente $J_i - X_i$ es el número de clasificaciones negativas para el sujeto i. El *CCI* se calcula del siguiente modo:

$$\kappa = r = \frac{BMS - WMS}{BMS + (J_0 - 1)WMS} \qquad [3.10]$$

siendo *BMS* los *cuadrados medios entre observadores*, *WMS* los *cuadrados medios dentro de los observadores* y siendo:

$$J_0 = \bar{J} - \frac{\sum_{i=1}^{N}(J_i - \bar{J})^2}{N(N-1)\bar{J}} \qquad [3.11]$$

Los cuadrados medios se estiman del siguiente modo:

$$BMS = \frac{1}{N} - \sum_{i=1}^{N} \frac{(X_i - J_i\bar{P})^2}{J_i} \qquad WMS = \frac{1}{N(\bar{J}-1)} \sum_{i=1}^{N} \frac{X_i(J_i - X_i)}{J_i} \qquad [3.12]$$

donde:

$$\bar{P} = \frac{\sum_{i=1}^{N} X_i}{N\bar{J}} \qquad \bar{J} = \frac{\sum_{i=1}^{N} J_i}{N} \qquad [3.13]$$

Fleiss y Cuzick [24] construyeron la distribución asintóticamente normal de este índice y derivaron su error estándar, lo que, en el caso de muestras grandes, permite contrastar hipótesis sobre él.

Como extensión de esta aproximación para clasificaciones multinomiales con *K* categorías, Landis y Koch [23] proponen calcular, para cada categoría *k=1,...,K*, un índice κ_k calculado por [3.10] y usar como índice global de acuerdo el siguiente:

$$\bar{\kappa} = \frac{\sum_{k=1}^{K} \bar{P}_k \bar{Q}_k \kappa_k}{\sum_{k=1}^{K} \bar{P}_k \bar{Q}_k} \qquad [3.14]$$

y para el caso particular de un número igual de observadores por sujeto, Fleiss, Nee y Landis [25] construyeron la distribución asintóticamente normal y derivaron el error estándar, tanto de κ_k como de $\bar{\kappa}$, por lo tanto también para [3.14], en el caso

particular aludido, es posible realizar contrastes de hipótesis cuando se dispone de muestras grandes.

Otra aproximación diferente, más elaborada, implica una definición explícita de qué se entiende por acuerdo entre varios observadores; se puede definir acuerdo por *unanimidad* (también llamado de *DeMoivre*) cuando todos los observadores coinciden, o por *mayoría o consenso* cuando una clara mayoría coinciden. Si por ejemplo hay 7 observadores, se define acuerdo por unanimidad si los 7 coinciden en la misma categoría y desacuerdo en cualquier otro caso; acuerdo por mayoría si, por ejemplo, coinciden al menos 5, y desacuerdo en caso contrario. Obviamente, es aconsejable una clara mayoría mejor que mayorías triviales como 4-3.

Landis y Koch [26] proponen, con esta aproximación, definir índices tipo kappa a partir de variables indicadores que representen la definición de acuerdo. Supóngase N observaciones, cada una de ellas realizada por J observadores, que clasifican en K categorías. Los datos se presentarían en una matriz de observaciones Y_{ij}, el subíndice i para las observaciones con los valores de 1 a N y el subíndice j para los observadores tomando valores de 1 a J. La variable Y_{ij} puede tomar valores de 1 a K. Se pueden crear tantas variables z_p como definiciones de acuerdo, del modo siguiente:

$$z_{0i}=\begin{cases}1 & \text{si para la observación i, todos los observadores coinciden}\\ 0 & \text{en otro caso}\end{cases}$$

$$z_{1i}=\begin{cases}1 & \text{si para la observación i, al menos J - 1 observadores coinciden}\\ 0 & \text{en otro caso}\end{cases} \qquad [3.15]$$

... ...

$$z_{pi}=\begin{cases}1 & \text{si para la observación i, al menos J - p observadores coinciden}\\ 0 & \text{en otro caso}\end{cases}$$

La proporción de acuerdos observados, para cada definición de acuerdo, se puede calcular a partir de estas variables, a través de:

$$P_{o(p)} = \frac{\sum_{i=1}^{N} z_{pi}}{N} \qquad [3.16]$$

Denominando $P_j(k)$ $(j=1,...,J\ k=1,...,K)$ a la frecuencia relativa de la categoría k para el observador j, es decir a la proporción de veces que el observador j clasifica en la categoría k, la proporción de acuerdos esperados en la hipótesis de independencia se calcula mediante:

$$P_{e(p)} = \sum_{V \in V_{K,J,p}} P_1(V) \cdots P_J(V) \qquad [3.17]$$

siendo $V_{K,J,p}$ el conjunto de variaciones con repetición de K elementos tomados de J en J, permaneciendo al menos J-p de ellos iguales.

A partir de estos valores de $P_{o(p)}$ y $P_{e(p)}$ se calcula el índice kappa usando [3.2].

Sin embargo, el índice de concordancia para múltiples observadores más usado es el índice kappa propuesto por Davies y Fleiss [27] a partir del acuerdo promedio de todos los posibles pares (*"pairwise agreement"*). Como se dijo antes, para cada observación hay $\binom{J}{2} = \frac{J(J-1)}{2}$ pares posibles de clasificaciones. Denominando X_{ik} al número de observadores que clasifican a la observación i en la categoría k, el número de pares de clasificaciones que están en acuerdo para la observación i es:

$$NA_i = \frac{1}{2}\sum_{k=1}^{K} X_{ik}(X_{ik}-1)$$

En consecuencia, la proporción de acuerdos para dicha observación será:

$$\frac{\sum_{k=1}^{K} X_{ik}(X_{ik}-1)}{J(J-1)} \qquad [3.18]$$

La proporción media de acuerdos observados para todas las observaciones, será la suma de [3.18] para todas las observaciones, dividida por el número de observaciones, es decir:

$$P_o = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{k=1}^{K} X_{ik}(X_{ik}-1)}{J(J-1)} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} X_{ik}(X_{ik}-1)}{NJ(J-1)} = \frac{1}{NJ(J-1)}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} X_{ik}^2 - NJ\right) \qquad [3.19]$$

ya que:

$$\sum_{k=1}^{K} X_{ik} = J \quad \text{y} \quad \sum_{i=1}^{N} J = NJ$$

Nótese que en el caso particular de 2 observadores (J=2) la expresión [3.18] vale 1 si ambos observadores coinciden y 0 en caso contrario, por tanto la expresión [3.19] se reduce a la dada para P_o en [3.5].

Denominando, como antes, $P_j(k)$ $(j=1,...,J\ k=1,...,K)$ a la proporción de veces que el observador j clasifica en la categoría k, la proporción de acuerdos esperados en la hipótesis de independencia entre el par de observadores l y m será:

$$\sum_{k=1}^{K} P_l(k)P_m(k)$$

y, por tanto, la proporción media de acuerdos de pares esperados será su suma para todos los posibles pares dividida por el número de pares posibles, es decir:

$$P_e = \frac{2}{J(J-1)} \sum_{m>l}^{J} \sum_{l=1}^{J} \sum_{k=1}^{K} P_l(k)P_m(k) \qquad [3.20]$$

y a partir de [3.19] y [3.20] se calcula el índice kappa usando [3.2]. Obsérvese que en el caso de 2 observadores la expresión [3.20] también se reduce a la dada en [3.5], por tanto esta formulación para varios observadores del índice κ incluye como caso particular la de 2 observadores. Además este índice puede verse (Dunn [22]) como una media ponderada de los κ de todos los pares posibles de observadores.

En un estudio con múltiples observadores es frecuente, bien por diseño (*diseño incompleto*) o por pérdida de datos, que haya datos que falten, es decir, que no todos los observadores clasifiquen a todas las observaciones. Es conveniente, por tanto, generalizar el índice kappa a diseños incompletos.

El índice kappa basado en el acuerdo de todos los posibles pares, fue generalizado a diseños incompletos por Schouten [28] y Gross [29]: considérese que del conjunto G de J observadores, cada observación i es clasificada por el subconjunto G_i de $J_i \leq J$ observadores, entonces el número de pares posibles de clasificaciones para dicha observación es $\frac{J_i(J_i-1)}{2}$ y por tanto la proporción de acuerdos observados para cada observación es, en lugar de [3.18], la siguiente:

$$\frac{\sum_{k=1}^{K} X_{ik}(X_{ik}-1)}{J_i(J_i-1)}$$

La proporción media de acuerdos observados será su suma dividida por el número de observaciones:

$$P_o = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{\sum_{k=1}^{K} X_{ik}(X_{ik}-1)}{-J_i(J_i-1)} \qquad [3.21]$$

siendo N_c el número de observaciones para las que hay más de un observador. En el caso particular de que todos los G_i sean iguales a G , [3.21] se convierte en la [3.19].

Para la observación i la proporción de acuerdos de pares esperados en la hipótesis de independencia es, análogamente a [3.20]:

$$\frac{2}{J_i(J_i-1)} \sum_{m>l}^{J} \sum_{l=1}^{J} \sum_{k=1}^{K} P_l(k)P_m(k)$$

donde la suma para m y l está restringida al conjunto G_i de observadores que han clasificado a la observación i. La proporción media de pares esperados será su suma para todas las observaciones dividida por el número de observaciones:

$$P_e = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{2}{J_i(J_i-1)} \sum_{m>l}^{J} \sum_{l=1}^{J} \sum_{k=1}^{K} P_l(k)P_m(k) \qquad [3.22]$$

que en el caso particular de que todos los G_i sean iguales a G también se convierte en la [3.20]. A partir de [3.21] y [3.22] se calcula el índice kappa usando [3.2].

4. GENERALIZACIONES DEL ÍNDICE KAPPA

Proponemos generalizaciones del índice kappa en varios sentidos:

4.1 Inclusión de diseños incompletos en el índice kappa basado en las definiciones explícitas de acuerdo propuestas por Landis y Koch [10] de un modo más sencillo al considerado por los autores (Koch, Imrey y Reinfurt [30]) y siguiendo el razonamiento de Schouten [28].

Considérese como antes que, siendo G el conjunto de J observadores, cada observación i es clasificada sólo por un subconjunto G_i de $J_i \leq J$ observadores. Aceptando como definición de acuerdo cualquiera de los representados por las variables [3.15], la proporción de acuerdos observados se calculará del mismo modo que en el diseño completo, es decir con [3.16], pero usando sólo aquellas observaciones en las que puede observarse el acuerdo definido, es decir en las que hay, al menos, J-p observadores. Sin embargo, ahora, la proporción de acuerdos esperados en la hipótesis de independencia es distinta para cada observación, puesto que cada una de ellas es clasificada por un subconjunto distinto de observadores. Para cada observación, del mismo modo que en [3.17], tal proporción es:

$$\sum_{V \in V_{K, J_i, p}} P_1(V) \cdots P_{J_i}(V)$$

donde, ahora, el conjunto de variaciones es sólo sobre los J_i observadores que clasifican a la observación i. La proporción media de acuerdos esperados será su suma para todas las observaciones, dividida por el número de observadores para los que puede observarse el acuerdo definido:

$$P_{e(p)} = \frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{V \in V_{K, J_i, p}} P_1(V) \cdots P_{J_i}(V) \qquad [4.1]$$

siendo N_c el número de observaciones para los que hay, al menos, J-p observadores. A partir de aquí se calculará el índice kappa como siempre.

4.2 Inclusión de pesos en el caso de múltiples observadores para variables ordinales y acuerdo por pares que ponderen la importancia de los distintos desacuerdos. Como las fórmulas [3.19] y [3.20] son un caso particular de las más generales [3.21] y [3.22] se van a usar estas últimas para la incorporación de los pesos.

Sea w_{lm} el peso correspondiente al acuerdo-desacuerdo entre las categorías l y m con las mismas condiciones que en el capítulo 3, es decir:

$$w_{mm} = 1; \qquad 0 \leq w_{lm} < 1 \quad \forall l \neq m; \qquad w_{lm} = w_{ml}$$

El número de acuerdos ponderados por estos pesos para la observación i es:

$$NA_i = \frac{1}{2} \sum_{k=1}^{K} w_{kk} X_{ik} (X_{ik} - 1) + \sum_{m>l}^{K} \sum_{l=1}^{K} w_{lm} X_{il} X_{im}$$

consecuentemente, la proporción de acuerdos ponderados para dicha observación es:

$$\frac{\sum_{k=1}^{K} w_{kk} X_{ik} (X_{ik} - 1) + 2 \sum_{m>l}^{K} \sum_{l=1}^{K} w_{lm} X_{il} X_{im}}{J_i (J_i - 1)} = \frac{\sum_{m=1}^{K} \sum_{l=1}^{K} w_{lm} X_{il} X_{im} - J_i}{J_i (J_i - 1)}$$

ya que:

$$\sum_{k=1}^{K} w_{kk} X_{ik} = \sum_{k=1}^{K} X_{ik} = J_i$$

y la proporción media de acuerdos ponderados para todas las observaciones es:

$$P_{o(w)} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{\sum_{m=1}^{K} \sum_{l=1}^{K} w_{lm} X_{il} X_{im} - J_i}{J_i (J_i - 1)} \qquad [4.2]$$

Para la observación i la proporción de acuerdos ponderados esperados bajo la hipótesis de independencia entre el par de observadores l y m será:

$$\sum_{u=1}^{K} \sum_{k=1}^{K} w_{uk} P_l(u) P_m(k)$$

y la proporción media de acuerdos ponderados esperados para todos las observaciones:

$$P_{e(w)} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{2}{J_i (J_i - 1)} \sum_{m>l}^{J} \sum_{l=1}^{J} \sum_{u=1}^{K} \sum_{k=1}^{K} w_{uk} P_l(u) P_m(k) \qquad [4.3]$$

donde, como en [3.22], las sumas para m y l están restringidas en cada observación i al conjunto G_i de observadores que la han clasificado. Esta formulación, que incluye pesos, clasificación multinomial, y múltiples observadores con diseño incompleto o datos perdidos, es la más general del índice κ e incluye las anteriores como casos

particulares.

4.3 Otras definiciones de acuerdo: Además de las definiciones de acuerdo por mayoría o para todos los posibles pares, últimamente en los estudios sobre uso apropiado de procedimientos médicos, se emplean definiciones más elaboradas de acuerdo y desacuerdo entre observadores que, además, no son complementarias entre sí. En estos estudios (Lázaro y Kitch [31]), un número impar, generalmente 9, de expertos examinan una serie de situaciones clínicas en las que un procedimiento médico, p.e. angioplastia coronaria, puede estar indicado. Cada uno de ellos puntúa lo apropiado de la indicación desde 1 (totalmente inapropiado) a 9 (totalmente apropiado) para, a partir de ciertas definiciones de acuerdo entre los observadores, generar guías de actuación ("*guidelines*") para dicho procedimiento. Las definiciones de acuerdo y desacuerdo propuestas por Brook et al. [32] son:

Para el acuerdo:

AE (estadístico): Con un nivel de seguridad del 33%, no se puede rechazar la hipótesis de que el 80% de la población hipotética de puntuaciones se encuentra dentro del mismo intervalo 1-3, 4-6 ó 7-9 que la mediana obtenida. Para el caso de 9 observadores, esta condición se cumple cuando menos de 3 observadores puntúan fuera del intervalo que contiene la mediana.

A9S (estricto): Las 9 puntuaciones caen dentro de uno de los siguientes intervalos: 1-3, 4-6, 7-9.

A9R (relajado): Las 9 puntuaciones caen dentro de cualquier intervalo de 3 puntos.

A7S (estricto): Tras descartar la puntuación más alta y más baja, las 7 puntuaciones restantes caen dentro de uno de los siguientes intervalos: 1-3, 4-6, 7-9.

A7R (relajado): Tras descartar la puntuación más alta y más baja, las 7 puntuaciones restantes caen dentro de un intervalo cualquiera de 3 puntos.

Para cada definición de acuerdo, se usan las siguientes definiciones de desacuerdo, que, evidentemente, no son complementarias de su acuerdo correspondiente:

DE (estadístico): Con un nivel de seguridad del 10% se puede rechazar la hipótesis de que el 90% de la población hipotética de puntuaciones se encuentra en uno de los dos intervalos 1-6 ó 4-9. Para el caso de 9 observadores esta definición se cumple cuando hay tres o más puntuaciones en el intervalo 1-3 y tres o más en el intervalo 7-9.

D9S (estricto): De las 9 puntuaciones, hay por lo menos una puntuación de 1 y otra de 9.

D9R (relajado): De las 9 puntuaciones, hay por lo menos una en el intervalo 1-3 y otra en el intervalo 7-9.

D7S (estricto): Tras descartar la puntuación más alta y más baja, de las 7 restantes hay por lo menos una puntuación de 1 y otra de 9.

D7R (relajado): Tras descartar la puntuación más alta y más baja, de las 7 restantes hay por lo menos una puntuación en el intervalo 1-3 y otra en el intervalo 7-9.

Para analizar la concordancia entre observadores, que puede ser útil para seleccionar el panel de expertos, proponemos usar, para cada definición de acuerdo, índices tipo κ similares a [3.2]:

$$\kappa = \frac{P_{o(a)} - P_{e(a)}}{1 - P_{e(a)}} \qquad [4.4]$$

Las proporciones observadas de acuerdos $P_{o(a)}$ se pueden calcular, para cada definición de acuerdo, a partir de variables indicadoras similares a [3.15]:

$$z_{AEi} = \begin{cases} 1 & \text{si para la observación } i, \text{ se cumple AE} \\ 0 & \text{si no se cumple} \end{cases}$$

$$z_{A9Si} = \begin{cases} 1 & \text{si para la observación i, se cumple A9S} \\ 0 & \text{si no se cumple} \end{cases} \qquad [4.5]$$

... ...

$$z_{A7Ri} = \begin{cases} 1 & \text{si para la observación i, se cumple A7R} \\ 0 & \text{si no se cumple} \end{cases}$$

Por ejemplo, para AE y A9S:

$$P_{o(AE)} = \frac{\sum_{i=1}^{N} z_{AEi}}{N} \qquad P_{o(A9S)} = \frac{\sum_{i=1}^{N} z_{A9Si}}{N} \qquad [4.6]$$

siendo N el número de situaciones clínicas consideradas.

Denominando, como siempre $P_p(k)$ $(p=1,...,J\ k=1,...,K)$ a la frecuencia relativa de la categoría k para el observador p, la proporción de acuerdos AE esperados bajo la hipótesis de independencia se calculará mediante:

$$P_{e(AE)} = \sum_{V \in V_{K,J,AE}} P_1(V) \cdots P_J(V) \qquad [4.7]$$

donde $V_{K,J,AE}$ representa al conjunto de variaciones con repetición de K elementos tomados de J en J, que cumplen la definición AE.

De modo similar se calculan la proporción esperada de A9S, A9R, A7S y A7R; y evidentemente, de un modo similar se pueden definir índices κ para otro número de observadores, u otro sistema de puntuaciones, u otras definiciones de acuerdo.

5. DISTRIBUCIÓN MUESTRAL DEL ÍNDICE KAPPA

Todos los índices κ vistos hasta ahora se calculan a partir de muestras, que producen estimaciones del verdadero valor de κ en la población. Es necesario estudiar su distribución muestral para poder construir intervalos de confianza y realizar contrastes de hipótesis.

En el caso de dos observadores clasificando en K categorías se puede asumir el denominado *modelo de Kullback* (Hubert [33]) según el cuál un observador clasifica siguiendo una distribución multinomial de parámetros $\pi_1, \pi_2, \ldots, \pi_k$ y el otro con una multinomial independiente de parámetros $\pi'_1, \pi'_2, \ldots, \pi'_k$ donde $\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \pi'_k = 1$. Evidentemente $P_{.i} = \frac{X_{.i}}{N}$ es un estimador de π_i y $P_{i.} = \frac{X_{i.}}{N}$ lo es de π'_i. Con esta notación la expresión [3.5] de P_e se puede escribir como $P_e = \sum_{k=1}^{K} P_{k.} P_{.k}$. Con este modelo, el número de acuerdos observados NA es una variable binomial de parámetros N y $\sum_{k=1}^{K} \pi_k \pi'_k$, en consecuencia su media y varianza son, respectivamente:

$$\mu_{NA} = N \sum_{k=1}^{K} \pi_k \pi'_k \qquad \sigma^2_{NA} = N \left(\sum_{k=1}^{K} \pi_k \pi'_k \right) \left(1 - \sum_{k=1}^{K} \pi_k \pi'_k \right) \qquad [5.1]$$

Si en [5.1] se sustituyen π_k y π'_k por sus estimadores, se obtienen los estimadores de μ_{NA} y σ^2_{NA} y, usando resultados asintóticos de la teoría de grandes muestras, se puede obtener (Fleiss, Cohen y Everitt [34]):

$$\mu_\kappa = 0 \qquad \hat{\sigma}^2_\kappa = \frac{P_e + P_e^2 - \sum_{k=1}^{K} P_{k.}P_{.k}(P_{k.} + P_{.k})}{N(1 - P_e)^2} \qquad [5.2]$$

Si se cumplen las condiciones de aproximación de la binomial a la normal, el estadístico $z = \frac{\hat{\kappa}}{\hat{\sigma}_\kappa}$ se distribuye como una normal tipificada y puede usarse para contrastar la H_0: $\kappa = 0$ con una región crítica para un contraste lateral $z > z_\alpha$.

En general, sin embargo, estos contrastes no tienen mucho interés. El objetivo de un estudio de concordancia no es tanto contrastar si hay más acuerdo que el esperado en la hipótesis de independencia, sino cuantificar el mismo. Si en un estudio se encuentra $\kappa = 0{,}1$, aunque sea significativamente distinto de 0, ello revela un acuerdo insignificante. Lo que, en general, tiene interés es la estimación por intervalos. Aquí debe observarse que la varianza dada por [5.2] se ha obtenido con el modelo de Kullback de independencia entre observadores al que le corresponde un $\kappa = 0$ y consecuentemente no sirve para construir intervalos de confianza en la hipótesis de no independencia. En esta hipótesis se asume que la tabla $K \times K$ es generada por una distribución multinomial de parámetros π_{ij}, $i=1,...,K$ $j=1,...,K$. Evidentemente $\sum_{i=1}^{K}\sum_{j=1}^{K} \pi_{ij} = 1$. El modelo de Kullback es un caso particular de éste, en el que $\pi_{ij} = \pi_{i.}\pi_{.j}$. El índice kappa se puede definir del siguiente modo:

$$\kappa = \frac{\sum_{k=1}^{K} \pi_{kk} - \sum_{k=1}^{K} \pi_{k.}\pi_{.k}}{1 - \sum_{k=1}^{K} \pi_{k.}\pi_{.k}} \qquad [5.3]$$

del que la expresión [3.2] es su estimador muestral ($\hat{\kappa}$).

Fleiss, Cohen y Everitt [34] demuestran que su varianza asintótica está dada por la expresión:

$$\hat{\sigma}^2_{\hat{\kappa}} = \frac{A + B - C}{N(1 - P_e)^4} \qquad [5.4]$$

donde:

$$A = \sum_{k=1}^{K} P_{kk}\Big((1 - P_e) - (P_{.k} + P_{k.})(1 - P_o)\Big)^2$$

$$B = (1 - P_o)^2 \sum_{j \neq k=1}^{K} \sum_{j=1}^{K} P_{jk}(P_{.j} + P_{k.})^2$$

$$C = (P_o P_e - 2P_e + P_o)^2$$

De [3.2] se deduce:

$$1 - \hat{\kappa} = \frac{1 - P_o}{1 - P_e}$$

Dividiendo A, B y C por $(1 - P_e)^2$ y teniendo en cuenta la expresión anterior, se obtiene:

$$A' = \frac{A}{(1 - P_e)^2} = \sum_{k=1}^{K} P_{kk}\Big(1 - (P_{.k} + P_{k.})(1 - \hat{\kappa})\Big)^2$$

$$B' = \frac{B}{(1 - P_e)^2} = (1 - \hat{\kappa})^2 \sum_{j \neq k=1}^{K} \sum_{j=1}^{K} P_{jk}(P_{j.} + P_{k.})^2$$

$$C' = \frac{C}{(1 - P_e)^2} = \left(\hat{\kappa} - P_e(1 - \hat{\kappa})\right)^2$$

que da lugar a una expresión equivalente a [5.4] más fácil de calcular:

$$\hat{\sigma}^2_{\hat{\kappa}} = \frac{A' + B' - C'}{N(1 - P_e)^2} \qquad [5.5]$$

A partir de esta varianza y usando la normalidad también asintótica de la distribución de $\hat{\kappa}$, un intervalo de confianza aproximado, con un nivel de confianza de $100(1 - \alpha)\%$ es:

$$\hat{\kappa} - z_{\alpha/2}\hat{\sigma}_{\hat{\kappa}} \leq \kappa \leq \hat{\kappa} + z_{\alpha/2}\hat{\sigma}_{\hat{\kappa}} \qquad [5.6]$$

Fleiss, Cohen y Everitt [34] estudiaron también la distribución muestral del $\hat{\kappa}$ ponderado en las hipótesis de independencia y no independencia de los observadores, y encontraron que su distribución es también asintóticamente normal con varianzas, en la hipótesis de independencia:

$$\hat{\sigma}^2_{\hat{\kappa}_w} = \frac{\sum_{j=1}^{K} \sum_{k=1}^{K} P_{j.} P_{.k}\left(w_{jk} - (\overline{w}_{j.} + \overline{w}_{.k})\right)^2 - P^2_{e(w)}}{N(1 - P_{e(w)})} \qquad [5.7]$$

y

$$\hat{\sigma}^2_{\hat{\kappa}_w} = \frac{\sum_{j=1}^{K}\sum_{k=1}^{K} P_{jk}\left(w_{jk} - (\overline{w}_{j.} + \overline{w}_{.k})(1 - \hat{\kappa}_w)\right)^2 - \left(\hat{\kappa}_w - P_{e(w)}(1 - \hat{\kappa}_w)\right)^2}{N(1 - P_{e(w)})} \qquad [5.8]$$

en la hipótesis de no independencia entre observadores.

En ambas fórmulas:

$$\overline{w}_{j.} = \sum_{i=1}^{K} P_{.i} w_{ji} \qquad \overline{w}_{.k} = \sum_{i=1}^{K} P_{i.} w_{ik}$$

Es fácil ver que las fórmulas [5.2] y [5.5] son una caso particular de las [5.7] y [5.8], respectivamente, si w_{ii}=1 para todo i y w_{ij}=0 para todo $i \neq j$.

Para el caso de múltiples observadores, aunque hay algunas aproximaciones parciales (Fleiss [11]), no existe todavía una fórmula de uso general para la estimación de $\sigma^2_{\hat{\kappa}}$. Una técnica muy general para construir intervalos de confianza a partir de estadísticos de distribución muestral desconocida es la denominada *"técnica jackknife"*. Esta técnica fue introducida por Quenouille [35] como método para reducir el sesgo en la estimación de parámetros y más tarde Tukey [36] propuso su uso para obtener un estadístico T que permita contrastar hipótesis y construir intervalos de confianza. Distintos autores (p.e. Kraemer [37], Fleiss y Davies [38], Schouten [28]) la han aplicado a diversas extensiones del índice kappa. Una buena descripción, así como su relación con técnicas similares puede verse en Efron y Gong [39].

La técnica *jackknife* consiste en lo siguiente: dada una muestra de N observaciones, se construyen N submuestras, cada una de ellas de $N-1$ elementos, por eliminación sucesiva de cada una de las observaciones[5]. Para cada submuestra se calcula el estadístico de interés, en este caso el índice kappa. Sea $\hat{\kappa}$ el estimador del índice kappa a partir de la muestra y $\hat{\kappa}_i$ el estimador a partir de la submuestra i, es decir de la submuestra de la que se ha eliminado la observación i. Se denomina *pseudovalor* de esa submuestra a:

$$J_i(\kappa) = N\hat{\kappa} - (N-1)\hat{\kappa}_i \qquad i = 1,2,...,N$$

El *estimador jackknife* de κ es la media aritmética de los pseudovalores:

$$J(\kappa) = \frac{\sum_{i=1}^{N} J_i(\kappa)}{N} \qquad [5.9]$$

En una gran variedad de situaciones se ha demostrado (Arvesen [40] y Arvesen y Schmitz [41]), y en particular (Parr y Tolley [42]) para toda función real de proporciones multinomiales (tal como los distintos índices κ) con primera y segunda derivada parciales continuas, que, para grandes muestras, el estimador jackknife tiene una distribución aproximadamente normal y que un estimador de su varianza es la varianza de los pseudovalores. Por lo tanto, el estadístico:

[5] Aunque en la propuesta inicial de Quenouille [35] y Tukey [36] se divide la muestra en n submuestras de k elementos cada una ($N=kn$), en la actualidad (Efron y Gong [39], Parr y Tolley [42]) es más usada la formulación presentada aquí.

$$T = \frac{J(\kappa) - \kappa}{\frac{S_J}{\sqrt{N}}} \qquad [5.10]$$

se distribuye aproximadamente como una t con $N-1$ grados de libertad. En [5.9], S_J es la desviación estándar de los pseudovalores, es decir:

$$S_J^2 = \frac{\sum_{i=1}^{N} \left(J_i(\kappa) - J(\kappa)\right)^2}{N - 1}$$

Por lo tanto, el error estándar estimado por esta técnica para el índice kappa está dado por:

$$\hat{EE}(\hat{\kappa}) = \frac{S_j}{\sqrt{N}} \qquad [5.11]$$

y entonces un intervalo aproximado, con un nivel de confianza de $100(1 - \alpha)\%$, es:

$$J(\kappa) - t_{\alpha/2,(N-1)} \frac{S_j}{\sqrt{N}} \leq \kappa \leq J(\kappa) + t_{\alpha/2,(N-1)} \frac{S_j}{\sqrt{N}} \qquad [5.12]$$

Se empleará este procedimiento para construir los intervalos de confianza de las distintas generalizaciones del índice kappa propuestas en el capítulo 4.

6. DESCRIPCIÓN DE LOS PROGRAMAS

Para implementar los cálculos de los índices descritos se escribió un nuevo programa, en FORTRAN 77, y se modificó otro de los existentes en el paquete estadístico PRESTA [3]. Los listados de ambos, así como los de las subrutinas específicas correspondientes figuran en el **Apéndice II**.

Se modificó el programa de *Análisis de Tablas de Contingencia (CONTIN)*, que calculaba los estadísticos para el análisis de independencia de dos caracteres cualitativos (pruebas ji-cuadrado y exacta de Fisher) y de la validez de una clasificación binomial (cálculo de la sensibilidad y especificidad con sus intervalos de confianza al 95%), para incluir el cálculo del índice kappa (fórmula [3.2] con [3.5]) en el caso de dos observadores. El programa calcula también los errores estándar en los supuestos de independencia (fórmula [5.2]) y de no independencia (fórmula [5.5]) así como los estadísticos de McNemar [43], en el caso de clasificaciones binomiales, y el de Stuart-Maxwell [44], en el de clasificaciones multinomiales, para contrastar la homogeneidad de las frecuencias marginales que, como se comentó en el **Capítulo 3**, facilita la interpretación del índice kappa.

Se escribió el programa de *Cálculo del índice kappa para múltiples observadores (KAPMUL)*, que incluye diseños incompletos, y acuerdo por pares (fórmulas [4.2] y [4.3]), por mayoría (fórmulas [3.16] modificada según apartado 4.1 y [4.1]) y "acuerdos RAND" (fórmulas [4.6] y [4.7]). En el caso de acuerdo por pares, acepta "pesos" para los distintos desacuerdos, ofreciendo por defecto los pesos bicuadrados (fórmula [3.9]). En todos los casos estima el error estándar por la técnica *jackknife* (fórmula [5.11]) y calcula tanto el índice kappa como su estimador *jackknife* (fórmula [5.9]) y

el intervalo de confianza al 95% construido en base a dicho estimador (fórmula [5.12]). La proximidad entre el índice kappa y su estimador *jackknife* puede usarse como indicador del grado de aproximación de esta técnica.

Para contrastar la homogeneidad de las frecuencias marginales; y dado que la generalización del estadístico de McNemar para múltiples observadores, debida a Cochran [45], cuya validez en el caso de medidas sobre los mismos sujetos fue establecida por Fleiss [46], sólo es aplicable a clasificaciones binomiales; se implementó el método desarrollado por Landis y Koch [10] y Koch et al. [47] usando ajustes a modelos lineales GSK (Grizzle, Starmer y Koch [48]), aunque requiere diseños completos. Existe una extensión del método para diseños incompletos, debida a Koch, Imrey y Reinfurt [30], pero se basa en la hipótesis de independencia entre el mecanismo que determina los datos faltantes y las variables del problema, que parece, en general, demasiado restrictiva como para implementarla en un programa de uso general.

El método consume mucha memoria, ya que trabaja con matrices de orden $((J-1)\times(K-1))\times(J\times(K-1))$, $(J\times(K-1))\times((J-1)\times(K-1))$, $(J\times(K-1))\times(J\times(K-1))$, siendo J el número de observadores y K el de categorías. El programa que se presenta en el **Apéndice II** está ajustado para los 640 Kb que permite manejar el DOS y calcula el estadístico GSK sólo si $K^J < 82$. Para ampliar ese límite, si la memoria disponible lo permitiera, hay que modificar las dimensiones de los *array* incluidos en la sentencia DOUBLE PRECISION y el IF de control inmediatamente anterior al bucle de la línea 7000.

Debido a la gran cantidad de tiempo de cálculo requerido; sobre todo para

generar las permutaciones en el cálculo del acuerdo esperado por mayoría (p.e. la generación de las permutaciones para el acuerdos A9S y AE necesitan aproximadamente 40 y 65 horas, respectivamente, con un procesador 486 a 70 Mhz) y los estimadores *jackknife* con diseño incompleto; el programa avisa del tiempo de ejecución restante estimado, si éste es superior a 0,5 minutos, cuando calcula los acuerdos esperados y el estimador *jackknife*. Además, y en orden a optimizar este tiempo, la primera vez que se ejecuta con un acuerdo RAND guarda en un archivo las permutaciones válidas para ese acuerdo. La siguiente vez que se ejecuta, si el archivo existe, las lee del mismo, reduciendo el tiempo necesario (p.e. a 73 minutos en el mismo ordenador para el acuerdo A9S, aunque sigue necesitando 108 horas para el acuerdo AE, para los datos del estudio sobre uso apropiado de la angioplastia coronaria presentado en el próximo capítulo).

7. ESTUDIOS OBSERVACIONALES

Se presentan tres estudios de precisión de clasificaciones clínicas, usando las generalizaciones del índice κ propuestas. Sus objetivos y diseño son:

7.1 Primer estudio:

Se trata de un estudio de factibilidad previo a otro estudio para determinar el estado actual de salud de la cohorte de afectados por el envenenamiento producido por el consumo de aceite de colza desnaturalizado conocido como Síndrome del Aceite Tóxico (SAT). Una descripción inicial del SAT puede verse en Grandjean y Tarkowski [49] y una revisión más actual en Nadal y Tarkowski [50]. El número de personas afectadas es aproximadamente 20.000. El estudio para determinar el estado actual de salud (actualmente en fase de preparación del manuscrito), se realizó en los 4.015 afectados pertenecientes a las 7 Unidades de Seguimiento del SAT de la provincia de Madrid; se usaron como fuentes de datos la historia clínica y la exploración directa de los pacientes realizada por los médicos de las Unidades respectivas. Como las exploraciones fueron llevadas a cabo por diferentes médicos, se planteó un estudio previo de factibilidad (el que aquí se presenta) que incluyó la evaluación de la precisión y la validez de dichas exploraciones para las variables más potencialmente influidas por la subjetividad de los observadores. Además de la presión arterial, que al ser una variable continua no se presenta, las variables del estudio fueron: neuropatía periférica, clasificada en tres niveles: ausente, dudosa y clara, codificadas respectivamente como 1, 2 y 3; severidad de los cambios esclerodermiformes de la piel, clasificada en cuatro niveles: no esclerodermia, esclerodermia leve, esclerodermia moderada y piel atrófica, codificadas de 1 a 4; y contracturas articulares clasificada como sí y no y codificada como 1 y 2 respectivamente.

Diseño:

Pacientes: Una muestra de conveniencia de 10 pacientes afectados por el SAT y elegidos para cubrir todo el rango de grados de manifestaciones clínicas de la enfermedad.

Observadores: Una muestra aleatoria de 6 médicos elegidos entre el conjunto de los 9 que formaron parte del estudio final sobre el estado de salud.

Procedimiento: Los seis médicos elegidos participaron, antes del estudio, en un seminario de 5 horas realizado en el Departamento de Reumatología del Hospital 12 de Octubre de Madrid, en el que médicos especialistas del Hospital les instruyeron en el protocolo de recogida de las variables. El seminario incluyó el examen clínico de varios pacientes, distintos a los que luego participaron en el estudio de factibilidad, afectados por el SAT.

El estudio se realizó en la Unidad de Seguimiento del SAT de Vallecas. Para evitar que un mismo paciente tuviera que ser visto demasiadas veces por un mismo signo en un breve intervalo de tiempo, lo que le provocaría molestias y cansancio, que podría además afectar al resultado de la exploración, se optó por un diseño de bloques incompletos balanceados (Fleiss [51]) en el que cada paciente es examinado por 3 médicos, cada médico examina a 5 pacientes, y cada uno de todos los posibles pares de médicos examina a 2 pacientes. En la Tabla 11 se muestra el esquema de los exámenes realizados. Con este diseño, el factor de eficiencia (Fleiss [51]) para la estimación del coeficiente de repetibilidad es 0,8 que aparece como un compromiso razonable. Este estudio, con diseño incompleto e incluyendo variables multinomiales ordinales, fue el origen de esta tesis al plantear la necesidad de la generalización del índice κ expuesta en el apartado 4.2.

Tabla 11

Diseño de bloques incompletos balanceados usado en el Estudio del SAT

Paciente	Médico 1	Médico 2	Médico 3	Médico 4	Médico 5	Médico 6
1	x			x		x
2			x	x	x	
3			x	x		x
4	x		x		x	
5	x				x	x
6	x	x	x			
7		x	x			x
8		x			x	x
9	x	x		x		
10		x		x	x	

Además, cada paciente fue examinado conjuntamente por otros dos médicos (un investigador "senior" del SAT y un especialista del Hospital 12 de Octubre) que establecieron por consenso el estado de cada variable de estudio. Esta medición se usó como patrón de oro para evaluar la validez de las mediciones; aunque como el estudio de la validez no es objeto de esta tesis, los datos correspondientes no se presentan. La secuencia de los cuatro exámenes en cada paciente se fijó aleatoriamente usando una tabla de permutaciones. Los datos registrados para las tres

variables de interés figuran en el **Apéndice III.**

Por imperativos éticos y legales, los pacientes fueron informados por escrito de que se trataba de un estudio, así como de sus objetivos; de que serían informados de los resultados; de que su participación era voluntaria y que se podían retirar en cualquier momento, sin tener que dar explicaciones y sin que ello significara ninguna renuncia a su derecho a la asistencia; y se les solicitó su consentimiento también por escrito y firmado. Para garantizar la privacidad de los datos, no se introdujo en los archivos informáticos del estudio ningún dato que permitiera identificar a los pacientes (nombre, dirección, etc.), sino sólo un número de identificación asignado en el momento de introducir los datos en el ordenador. El listado con los nombres de los pacientes y el número asignado es custodiado por el investigador responsable del conjunto del proyecto.

7.2 Segundo estudio:

El *nevus flammeus* (NF) o *mancha en vino de Oporto* es una malformación caracterizada por la presencia de capilares ectásicos en dermis, que se manifiesta clínicamente en forma de manchas de color rojizo, presentes casi siempre desde el nacimiento. Su incidencia se estima entre el 0,3 y el 0,5% de los recién nacidos, si bien la incidencia real podría ser mayor ya que en ocasiones el eritema fisiológico del neonato enmascara la lesión en los primeros días. El tratamiento de elección en la actualidad para estas lesiones es la laserterapia (Enjolras y Mulliken [52], Geronemus [53]), en particular el láser de colorante pulsado (LCP) con una duración de pulso de 360 ms y una longitud de onda de 585 nm. El resultado es un daño específico a los capilares dérmicos, sin cicatrices y, generalmente, sin alteraciones pigmentarias. La valoración de la respuesta del NF al tratamiento se basa en el cambio cromático del angioma;

aunque se han intentado otros procedimientos más objetivos como la espectrofotometría o la microscopía transcutánea; cuantificado de modo subjetivo en porcentaje de aclaramiento respecto al color inicial. Entre los factores pronósticos de esta respuesta se ha descrito el color de la lesión (Noe et al. [54]) también apreciado generalmente de manera subjetiva.

Este estudio (Pérez, Abraira et al. [55]) pretende evaluar la concordancia entre las clasificaciones, realizadas por distintos dermatólogos, tanto del color inicial de la lesión como del aclaramiento producido por el tratamiento.

Diseño:

Observaciones: Una muestra aleatoria de 80 historias clínicas elegidas entre las 1.500 correspondientes a los pacientes de NF tratados por LCP en el Servicio de Dermatología del Hospital Ramón y Cajal de Madrid entre 1989 y 1995. A falta de otro procedimiento para estimar el tamaño muestral, se usó, sobreestimándolo, el conseguido por Cicchetti [56], por el método de simulación de Monte Carlo, para dos observadores ($N=2K^2$, siendo K el número de categorías).

Observadores: Seis médicos del Servicio de Dermatología; de ellos 3 son los que habitualmente tratan a estos pacientes y los otros 3 fueron voluntarios con experiencia en el diagnóstico del NF.

Procedimiento: El tratamiento se realiza en sesiones sucesivas y finaliza cuando: i) el paciente lo decide, bien porque se encuentra satisfecho con el aclaramiento conseguido o porque considera que no merece la pena seguir; ii) el médico considera que no se va a conseguir más aclaramiento. En la visita inicial antes de comenzar el tratamiento, y después de cada sesión, se toman fotografías de la lesión, todas con la misma cámara y los mismos parámetros (Nixon F-601 con objetivo AF micrinikon

60 m 1/2.8, flash anular y película Ektacrome Kodak 100 ASA) como ayuda en esta decisión. De las historias clínicas seleccionadas, se tomaron 2 fotografías: la correspondiente a la visita inicial y otra post-sesión elegida al azar de entre todas las sesiones realizadas para ese paciente, para cubrir el espectro más amplio de aclaramientos posibles. Las fotos iniciales fueron vistas y clasificadas, de modo ciego e independiente por los 6 dermatólogos, según su color en Rosa, Rojo y Violeta, codificados como 1, 2 y 3 respectivamente. Los 3 dermatólogos implicados en el tratamiento de los pacientes, también de modo ciego e independiente, compararon la fotos iniciales y la post-sesión de cada paciente y cuantificaron el aclaramiento en una escala de 0 (ningún aclaramiento) a 100 (aclaramiento total). Posteriormente, esta escala fue categorizada en los 4 niveles usados habitualmente en este tipo de estudios, codificados de 1 a 4: 0 a 24, ausencia de aclaramiento; 25 a 49, aclaramiento pobre; 50 a 74, bueno; y 75 a 100 excelente. Los datos recogidos figuran en el **Apéndice IV.**

Como el estudio no se realiza sobre pacientes, sino sobre fotografías recogidas en la historia clínica, realizadas como parte del tratamiento, no se consideró que planteara problemas éticos y, por ello, no se solicitó el consentimiento informado.

7.3 Tercer estudio:

Se enmarca en el Proyecto sobre "Uso apropiado de la angioplastia coronaria transluminal percutánea (ACTP) en España" [56], en el que, siguiendo el método propuesto por Brook et al.. [32], se pretende evaluar el uso apropiado de esta técnica de cardiología intervencionista en España. El método, resumiendo, se estructura en tres fases:

Fase 1: Un grupo de médicos, especialmente entrenado, hace una revisión crítica de

la literatura científica sobre ACTP producida hasta la fecha de inicio del proyecto. Basándose en esta revisión se realiza una síntesis de la evidencia sobre utilización, eficacia, efectividad, riesgos y costes del procedimiento. Paralelamente, dos cardiólogos y un experto en el método, elaboran una lista de circunstancias clínicas específicas que un paciente puede presentar para valorar la posible realización de la ACTP. Esta lista, al constituir una clasificación, debe ser exhaustiva (todo paciente debe poder ser clasificado en una circunstancia clínica) y excluyente (cada paciente sólo puede ser clasificado en una). La lista construida en este caso consta de 467 circunstancias, consideradas en dos posibles situaciones: "Riesgo quirúrgico BAJO/MODERADO" y "Riesgo quirúrgico ALTO". A modo de ejemplo, en el **Apéndice V** se muestran las 18 primeras.

Fase II: Un panel de 9 expertos; en este caso tres cardiólogos intervencionistas (que practican la ACTP), tres cardiólogos no intervencionistas (que no la practican), y tres cirujanos cardíacos; usando la síntesis de la evidencia obtenida en la fase anterior, puntúan de 1 (totalmente inapropiado) a 9 (totalmente apropiado) el grado en que, para cada circunstancia, el procedimiento es apropiado. En este proyecto, y como modificación del método original, a los panelistas se les pide comparar el procedimiento con su alternativo, primero tratamiento médico (1) contra revascularización (9) y después las dos técnicas de revascularización: ACTP (1) contra la cirugía aórtico coronaria (CAC) (9) para las 445 primeras "circunstancias clínicas", mientras que para las 22 restantes sólo tratamiento médico contra CAC. Si considera que ninguna técnica es apropiada debe responder 5. Como el resultado final es el del panel en su conjunto, a cada panelista se le pide que compare las dos técnicas de revascularización, incluso aunque considere que lo apropiado es el tratamiento médico. En una primera ronda los expertos puntúan de modo independiente y en una segunda, y después de conocer los resultados de la primera, pueden modificar

su puntuación.

Fase III: En una muestra aleatoria de historias clínicas de pacientes que hayan recibido ACTP, se valora lo apropiado de su uso, de acuerdo con las definiciones obtenidas en la fase anterior.

El estudio se realizó entre las dos rondas de la Fase II, para evaluar el acuerdo entre los expertos, actuándose sobre las 445 circunstancias en que se comparaba el tratamiento médico contra la revascularización en las dos situaciones de riesgo quirúrgico. Se usaron como definiciones de acuerdo la más estrica (A9S) y la más amplia (AE).

8. RESULTADOS

Los resultados para los tres estudios de precisión presentados son:

8.1 Resultados del estudio del SAT

Para todas las variables de este estudio, como el número de observadores es 6, $K^J \geq 82$ *y por lo tanto, no se pudo calcular el estadístico GSK para contrastar la* homogeneidad de las frecuencias marginales. Como consecuencia de ello, se juzgará su homogeneidad a partir de las frecuencias estimadas.

8.1.1 Contracturas articulares

```
P R E S T A  PC   V2.2                              12-JUL-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

 NOMBRE DE LOS DATOS: contrac

 CONDICIONES DEL DISEÑO
 Nº CASOS:                10
 Nº CASOS ELIMINADOS:      0
 Nº VARIABLES:             6    DISEÑO INCOMPLETO
 Nº CATEGORIAS:            2


          FRECUENCIAS MARGINALES
                CATEGORIA
 CONDICION       1       2
 MED2        1.000    .000
 MED4         .400    .600
 MED5         .400    .600
 MED6         .400    .600
 MED7         .600    .400
 MED8         .400    .600

 PROMEDIO     .533    .467
```

```
ACUERDO POR PARES

PROPORCION DE ACUERDO OBSERVADO:    .6667
PROPORCION DE ACUERDO ESPERADO:     .4827

INDICE KAPPA:                       .3557
ESTIMADOR "JACKKNIFE":              .3827   ERROR ESTANDAR:     .2267
INTERVALO DE CONFIANZA AL 95%:     -.1257   .8910


ACUERDO POR MAYORIA DE   3 OBSERVADORES

PROPORCION DE ACUERDO OBSERVADO:    .5000
PROPORCION DE ACUERDO ESPERADO:     .2240

INDICE KAPPA:                       .3557
ESTIMADOR "JACKKNIFE":              .3827   ERROR ESTANDAR:     .2267
INTERVALO DE CONFIANZA AL 95%:     -.1257   .8910
```

En este caso, en que sólo 3 observadores clasifican a cada observación, no sería aconsejable definir una mayoría de sólo 2 observadores; en apoyo de esta opinión, si se definiera, se encontraría una proporción de acuerdo observado de 1 (con la que no se puede calcular el índice kappa) que, comparándolo con los anteriores, claramente sobrevalora el acuerdo.

Se observa para ambas definiciones de acuerdo, un índice kappa mediano en la valoración de Landis y Koch [10] y pobre en la de Fleiss [11], comentadas en el capítulo 3. La prevalencia media observada es cercana a 0,5 y, salvo para el primer observador (MED2 en el listado), las frecuencias marginales son similares, por lo tanto, el índice kappa no está alterado por las paradojas de Feinstein y Cicchetti [14]. Se observa también que a pesar del pequeño tamaño muestral, la diferencia entre el índice kappa y su estimación por el método "jackknife" no llega al 8% y que, por lo tanto, se puede aceptar con confianza la estimación del error estándar por este

método. Como era de esperar, con tan pequeño tamaño muestral, la estimación del índice kappa es imprecisa. Nótese, sin embargo, que el tamaño muestral se corresponde con el necesario según la estimación de Cicchetti [57], (N=8) corregido por el factor de eficiencia del diseño del estudio. Es necesario aclarar que la fórmula de Cicchetti se estableció para dos observadores y para más de dos categorías. En base a este resultado, para el estudio del *nevus flammeus* se sobredimensionó el tamaño estimado por la misma fórmula.

Si se elimina del análisis al observador más discrepante según las frecuencias marginales (MED2), los índices aumentan (0,5296 con EE=0,2559 para el acuerdo por pares y 0,4545 con EE=0,4036 para el acuerdo por unanimidad).

8.1.2 Neuropatía periférica

```
P R E S T A  PC   V2.2                              12-JUL-1996

NALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: neuro

CONDICIONES DEL DISEÑO
Nº CASOS:                10
Nº CASOS ELIMINADOS:      0
Nº VARIABLES:             6    DISEÑO INCOMPLETO
Nº CATEGORIAS:            3

          FRECUENCIAS MARGINALES
                   CATEGORIA
CONDICION       1       2       3
MED2         .400    .200    .400
MED4         .400    .600    .000
MED5         .600    .200    .200
MED6         .400    .200    .400
MED7         .400    .200    .400
MED8         .400    .400    .200

PROMEDIO     .433    .300    .267
```

```
A) ACUERDO POR PARES SIN PESOS

PROPORCION DE ACUERDO OBSERVADO:    .6667
PROPORCION DE ACUERDO ESPERADO:     .3387

INDICE KAPPA:                       .4960
ESTIMADOR "JACKKNIFE":              .4995   ERROR ESTANDAR:     .1387
INTERVALO DE CONFIANZA AL 95%:      .1885   .8105


B) ACUERDO POR PARES CON PESOS

PROPORCION DE ACUERDO OBSERVADO:    .8667
PROPORCION DE ACUERDO ESPERADO:     .6607

INDICE KAPPA:                       .6071
ESTIMADOR "JACKKNIFE":              .6095   ERROR ESTANDAR:     .1738
INTERVALO DE CONFIANZA AL 95%:      .2197   .9992

MATRIZ DE PESOS USADA

             1       2       3
     1    1.00     .75     .00
     2     .75    1.00     .75
     3     .00     .75    1.00


C) ACUERDO POR MAYORIA DE   3 OBSERVADORES

PROPORCION DE ACUERDO OBSERVADO:    .5000
PROPORCION DE ACUERDO ESPERADO:     .1176

INDICE KAPPA:                       .4334
ESTIMADOR "JACKKNIFE":              .4373   ERROR ESTANDAR:     .1622
INTERVALO DE CONFIANZA AL 95%:      .0736   .8009
```

Los pesos usados, en la opción b), para el acuerdo por pares son los bicuadrados (fórmula [3.9]). Con todas las opciones, se observa un mayor acuerdo en esta variable que en las contracturas articulares (moderado en las valoraciones usadas, para ambas definiciones de acuerdo). Las prevalencias medias observadas para las 3 categorías no son muy diferentes (recordemos que los pacientes se eligieron para tener representado todo el espectro de manifestaciones clínicas) y, salvo para el

segundo observador (MED4), las frecuencias marginales también parecen similares. El índice kappa, como era de esperar, es mayor para el acuerdo por pares con pesos, y menor para el acuerdo por unanimidad. Como la variable es ordinal parece razonable el uso de los pesos, sin embargo, el menor valor para el acuerdo por unanimidad está poniendo de manifiesto que hay algún observador discrepante. Si se elimina del análisis el observador MED4 (el más discrepante según las frecuencias marginales) todos los índices aumentan (0,7439 con EE=0,1727 para el acuerdo por pares sin pesos; 0,8888 con EE=0,08322 para el acuerdo por pares con pesos bicuadrados y 0,7379 con EE=0,2844 para el acuerdo por unanimidad).

No hay prácticamente diferencias entre los índices kappa estimados en la muestra y sus estimaciones por el método "jackknife", y las estimaciones son también muy imprecisas, sin embargo, y aunque para esta variable, por estar clasificada en tres categorías, el tamaño muestral necesario según la fórmula de Cicchetti sería mayor, los errores estándar son menores que para la variable anterior.

8.1.3 Cambios esclerodermiformes

```
P R E S T A  PC   V2.2                          17-JUL-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: esclero

CONDICIONES DEL DISEÑO
Nº CASOS:                10
Nº CASOS ELIMINADOS:      0
Nº VARIABLES:             6    DISEÑO INCOMPLETO
Nº CATEGORIAS:            4
```

FRECUENCIAS MARGINALES

	CATEGORIA			
CONDICION	1	2	3	4
MED2	.400	.200	.200	.200
MED4	.600	.200	.200	.000
MED5	.200	.000	.400	.400
MED6	.200	.000	.400	.400
MED7	.400	.600	.000	.000
MED8	.400	.000	.400	.200
PROMEDIO	.367	.167	.267	.200

A) ACUERDO POR PARES SIN PESOS

PROPORCION DE ACUERDO OBSERVADO: .6667
PROPORCION DE ACUERDO ESPERADO: .2507

INDICE KAPPA: .5552
ESTIMADOR "JACKKNIFE": .5757 ERROR ESTANDAR: .1343
INTERVALO DE CONFIANZA AL 95%: .2747 .8768

B) ACUERDO POR PARES CON PESOS

PROPORCION DE ACUERDO OBSERVADO: .9407
PROPORCION DE ACUERDO ESPERADO: .6868

INDICE KAPPA: .8108
ESTIMADOR "JACKKNIFE": .8401 ERROR ESTANDAR: .1062
INTERVALO DE CONFIANZA AL 95%: .6019 1.0782

MATRIZ DE PESOS USADA

	1	2	3	4
1	1.00	.89	.56	.00
2	.89	1.00	.89	.56
3	.56	.89	1.00	.89
4	.00	.56	.89	1.00

C) ACUERDO POR MAYORIA DE 3 OBSERVADORES

PROPORCION DE ACUERDO OBSERVADO: .5000
PROPORCION DE ACUERDO ESPERADO: .0656

```
INDICE KAPPA:                          .4649
ESTIMADOR "JACKKNIFE":                 .4825   ERROR ESTANDAR:     .1679
INTERVALO DE CONFIANZA AL 95%:         .1060   .8589
```

Resultados en todo análogos a los de la variable anterior, excepto que hay mayor diferencia, aunque sigue siendo pequeña, entre las estimaciones del índice kappa a partir de la muestra completa y las obtenidas por el método "jackknife".

Si se comparan los resultados de las tres variables, se observa que para las definiciones de acuerdo comunes a las tres (acuerdo por pares sin pesos y acuerdo por unanimidad) las proporciones de acuerdo observado son idénticas (0,6667 y 0,5 respectivamente) y que las diferencias encontradas en los índices kappa lo son a expensas de las proporciones de acuerdo esperado, que disminuyen, aumentando por ello el índice kappa, cuando aumenta el número de categorías de la clasificación. Esto no es un inconveniente del índice, ya que justamente se trata de corregir por el posible acuerdo por azar, pero pone de manifiesto otro aspecto a tener en cuenta en la interpretación de este índice y en el diseño de estudios sobre precisión de clasificaciones: las categorías usadas en la clasificación deben de ser las relevantes para el problema, ya que un número más alto, o más bajo, alteraría, vía la proporción de acuerdo esperado, el valor del índice kappa.

En base a estos resultados, antes de proceder con el estudio definitivo sobre el estado de salud de la cohorte de afectados por el Síndrome del Aceite Tóxico, se realizó otro Seminario de Entrenamiento con los nueve médicos participantes en él, en el que se presentaron estos resultados, comentándose el moderado acuerdo encontrado y se repitieron los criterios de clasificación de las variables, con una mayor participación, esta vez, de los médicos participantes.

8.2 Resultados del estudio del *nevus flammeus*

8.2.1 Color

Cuando se evalúa la concordancia entre los 6 dermatólogos, también $K^J \geq 82$ y por lo tanto, tampoco se puede calcular el estadístico GSK para contrastar la homogeneidad de las frecuencias marginales; sin embargo cuando se estudia la de los 3 que manejan a estos pacientes, sí se puede calcular. Como consecuencia, en el primer caso se juzga la homogeneidad a partir de las frecuencias estimadas y en el segundo se realiza el contraste.

Todos los dermatólogos

```
P R E S T A  PC   V2.2                          6-AGO-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: color

CONDICIONES DEL DISEÑO
Nº CASOS:                80
Nº CASOS ELIMINADOS:      0
Nº VARIABLES:             6    DISEÑO COMPLETO
Nº CATEGORIAS:            3


          FRECUENCIAS MARGINALES
                    CATEGORIA
CONDICION      1       2       3
A           .188    .637    .175
B           .275    .525    .200
C           .225    .587    .188
D           .325    .463    .213
E           .300    .463    .237
F           .213    .512    .275

PROMEDIO    .254    .531    .215
```

A) ACUERDO POR PARES SIN PESOS

PROPORCION DE ACUERDO OBSERVADO:	.7775		
PROPORCION DE ACUERDO ESPERADO:	.3914		
INDICE KAPPA:	.6344		
ESTIMADOR "JACKKNIFE":	.6380	ERROR ESTANDAR:	.0458
INTERVALO DE CONFIANZA AL 95%:	.5479	.7280	

B) ACUERDO POR PARES CON PESOS

PROPORCION DE ACUERDO OBSERVADO:	.9444		
PROPORCION DE ACUERDO ESPERADO:	.7661		
INDICE KAPPA:	.7622		
ESTIMADOR "JACKKNIFE":	.7665	ERROR ESTANDAR:	.0357
INTERVALO DE CONFIANZA AL 95%:	.6963	.8368	

MATRIZ DE PESOS USADA

	1	2	3
1	1.00	.75	.00
2	.75	1.00	.75
3	.00	.75	1.00

C) ACUERDO POR MAYORIA DE 5 OBSERVADORES

PROPORCION DE ACUERDO OBSERVADO:	.7500		
PROPORCION DE ACUERDO ESPERADO:	.1459		
INDICE KAPPA:	.7073		
ESTIMADOR "JACKKNIFE":	.7104	ERROR ESTANDAR:	.0585
INTERVALO DE CONFIANZA AL 95%:	.5954	.8254	

Para todos los índices, se encuentra un acuerdo "sustancial" en la valoración de Landis y Koch, siendo mayor, como siempre, cuando se usan los pesos; en este caso es notable lo alto del índice de acuerdo por mayoría; las frecuencias marginales son similares, excepto quizás para el médico A que tiende a clasificar una mayor proporción de "Rojo" y las prevalencias están alejadas de los valores extremos, por

lo que tampoco en este estudio el índice kappa está alterado por las paradojas de Feinstein y Cicchetti. Las diferencias entre los índices kappa y sus estimaciones *jackknife* son muy pequeñas, por lo tanto la estimación del error estándar por esta técnica es muy fiable; el error estándar es pequeño en todos los casos, lo que ilustra que con este tamaño muestral tenemos suficiente precisión de la estimación.

Sólo los dermatólogos que tratan esta patología

```
P R E S T A  PC   V2.2                              6-AGO-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: color

CONDICIONES DEL DISEÑO
Nº CASOS:                 80
Nº CASOS ELIMINADOS:       0
Nº VARIABLES:              3     DISEÑO COMPLETO
Nº CATEGORIAS:             3


          FRECUENCIAS MARGINALES
                    CATEGORIA
CONDICION       1       2       3
B            .275    .525    .200
C            .225    .587    .188
D            .325    .463    .213

PROMEDIO     .275    .525    .200

PRUEBA DE HOMOGENEIDAD DE LAS FRECUENCIAS MARGINALES
ESTADISTICO GSK:                    8.3479   G.L.:   4   p= .078560


A) ACUERDO POR PARES SIN PESOS

PROPORCION DE ACUERDO OBSERVADO:    .7833
PROPORCION DE ACUERDO ESPERADO:     .3891

INDICE KAPPA:                       .6454
ESTIMADOR "JACKKNIFE":              .6489   ERROR ESTANDAR:     .0594
INTERVALO DE CONFIANZA AL 95%:      .5321   .7657
```

```
B) ACUERDO POR PARES CON PESOS

PROPORCION DE ACUERDO OBSERVADO:    .9458
PROPORCION DE ACUERDO ESPERADO:     .7651

INDICE KAPPA:                       .7694
ESTIMADOR "JACKKNIFE":              .7739   ERROR ESTANDAR:     .0440
INTERVALO DE CONFIANZA AL 95%:      .6875   .8604


C) ACUERDO POR MAYORIA DE   3 OBSERVADORES

PROPORCION DE ACUERDO OBSERVADO:    .6750
PROPORCION DE ACUERDO ESPERADO:     .1707

INDICE KAPPA:                       .6081
ESTIMADOR "JACKKNIFE":              .6110   ERROR ESTANDAR:     .0651
INTERVALO DE CONFIANZA AL 95%:      .4830   .7390
```

En las condiciones de este diseño $K^1 = 27$ y sí se ha podido calcular el estadístico GSK, con el que no se puede rechazar la hipótesis de homogeneidad de las frecuencias marginales al nivel de significación habitual de α=0,05. Los pesos usados en la opción B) son los mismos que en el apartado anterior (pesos "bicuadrados").

Estos resultados son prácticamente superponibles a los del apartado anterior, salvo una ligera disminución en el kappa por mayoría; hay que hacer notar que ahora la mayoría es más estricta pues se trata de unanimidad ya que, como se señaló en el estudio del SAT, con sólo 3 observadores no es recomendable definir una mayoría de 2, mientras que para todos los dermatólogos se usó una mayoría de 5 sobre 6 observadores.

Concluyendo, la evaluación subjetiva del color de los angiomas por distintos

dermatólogos es suficientemente precisa y no está influida por la familiaridad con el manejo de los pacientes que los padecen.

8.2.2 Aclaramiento

En las condiciones de diseño de este estudio $K^J = 64$ y también se ha podido calcular el estadístico GSK.

```
P R E S T A  PC   V2.2                              6-AGO-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: aclar

CONDICIONES DEL DISEÑO
Nº CASOS:                80
Nº CASOS ELIMINADOS:      0
Nº VARIABLES:             3    DISEÑO COMPLETO
Nº CATEGORIAS:            4


        FRECUENCIAS MARGINALES
                   CATEGORIA
CONDICION      1      2      3      4
BC          .050   .175   .325   .450
CC          .050   .162   .412   .375
DC          .038   .213   .275   .475

PROMEDIO    .046   .183   .338   .433


PRUEBA DE HOMOGENEIDAD DE LAS FRECUENCIAS MARGINALES
ESTADISTICO GSK:                    8.2137   G.L.:   6   p= .221885


A) ACUERDO POR PARES SIN PESOS

PROPORCION DE ACUERDO OBSERVADO:    .6625
PROPORCION DE ACUERDO ESPERADO:     .3346

INDICE KAPPA:                       .4928
ESTIMADOR "JACKKNIFE":              .4955   ERROR ESTANDAR:     .0503
INTERVALO DE CONFIANZA AL 95%:      .3965   .5944
```

B) ACUERDO POR PARES CON PESOS

PROPORCION DE ACUERDO OBSERVADO:	.9611		
PROPORCION DE ACUERDO ESPERADO:	.8277		
INDICE KAPPA:	.7743		
ESTIMADOR "JACKKNIFE":	.7792	ERROR ESTANDAR:	.0355
INTERVALO DE CONFIANZA AL 95%:	.7093	.8490	

MATRIZ DE PESOS USADA

	1	2	3	4
1	1.00	.89	.56	.00
2	.89	1.00	.89	.56
3	.56	.89	1.00	.89
4	.00	.56	.89	1.00

C) ACUERDO POR MAYORIA DE 3 OBSERVADORES

PROPORCION DE ACUERDO OBSERVADO:	.5000		
PROPORCION DE ACUERDO ESPERADO:	.1232		
INDICE KAPPA:	.4298		
ESTIMADOR "JACKKNIFE":	.4321	ERROR ESTANDAR:	.0577
INTERVALO DE CONFIANZA AL 95%:	.3186	.5456	

Se observa que no se puede rechazar la hipótesis de frecuencias marginales homogéneas con el nivel de significación de α=0,05; sin embargo la prevalencia de la primera categoría de aclaramiento (de 0 a 24%) es muy baja por lo que el índice kappa resulta penalizado. Algunos autores, p.e. Elmore et al. [12], para evitar este problema recurren a un muestreo estratificado para "enriquecer" la muestra con las categorías de baja prevalencia. Aquí se ha preferido el muestreo aleatorio basado en nuestra opinión de que la primera paradoja de Feinstein [14] no es realmente una paradoja y en que es preferible evaluar el sistema de clasificación en las condiciones más parecidas a las de aplicación que sea posible: los evaluadores están acostumbrados a las prevalencias realmente existentes, si se les presentara una serie de fotografías en las que la proporción de "ausencia de aclaramiento" fuera varias

veces mayor que la habitual, es posible que cambiaran los criterios de clasificación que se pretenden evaluar. Los índices encontrados oscilan entre moderado y sustancial, siendo el más alto (0,7743) el calculado con los pesos bicuadrados que es el indicado para variables ordinales. Si se calcula el CCI para las variables originales, en la escala 0 a 100, que es la alternativa de análisis para variables continuas, resulta 0,8837. Es de destacar que el CCI presupone distribución normal y nuestros datos presentan una distribución muy sesgada a la derecha.

Si, para evitar el efecto de la baja prevalencia de la "ausencia de aclaramiento", se colapsaran las dos primeras categorías en una sola, los índices kappa resultan 0,5502 con EE=0,0541 para el acuerdo por pares sin pesos; 0,7653 con EE=0,0358 para el acuerdo por pares con pesos bicuadrados y 0,4977 con EE=0,0608 para el acuerdo por unanimidad, que no suponen cambios sustanciales con los calculados para el diseño original.

En resumen, nos inclinamos por el índice kappa con pesos bicuadrados para la clasificación original que revela un acuerdo sustancial entre dermatólogos.

8.3 Resultados del estudio sobre uso apropiado de la angioplastia

En las condiciones de este estudio $K^J = 9^9 \geq 82$ y, por lo tanto, no se ha podido calcular el estadístico GSK, y en consecuencia la homogeneidad de las frecuencias marginales se juzga a partir de las frecuencias estimadas.

8.3.1 Riesgo quirúrgico BAJO/ MODERADO

```
P R E S T A  PC   V2.2                              14-OCT-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: rbajoa

CONDICIONES DEL DISEÑO
Nº CASOS:                445
Nº CASOS ELIMINADOS:       0
Nº VARIABLES:              9     DISEÑO COMPLETO
Nº CATEGORIAS:             9
```

FRECUENCIAS MARGINALES

	CATEGORIA								
CONDICION	1	2	3	4	5	6	7	8	9
MED1	.020	.007	.009	.000	.000	.004	.022	.180	.757
MED2	.043	.052	.054	.063	.151	.112	.153	.166	.207
MED3	.007	.099	.043	.124	.070	.101	.119	.231	.207
MED4	.276	.022	.049	.022	.083	.004	.009	.036	.497
MED5	.288	.049	.025	.067	.079	.052	.103	.160	.178
MED6	.022	.101	.090	.072	.079	.139	.108	.220	.169
MED7	.076	.124	.022	.013	.070	.022	.079	.151	.443
MED8	.162	.054	.074	.002	.090	.018	.081	.061	.458
MED9	.234	.000	.025	.000	.135	.004	.052	.000	.551
PROMEDIO	.125	.056	.043	.040	.084	.051	.081	.134	.385

```
ACUERDO POR RAND:   A9S

PROPORCION DE ACUERDO OBSERVADO:    .2562
PROPORCION DE ACUERDO ESPERADO:     .0081

INDICE KAPPA:                       .2501
ESTIMADOR "JACKKNIFE":              .2502   ERROR ESTANDAR:    .0200
INTERVALO DE CONFIANZA AL 95%:      .2115   .2890
```

Se observa, como era de esperar para este tamaño muestral, que la estimación "jackknife" prácticamente coincide con la obtenida en la muestra y que el error estándar es muy pequeño, por lo tanto la estimación es muy precisa. El acuerdo

observado es mediano; las frecuencias marginales no parecen homogéneas, mostrando una llamativa distribución bimodal, con las modas en los extremos; resulta también llamativo la inversión de las modas del MED5. Se observa también que la proporción de acuerdo esperado en la hipótesis de independencia es muy bajo y, por lo tanto, no hay apenas diferencia entre el acuerdo esperado y el índice kappa. Teniendo en cuenta la fórmula [4.7] para la proporción de acuerdos esperados; con nueve categorías, no es sorprendente que con una definición muy estricta de acuerdo, el acuerdo esperado sea despreciable. Para evaluar este resultado se numeraron "las condiciones clínicas" y se extrajeron 10 muestras, de tamaños diferentes, de condiciones consecutivas empezando por una elegida al azar, y sobre ellas se hicieron los cálculos. Los resultados se muestran en la Tabla 11, donde se observa que en todas las muestras, salvo en una, se repite el mismo resultado, es decir, el acuerdo esperado en la hipótesis de independencia es cercano a 0, y por lo tanto, la proporción de acuerdo observado, menos costosa de calcular, puede ser un buen índice del acuerdo.

Usando la definición menos estricta de acuerdo (AE), los resultados son:

```
ACUERDO POR RAND:    AE

PROPORCION DE ACUERDO OBSERVADO:    .5236
PROPORCION DE ACUERDO ESPERADO:     .2190

INDICE KAPPA:                       .3900
ESTIMADOR "JACKKNIFE":              .3914   ERROR ESTANDAR:     .0262
INTERVALO DE CONFIANZA AL 95%:      .3405   .4424
```

donde se observa un considerable aumento del acuerdo observado, pero también del acuerdo esperado en la hipótesis de independencia, por lo que la corrección

introducida por el índice kappa ahora sí es apreciable.

Tabla 11

Proporción de acuerdo esperado para A9S en función del tamaño muestral

Tamaño muestral	Proporción de acuerdo		Indice kappa
	Observado	Esperado	
5	0,0000	0,0000	0,000
10	0,0000	0,0000	0,0000
15	0,6667	0,4770	0,3627
20	0,1500	0,0065	0,1444
25	0,1200	0,0205	0,1016
30	0,0333	0,0001	0,0333
35	0,1143	0,0003	0,1141
40	0,2000	0,0056	0,1955
50	0,1800	0,0026	0,1779
100	0,2200	0,0032	0,2175

8.3.2 Riesgo quirúrgico ALTO

```
P R E S T A  PC   V2.2                          14-OCT-1996

ANALISIS DE CONCORDANCIA. INDICE KAPPA

NOMBRE DE LOS DATOS: raltoa


CONDICIONES DEL DISEÑO
Nº CASOS:                445
Nº CASOS ELIMINADOS:       0
Nº VARIABLES:              9     DISEÑO COMPLETO
Nº CATEGORIAS:             9
```

FRECUENCIAS MARGINALES

	CATEGORIA								
CONDICION	1	2	3	4	5	6	7	8	9
MED1	.020	.016	.000	.000	.007	.076	.191	.674	.016
MED2	.058	.056	.094	.088	.173	.139	.133	.119	.139
MED3	.045	.121	.124	.052	.031	.128	.171	.187	.142
MED4	.364	.036	.031	.016	.058	.004	.004	.013	.472
MED5	.330	.058	.047	.061	.070	.049	.108	.133	.144
MED6	.022	.101	.090	.076	.076	.139	.119	.258	.117
MED7	.124	.124	.034	.020	.092	.063	.070	.218	.256
MED8	.258	.049	.074	.018	.090	.025	.067	.038	.380
MED9	.234	.000	.025	.000	.137	.004	.054	.004	.542
PROMEDIO	.162	.062	.058	.037	.082	.070	.102	.183	.245

```
ACUERDO POR RAND:  A9S


PROPORCION DE ACUERDO OBSERVADO:    .1753
PROPORCION DE ACUERDO ESPERADO:     .0025

INDICE KAPPA:                       .1732
ESTIMADOR "JACKKNIFE":              .1732   ERROR ESTANDAR:     .0178
INTERVALO DE CONFIANZA AL 95%:      .1388   .2077
```

Resultados en todo similares a la situación de riesgo bajo, tanto en la distribución de las frecuencias marginales, como en los acuerdos observado y

esperado, aunque aquí algo menores, la igualdad entre los estimadores muestral y "jackknife" y el pequeño error estándar del estimador.

Para el acuerdo estadístico (AE):

```
ACUERDO POR RAND:  AE

PROPORCION DE ACUERDO OBSERVADO:    .4539
PROPORCION DE ACUERDO ESPERADO:     .1144

INDICE KAPPA:                       .3834
ESTIMADOR "JACKKNIFE":              .3835   ERROR ESTANDAR:     .0231
INTERVALO DE CONFIANZA AL 95%:      .3387   .4284
```

Se repite la misma situación, es decir, el acuerdo para las situaciones de riesgo quirúrgico alto es ligeramente menor que para las de riesgo bajo y con la definición menos estricta de acuerdo aumentan tanto el acuerdo observado como el esperado.

9. CONCLUSIONES

1. La generalización 4.2 del índice kappa permitió analizar el acuerdo entre médicos en el estudio del SAT, en el que, por el carácter de las pruebas y de los pacientes, era necesario recurrir a un diseño incompleto.

2. La generalización 4.1 permite evaluar el acuerdo para diferentes definiciones del mismo, lo que supone una ayuda en su interpretación.

3. La generalización 4.3 tiene como principal incoveniente el tiempo de cálculo necesario para generar las permutaciones que permiten calcular la proporción de acuerdo esperado. Este inconveniente se matiza teniendo en cuenta que los ordenadores personales más recientes, con procesador Pentium, son aproximadamente 3 veces más rápidos que el ordenador usado en esta tesis (Intel 486 a 70 MHz) al que corresponden los tiempos de cálculo presentados.

   3.1 Para la definición más estricta (A9S) de acuerdo, además esta proporción es muy cercana a cero y, por lo tanto, no supone ninguna ventaja el uso del índice kappa.

   3.2 Para la definición más lasa (AE), en que el acuerdo esperado ya es importante, el tiempo de cálculo se dispara a 7 días (2,5 para generar las permutaciones y el acuerdo esperado y 4,5 para el estimador *jackknife*), obteniéndose un índice kappa similar al acuerdo observado para la definición A9S.

En definitiva nuestra propuesta es evaluar el acuerdo mediente el acuerdo observado con la definición más estricta (A9S).

4. La precisión de la técnica *jackknife*, estimada por la diferencia entre el estimador muestral y el estimador *jackknife*, es suficiente incluso para pequeños tamaños muestrales (n=10 en el estudio del SAT).

5. En cuanto a los resultados de los tres estudios presentados: el acuerdo entre médicos resultó moderado en el estudio del SAT, alto en el del *nevus flammeus* y mediano en la primera ronda de los panelistas para el uso apropiado de los procedimientos quirurgicos para la enfermedad coronaria, que ponen, una vez más, de manifiesto la necesidad de seguir investigando sobre el acuerdo entre médicos en la percepción de las variables clínicas y de disponer de herramientas metodológicas para ello.

APÉNDICE I

Tablas a partir de las que se han obtenido los datos de la Figura 1

Proporción de acuerdos observados igual a 0,80

		Prev. Media	Kappa
2	8	0,12	0,057
12	78		
10	14	0,20	0,381
6	70		
25	2	0,35	0,605
18	75		
40	8	0,50	0,601
12	40		

Proporción de acuerdos observados igual a 0,70

		Prev. Media	Kappa
0	3	0,15	-0,057
27	70		
5	18	0,20	0,068
12	65		
20	9	0,35	0,351
21	50		
35	11	0,50	0,404
19	35		

Proporción de acuerdos observados igual a 0,60

		Prev. Media	Kappa
0	4	0,20	-0,078
36	60		
15	7	0,35	0,167
33	45		
30	26	0,50	0,211
14	30		

APÉNDICE II

Listado de los programas:

CONTIN: Análisis de tablas de contingencia, que incluye el cálculo del índice kappa para dos observadores, con los errores estándar en los supuestos de independencia y no independencia, y los estadísticos de McNemar y Stuart-Maxwell.

```
$storage:2
C  CONTIN.v2
C
C  ANALIZA LA INDEPENDENCIA DE DOS CARACTERES CUALITATIVOS
C  MEDIANTE LA CHI-CUADRADO
C
C  V. ABRAIRA \ DPTO INVESTIGACION \ Hosp.  'RAMON Y CAJAL' - MADRID
C  ENE 81
c  MODIFICADO FEB 93 PARA INCLUIR KAPPA Y TESTS DE McNEMAR
c  MODIFICADO MAY 94 PARA EE PARA KAPA<>0 Y ESTADISTICO DE STUART-MAXWELL
C
      PROGRAM CONTIN
      DIMENSION D(2000),TR(50),TC(40),NE(2),DD(4)
      CHARACTER CLAVE*6,CLAPRO*6,NOM*9,NOMDAT*16
      COMMON ICOD,KKK,XNEP
      ICAR1=50
      ICAR2=40
      OPEN (8,FILE='CONTIN.SAL')
      CALL LEECLA(CLAVE)
      CALL ACTPRE(65)
C
C  LEE DATOS
C  ---------
310   WRITE(*, 100)
100   FORMAT(/' NOMBRE DE LOS DATOS (←┘  para acabar TABLAS DE CONTINGE
     1NCIA) : '\)
      READ(*, 110)NOM
110   FORMAT(A)
      IF (NOM(9:9).NE.' ') GOTO 310
      IF (NOM(1:1).EQ.' ') GOTO 1011
2200  WRITE(*, 120)
120   FORMAT(//' ESTOS DATOS SON UNA TABLA DE CONTINGENCIA (S/N) [N] : '
     1\)
      IRES=RUM(66,78.,0,0.,0.)
      IF (IRES.GT.90) IRES=IRES-32
      IF (IRES.NE.83.AND.IRES.NE.78) GOTO 2200
      IF (MIRA(NOM,1,NOMDAT,0,0).EQ.1) GOTO 1000
      PAUSE '<P> NO EXISTEN DATOS CON ESE NOMBRE (←┘ para seguir)'
      GOTO 310
1000  OPEN(1,FILE=NOMDAT,ACCESS='DIRECT',RECL=26)
      READ(1,REC=1)CLAPRO,NC,NV,NX,ICOD,NCA,LONAR,XNEP,NVNE
      CLOSE(1)
      NVNR=NV-NVNE
      IF (NVNE.EQ.0.OR.IRES.EQ.78) GOTO 222
      WRITE(*, 152)
      CALL PITO(2)
152   FORMAT(//' ESTE ARCHIVO TIENE VARIABLES CON CODIGOS NO EXCLUYENTES
     1'/' NO PUEDE SER UNA TABLA DE CONTINGENCIA')
      GOTO 901
222   IF (ICOD.EQ.1.OR.IRES.EQ.78) GOTO 2010
      WRITE(*, 151)
      CALL PITO(2)
151   FORMAT(//' ESTOS DATOS SON REALES. LAS TABLAS DE CONTINGENCIA DEBE
     1N TENER DATOS ENTEROS')
```

```
      GOTO 901
2010  LONRE=NV*ICOD*2
      LONPRO=(25+LONRE)/LONRE
      IF (CLAVE.EQ.CLAPRO) GOTO 1002
      PAUSE '<P> LA PALABRA CLAVE DE ESOS DATOS NO ES LA SUYA (◄┘ para
     1seguir)'
      GOTO 310
1002  IF (IRES.NE.83) GOTO 2
      IF (NC.LE.ICAR2) GOTO 1
      WRITE(*, 130)NC,2,ICAR2
130   FORMAT(//' ESTE ARCHIVO TIENE',I4,' ESTADOS EN EL CARACTER',I2
     1/' EL PROGRAMA SOLO ADMITE',I3,' . LO SIENTO')
      GOTO 901
1     IF (NV.LE.ICAR1) GOTO 2
      WRITE(*, 130)NV,1,ICAR1
      GOTO 901
2     IF (NCA.EQ.NC) GOTO 1003
      WRITE(*, 1010)NCA,NC
1010  FORMAT(//' ESTE ARCHIVO NO ESTA COMPLETO. TIENE',I5,
     1' CASOS Y DECLARO',I5)
      PAUSE '(◄┘ para seguir)'
      NC=NCA
1003  OPEN(1,FILE=NOMDAT,ACCESS='DIRECT',RECL=LONRE)
      KKK=1+LONPRO
      IF (IRES.NE.83) GOTO 2100
      DO 5 I=1,NC
      IJ=(I-1)*NV+1
      CALL DATA(D(IJ),NV,NVNR,1,1)
5     CONTINUE
      NE(2)=NC
      NE(1)=NV
      CALL CABECE(8)
      WRITE(8,555,ERR=999)NOM
      WRITE(*,555)NOM
555   FORMAT(/' ANALISIS DE TABLAS DE CONTINGENCIA'//' ARCHIVO: 'A//
     1' PRUEBA DE ASOCIACION')
      GOTO 2020
2100  CALL TABLA(NOM,NVNR,NC,NV,NE,D)
C
C  CALCULA CHI-CUADRADO
C  --------------------
2020  CLOSE(1)
      IF (NE(1).NE.2.OR.NE(2).NE.2) GOTO 1100
      IERR=0
1110  WRITE(*, 1101)
1101  FORMAT(//\' QUIERE JI-CUADRADO (J) O FISHER (F) [J] : ')
      IRES=RUM(66,74.,0,0.,0.)
      IF (IRES.GT.90) IRES=IRES-32
      IF (IRES.NE.74.AND.IRES.NE.70) GOTO 1110
      IF (IRES.EQ.70) GOTO 1020
1100  CALL CHISQ(D,NE(1),NE(2),CC,NGL,IERR,TR,TC)
      I=1
      GOTO (10,20,30,30,40) IERR+1
30    CALL PITO(2)
      WRITE(8,210,ERR=999)
      WRITE(*,210)
210   FORMAT(//' ALGUN CARACTER SOLO TIENE UN ESTADO Y/O'/
     1' ALGUN(OS) VALOR(ES) TEORICO(S) ES = 0.')
      GOTO 900
20    CALL PITO(2)
      WRITE(8,200,ERR=999)
      WRITE(*,200)
200   FORMAT(//' ALGUN(OS) VALOR(ES) TEORICO(S) ES < 1'/
     1' O MAS DEL 20% SON < 5')
      GOTO 900
40    CALL PITO(2)
      WRITE(8,231,ERR=999)
```

```
      WRITE(*,231)
231   FORMAT(//' LA PRUEBA JI-CUADRADO ES INAPLICABLE, VOY A USAR LA DE
     1FISHER')
1020  DO 1021 JJ=1,4
1021  DD(JJ)=D(JJ)
      CALL FISHER(DD,XP)
      NCOL=1
      IF (XP.LT.2.) GOTO 3000
      CALL PITO(2)
      WRITE(8,3010,ERR=999)
      WRITE(*,3010)
3010  FORMAT (//' HAY DESBORDES EN EL CALCULO DEL TEST DE FISHER')
      IF (IERR.EQ.4) GOTO 900
      GOTO 1110
999   STOP 10
10    WRITE(*,220)CC,NGL
      WRITE(8,220,ERR=999)CC,NGL
220   FORMAT(/' JI-CUADRADO :',F8.2,' CON',I4,' GRADOS DE LIBERTAD')
      NCOL=2
      XP=PROB(1,CC,NGL,0)
3000  IF (XP.GT..05) GOTO 500
      WRITE(*,230)XP,NCOL
      WRITE(8,230,ERR=999)XP,NCOL
230   FORMAT(/' EXISTE ASOCIACION p=',F9.7,' (',I2,' COLA/S)')
      GOTO 900
500   WRITE(*,240)XP,NCOL
      WRITE(8,240,ERR=999)XP,NCOL
240   FORMAT(/' ESTE TEST NO DEMUESTRA ASOCIACION p=',F9.7,' (',I2,
     1' COLA/S )')
C
c  CALCULA SENSIBILIDAD Y ESPECIBILIDAD CON SUS IC AL 95% SI TABLA 2x2
c  ASUME EL GOLD STANDAR EN EL CARACTER 1 Y EL ORDEN COMO - Y +
C
900   IF (NE(1).EQ.NE(2).AND.NE(1).EQ.2) THEN
      WRITE(*,5010)
      WRITE(8,5010,ERR=999)
5010  FORMAT(//' ANALISIS DE VALIDEZ'/' (Asume que el carácter 1 es la'
     1' referencia y que el 1º estado es el negativo)')
      write(*,5020)
      write(8,5020,ERR=999)
5020  FORMAT(/30X,' IC AL 95%')
      if (INT(D(2)+D(4)).EQ.0) GOTO 5070
      CALL INTCON(INT(D(2)+D(4)),int(D(4)),XI,XS,IER)
      IF (IER.EQ.0) THEN
      write(*,5030) D(4)/(D(2)+D(4)),XI,XS
      write(8,5030,ERR=999) D(4)/(D(2)+D(4)),XI,XS
5030  FORMAT(' Sensibilidad =',f6.3,5x,2f7.3)
      else
      write(*,5040) D(4)/(D(2)+D(4))
      write(8,5040,ERR=999) D(4)/(D(2)+D(4))
5040  FORMAT(' Sensibilidad =',f6.3,9X,'Ver tablas')
      endif
5070  if (INT(D(1)+D(3)).EQ.0) GOTO 5080
      CALL INTCON(INT(D(1)+D(3)),int(D(1)),XI,XS,IER)
      IF (IER.EQ.0) THEN
      write(*,5050) D(1)/(D(1)+D(3)),XI,XS
      write(8,5050,ERR=999) D(1)/(D(1)+D(3)),XI,XS
5050  FORMAT(' Especificidad=',f6.3,5x,2f7.3)
      else
      write(*,5060) D(1)/(D(1)+D(3))
      write(8,5060,ERR=999) D(1)/(D(1)+D(3))
5060  FORMAT(' Especificidad=',f6.3,9X,'Ver tablas')
      endif
      ENDIF
C
C  CALCULA KAPPA SI TABLA CUADRADA Y MCNEMAR SI 2x2 O STUART-MAXWELL SI NXN
c
```

```
5080  IF (NE(1).EQ.NE(2)) THEN
        CALL KAPPA(D,NE(1),XKAP,EKAP0,EKAP1,XMAC,IER)
        IF (IER.EQ.0.OR.IER.EQ.2) THEN
          IF (EKAP0.NE.0.) THEN
            write(*,700) XKAP,EKAP0,PROB(3,XKAP/EKAP0,0,0)
            write(8,700,ERR=999) XKAP,EKAP0,PROB(3,XKAP/EKAP0,0,0)
700   FORMAT(//' ANALISIS DE CONCORDANCIA'/
     1' ESTADISTICO KAPPA:',F12.3,5X/' E.E.'
     1' para KAPPA= 0    :',F8.3,15X,'p=',f9.7)
           ELSE
            write(*,705) XKAP,EKAP0
            write(8,705,ERR=999) XKAP,EKAP0
705   FORMAT(//'ANALISIS DE CONCORDANCIA'/
     1' ESTADISTICO KAPPA:',F12.3,5X/' E.E. para KAPPA= 0    :',F8.3)
          ENDIF
         write(*,707) EKAP1,XKAP-1.96*ekap1,XKAP+1.96*EKAP1
         write(8,707,ERR=999) EKAP1,XKAP-1.96*EKAP1,XKAP+1.96*EKAP1
707   FORMAT(' E.E. para KAPPA ESTIMADO:',F5.3,5X,'I.C. al 95%:'
     1,2f7.3)
       ENDIF
        IF (IER.EQ.0.OR.IER.EQ.1) THEN
          WRITE (*,711)
          WRITE (8,711,ERR=999)
711   FORMAT(//' PRUEBA DE HOMOGENEIDAD DE MARGINALES')
          IF (NE(1).EQ.2) THEN
            write(*,710) XMAC,PROB(1,XMAC,1,0)
            write(8,710,ERR=999) XMAC,PROB(1,XMAC,1,0)
710   FORMAT(' ESTADISTICO DE McNEMAR:',F8.3,14X,'p=',f9.7)
           ELSE
            write(*,713) XMAC,PROB(1,XMAC,NE(1)-1,0)
            write(8,713,ERR=999) XMAC,PROB(1,XMAC,NE(1)-1,0)
713   FORMAT(' ESTADISTICO DE STUART-MAXVELL:',F8.3,7X,'p=',f9.7)
          ENDIF
        ENDIF
      ENDIF
901   WRITE(*, 300)
300   FORMAT(/\' OTRA TABLA DE CONTINGENCIA (S/N) [N] : ')
      IRES=RUM(66,78.,0,0.,0.)
      IF (IRES.GT.90) IRES=IRES-32
      IF (IRES.EQ.83) GOTO 310
      CLOSE(8)
      WRITE(*,'(//\'' QUIERES COPIA EN PAPEL (S/N) [S] : '')')
      IQ=RUM(66,83.,0,0.,0.)
      IF (IQ.GT.90)IQ=IQ-32
      IF (IQ.EQ.83) CALL COPIA('CONTIN.sal')
1011  CONTINUE
      END

$STORAGE:2
      SUBROUTINE KAPPA(A,N,XK,EKAP0,EKAP1,XMAC,IER)
C
C  CALCULA EL KAPPA DE UNA TABLA DE CONTINGENCIA
C  DESCRIPCIÓN DE LOS PARAMETROS
C     A     - MATRIZ N x N CONTENIENDO LA TABLA DE CONTINGENCIA
C     N     - DIMENSION DE A
C     XK    - KAPPA
C     EKAP0 - ERROR ESTANDAR PARA KAPPA=0
C     EKAP1 - ERROR ESTANDAR PARA KAPPA <>0
C     XMAC  - ESTADISTICO DE MCNEMAR SI N=2 O STUART-MAXVELL SI N<>2
C     IER    - CODIGO DE ERROR
C              0 - SIN ERROR
C              1 - ACUERDOS ESPERADOS CON PROB 1
C              2 - NO SE PUEDE CALCULAR MCNEMAR O STUART-MAXVELL
C              3 - NO HAY ELEMENTOS EN LA TABLA
C V.Abraira \ Unidad de Biosestadística Clínica
C Hosp. Ramón y Cajal. Madrid
C
```

```
      DIMENSION A(1),TR(40),TC(40),L(39),M(39)
      DOUBLE PRECISION D(39),VAR(39),DET,MM(39),DXMAC
      IER=0
C
C  CALCULA LA SUMA DE LAS FILAS
C
10    DO 95 I=1,N
      TR(I)=0.
      IJ=I-N
      DO 90 J=1,N
      IJ=IJ+N
90    TR(I)=TR(I)+A(IJ)
95    CONTINUE
C
C  CALCULA LA SUMA DE COLUMNAS
C
      IJ=0
      DO 105 J=1,N
      TC(J)=0.
      DO 100 I=1,N
      IJ=IJ+1
100   TC(J)=TC(J)+A(IJ)
105   CONTINUE
C
C  CALCULA SUMA TOTAL
C
      GT=0.
      DO 110 I=1,N
110   GT=GT+TR(I)
      IF (GT.NE.0.) GOTO 170
      IER=3
      RETURN
C
C  CALCULA PROPORCION DE ACUERDOS OBSERVADOS Y ESPERADOS
C
170   XPO=0.
      XPE=0.
      RPE=0.
      DO 200 I=1,N
      XPO=XPO+A((I-1)*N+I)/GT
      XPE=XPE+TR(I)*TC(I)/GT**2.
200   RPE=RPE+TR(I)*TC(I)*(TR(I)+TC(I))/GT**3.
      IF (XPE.NE.1.) GOTO 300
      IER=1
      RETURN
300   XK=(XPO-XPE)/(1.-XPE)
      XA=0.
      XB=0.
      DO 400 I=1,N
      J=(I-1)*N+I
      XA=XA+(A(J)/GT)*(1.-(TR(I)+TC(I))*(1.-XK)/GT)**2.
      DO 410 II=1,N
      IF (I.EQ.II) GOTO 410
      XB=XB+A((I-1)*N+II)/GT*((TR(I)+TC(II))/GT)**2.
410   CONTINUE
400   CONTINUE
      XB=XB*(1.-XK)**2.
      XC=(XK-XPE*(1.-XK))**2.
C PRECUACION POR ERRORRES DE REDONDEO
      IF (XPE+XPE**2.-RPE.LE.0) THEN
        EKAP0=0.
      ELSE
        EKAP0=SQRT((XPE+XPE**2.-RPE)/GT)/(1.-XPE)
      ENDIF
      EKAP1=SQRT((XA+XB-XC)/GT)/(1.-XPE)
      IF (N.EQ.2) THEN
C MCNEMAR
```

```
          IF (A(2)+A(3).NE.0.) GOTO 310
          IER=2
          RETURN
310       XMAC=(A(2)-A(3))**2./(A(2)+A(3))
       ELSE
C STUART-MAXVELL
          LL=0
          DO 500 I=1,N-1
             D(I)=DBLE(TR(I)-TC(I))
             DO 600 KK=1,I
               IF (I.EQ.KK) THEN
                  VAR(LL+KK)=DBLE(TR(I)+TC(I)-2.*A((I-1)*N+I))
               ELSE
                  VAR(LL+KK)=-DBLE(A((I-1)*N+KK)+A((KK-1)*N+I))
               ENDIF
600          CONTINUE
           LL=LL+I
500    CONTINUE
       CALL DCOMPL(VAR,N-1)
       CALL DMINV(VAR,N-1,DET,L,M)
          IF (DET.NE.0.) GOTO 510
          IER=2
          RETURN
510    CALL DGMPRD(D,VAR,MM,1,N-1,N-1)
       CALL DGMPRD(MM,D,DXMAC,1,N-1,1)
       XMAC=SNGL(DXMAC)
       ENDIF
       RETURN
       END

$STORAGE:2
       SUBROUTINE DCOMPL(A,M)
C
C  VERSION DOBLE PRECISION DE COMPLE
C
C  Dada una matriz simetrica como triangular inferior en A, la devuelve
c  completa tambien en A. M es el orden de la matriz ==> A tendra
c  (M-1)*M/2 elementos en la entrada y M*M en la salida.
c  El orden maximo es 50. Para cambiarlo poner en la siguiente DIMENSION
C  el cuadrado del nuevo orden maximo
C
       DIMENSION T(2500),A(1)
       DOUBLE PRECISION T,A
       IF (M.GT.50) STOP 33
       DO 1 I=1,M
         JJ=(I-1)*M
         DO 1 J=1,M
           IF (I-J) 3,4,4
3          L=I+(J*J-J)/2
           GOTO 1
4          L=J+(I*I-I)/2
1      T(JJ+J)=A(L)
       MM=M*M
       DO 2 I=1,MM
2      A(I)=T(I)
       RETURN
       END
```

Las subrutinas DMINV y DGMPRD, usadas en éste y en el siguiente programa, que invierte una matriz cuadrada y multiplica dos matrices generales, respectivamente, han sido tomadas de la libreria SSP [58].

KAPMUL: Programa para el cálculo del índice kappa para múltiples observadores, aceptando diseños incompletos, distintos "pesos" y distintas definiciones de acuerdo.

```
$storage:2
C  KAPMUL.v2
C  ANALISIS DE CONCORDANCIA
C  CALCULA EL INDICE KAPPA PARA VARIABLES MULTINOMIALES Y VARIOS OBSERVADORES
C  ACEPTA DATOS NO ESPECIFICADOS
C  CALCULA ERROR ESTANDAR POR LA TECNICA JACKKNIFE
C
C  V. ABRAIRA \ Unidad de Bioestadística Clínica
C  Hosp. Ramón y Cajal. Madrid
C  SEPTIEMBRE 92
C  MODIFICADO MAYO 94 PARA OPTIMIZAR JACKKNIFE Y CORREGIR CALCULO E.E.
C  modificado JUN 94 PARA INCLUIR HOMOGENEIDAD DE MARGINALES POR GSK
C               SI CAT**IVG<82
c  modificado SEP 94 PARA INCLUIR ACUERDO POR MAYORIA
c  MODIFICADO OCT 94 PARA INCLUIR ACUERDOS DE RAND
c  MODIFICADO JUL 96 PARA INCLUIR JACKKNIFE EN RAND
C
      PROGRAM KAPMUL
      DIMENSION X(100),N(100),IY(30),IC(100,30),IB(100,30),ICV(100)
     1;PT(30),PES(30,30),NN(100),NM(100),NNC(100),IPP(81),IVAR(100)
      DOUBLE PRECISION CON(128),CB(8),VF(256),TEM(1296),AAA(1296)
     1,VP(6561),DET
C LAS DIMENSIONES DE IY(),PT(), PES(,) Y LA SEGUNDA DE IC(,) Y IB(,)
C DEBEN SER MAYORES O IGUALES QUE NMAXC
      CHARACTER NOMCV*6(100),CLAVE*6,NOM*9,NOMC*16,PCP*6,NOMBRE*17
     1,DIS*10(2),MAY*7(3),DRAND*3(5)
      COMMON ICOD,KKK,XNEP
      DATA NMAXC/30/EPS/.00001/DIS/'INCOMPLETO','COMPLETO'/
      DATA MAY/'MAYORIA','PARES','RAND:'/
      DATA DRAND/'A9S','A9R','A7S','A7R','AE '/
C
      NOMBRE(17:17)='$'
      CALL LEECLA(CLAVE)
      CALL ACTPRE(81)
      write(*,'('' ANALISIS DE CONCORDANCIA. VARIABLES MULTINOMIALES''
     1)')
C
C  PREGUNTA POR ARCHIVO Y VARIABLES
C  --------------------------------
551   ip=78
      WRITE(*, 100)
100   FORMAT(/' NOMBRE DE LOS DATOS (◄┘ para terminar) :'\)
      READ(*, 200) NOM
200   FORMAT(A)
      IF (NOM(1:1).EQ.' ') GOTO 444
      IF (NOM(9:9).NE.' ') GOTO 551
      IF (MIRA(NOM,1,NOMC,0,0).EQ.1) GOTO 552
      PAUSE '<P> NO EXISTEN DATOS CON ESE NOMBRE (◄┘ para seguir)'
      GOTO 551
552   OPEN (2,FILE=NOMC,ACCESS='DIRECT',RECL=26)
      READ(2,REC=1,ERR=991) PCP,NC,NV,MS,ICOD,NA,NR,XNEP,ncne,ICMX
      IF (CLAVE.EQ.PCP) GOTO 84
      PAUSE'<P> LA PALABRA CLAVE DE ESTOS DATOS NO ES LA SUYA (◄┘ para
     1seguir)'
      goto 572
84    IF (NA.EQ.NC) GOTO 1000
      WRITE(*, 556) NA,NC
556   FORMAT(/' <P> ESTE ARCHIVO NO ESTA COMPLETO. TIENE',I5,
     1' CASOS Y DECLARO',I5)
      pause '(◄┘ para seguir)'
1000  CLOSE(2)
11    nvr=nv-ncne
        IF (NVR.LE.1) THEN
          WRITE(*, 557)
557     FORMAT(/' <P> ESTE ARCHIVO TIENE MENOS DE 2 VARIABLES NORMALES')
          GOTO 572
        ENDIF
```

```
        NP=NV*ICOD
      NP2=NP*2
      MP=(25+NP2)/NP2
      OPEN(2,FILE=NOMC,ACCESS='DIRECT',RECL=NP2)
22    WRITE(*,111) NVR
111   FORMAT(/' ¿Cuántas variables va a analizar? (máximo:',i3,'): '\)
        IVG=RUM(78,0.,1,2.,FLOAT(nvr))
      DO 4 I=1,IVG
45       WRITE(*,121)I
121     FORMAT(/I5,'ª variable :'\)
        CALL PREVAR(NOM(1:8),NV,NOMCV(I),ICV(I),0)
        IF (ICV(I).LE.0.OR.ICV(I).GT.NVR) THEN
        WRITE(*,'('' Respuesta incorrecta'')')
        GOTO 45
        ENDIF
        IF (I.EQ.1) GOTO 4
        DO 34 J=1,I-1
          IF (ICV(I).EQ.ICV(J)) GOTO 37
34        continue
      GOTO 4
37      CALL PITO(2)
        WRITE(*,'('' VARIABLE REPETIDA'')')
        GOTO 45
4       continue
C
C  PREGUNTA POR TIPO DE ACUERDO Y PESOS
C  ------------------------------------
500   JJ=MIRA(NOM,2,NOMBRE(1:16),NA,212)
      IF (JJ.EQ.1) THEN
          IF (IDELET(NOMBRE).NE.0) STOP 6
        GOTO 500
      ENDIF
      IF (JJ.EQ.4) STOP 2
C ARCHIVO *.TMP PARA JACKKNIFE: En cada caso guarda
c IA      : 1 caso eliminado; los demás valores irrelevantes
c JJ      : nº observadores para ese caso
C IR      : 1 ACUERDO
c  B      : CONTRIBUCION A XP0S DE ESE CASO
c  C      : XPE SIN ESE CASO
c  NN(J)  : 0 esa variable no está; 1 sí
        OPEN(1,FILE=NOMBRE(1:16),ACCESS='DIRECT',RECL=214)
      DO 501 I=1,NA
501      WRITE(1,REC=I)0,0,0,0.,0.,(0,II=1,IVG)
      WRITE(*,38)NMAXC
38    FORMAT(/'¿Cuántas categorías (máximo',i3,'): '\)
      ICAT=RUM(78,0.,1,2.,FLOAT(NMAXC))
      IM=2
      IOM=2
      XLIM=IVG*LOG10(FLOAT(ICAT))
      IF (IVG.GT.2.AND.XLIM.LE.38.) THEN
        WRITE(*,2000)
2000  FORMAT(/' Acuerdo por mayoría (1) o todos los pares (2) [2]: '\)
        IM=RUM(78,2.,1,1.,2.)
        IF (IM.EQ.1) THEN
          IF (IVG.EQ.9.AND.ICAT.EQ.9) THEN
2005        WRITE(*,2010)
2010    FORMAT(/' Definiciones RAND (3) o estándar (1) [1]: '\)
            IM=RUM(78,1.,1,1.,3.)
            if (im.eq.2) GOTO 2005
           endif
           IF (IM.EQ.1) THEN
             WRITE(*,2020)IVG,ivg
2020  FORMAT(/' ¿Cuántos observadores forman la mayoría? (mínimo 2,',
     1' máximo',i3,') [',i3,']: '\)
              IOM=RUM(78,FLOAT(IVG),1,2.,FLOAT(IVG))
            ELSE
              WRITE(*,2030)
```

```
2030  FORMAT(/' Definiciones de ACUERDO:'//
     1'    9 estricto  (1)'/'    9 relajado  (2)'/'    7 estricto  (3)'/
     2'    7 relajado  (4)'/'    estadístico (5)  [1]: '\)
            IRAND=RUM(78,1.,1,1.,5.)
           ENDIF
         ENDIF
       ENDIF
       IF (IM.EQ.2) THEN
       WRITE(*,'(/\'' QUIERES PESOS (S/N) [N] : '')')
       IP=RUM(66,78.,0,0.,0.)
       IF (IP.GT.90)Ip=Ip-32
       IF (IP.NE.83) THEN
       DO 300 I=1,ICAT
         DO 300 J=1,ICAT
             IF (I.EQ.J) THEN
               PES(I,J)=1.
             ELSE
               PES(I,J)=0.
300      ENDIF
       ELSE
       DO 310 I=1,ICAT
         DO 310 J=I,ICAT
           PX=1.-FLOAT(J-I)**2./(FLOAT(ICAT)-1.)**2.
           WRITE(*,378)I,J,PX
378    FORMAT(' PESO PARA LA COINCIDENCIA'I3,' -',I3,' [',F5.3,'] :'\)
         PES(I,J)=RUM(82,PX,0,0.,0.)
310      PES(J,I)=PES(I,J)
       ENDIF
       ENDIF
       KKK=1+MP
       NCQ=0
       JPP=100
c comprueba límites para estadística GSK
       IF (XLIM.LE.LOG10(81.)) JPP=ICAT**IVG
       IF (JPP.LE.81) THEN
           IGSK=1
         DO 7000 I=1,JPP
7000       IPP(I)=0
       ELSE
         IGSK=0
       ENDIF
C
C  LEE DATOS Y CALCULA FRECUENCIAS Y ACUERDO
C  ----------------------------------------
       DO 12 I=1,IVG
         N(I)=0
C  N(I): Nº DE CASOS
         DO 12 J=1,ICAT
12         IC(I,J)=0
C  IC(I,J): FRECUENCIA DE CADA CATEGORIA EN CADA VARIABLE
         XP0S=0.
       IDC=1
       IF (IM.NE.2) IACU=0
       DO 10 I=1,NA
           DO 13 J=1,ICAT
             PT(J)=0.
13         IY(J)=0
         CALL DATA(X,NV,NVR,2,0)
         IF (IM.EQ.3) THEN
             CALL RAND(X,ICV,IRAND,IR,IER)
             IF (IER.NE.0) THEN
               WRITE(*,979)I
979            FORMAT(/' EL CASO',I5,' NO ESTA COMPLETO')
               CLOSE(2)
               GOTO 551
             ENDIF
             IACU=IACU+IR
```

```
          ENDIF
        IK=0
        DO 123 J=1,IVG
            NN(J)=0
            IF (ABS(X(ICV(J))-XNEP).LE.EPS) GOTO 123
          IF (ABS(X(ICV(J))-FLOAT(INT(X(ICV(J))))).GT.EPS) GOTO 1100
          IF (X(ICV(J)).LT.1..OR.X(ICV(J)).GT.FLOAT(ICAT)) GOTO 1100
C  IK:        Nº OBSERVADORES EN ESTA OBSERVACION
C  IVAR(IK): RESPUESTA DEL OBSERVADOR IK
C  NN(J):     1 SI LA VARIABLE ENTRA
C  IY(CAT):  FRECUENCIA DE CADA CATEGORIA
            IK=IK+1
          IVAR(IK)=INT(X(ICV(J)))
          NN(J)=1
          N(J)=n(J)+1
          IC(J,INT(X(ICV(J))))=IC(J,INT(X(ICV(J))))+1
          IY(INT(X(ICV(J))))=IY(INT(X(ICV(J))))+1
123       CONTINUE
        IF (IK.NE.IVG) IDC=0
        IF (IM.EQ.1.AND.IK.GE.IOM) THEN
            IR=IREPET(IK,IOM,IVAR)
            IACU=IACU+IR
        ENDIF
        IF (IGSK.EQ.1.AND.IK.EQ.IVG) THEN
C  MATRIZ PASA GSK
          IND=1
          DO 7010 J=1,IVG
7010        IND=IND+INT(X(ICV(J))-1.)*ICAT**(IVG-J)
          IPP(IND)=IPP(IND)+1
          ENDIF
          IF (IK.LE.1.OR.(IM.EQ.1.AND.IK.LT.IOM)) THEN
            WRITE(1,REC=I)1,0,0,0.,0.,(0,J=1,IVG)
            NCQ=NCQ+1
            WRITE(*,988)I,IOM
988   FORMAT(/' EN EL CASO',I5,' HAY MENOS DE',I3,' OBSERVADORES.',
     1' LO ELIMINO')
            IF (IK.EQ.0) GOTO 10
            DO 128 J=1,ICAT
              IF (IY(J).EQ.0) GOTO 128
              DO 129 JJ=1,IVG
                IF (INT(X(ICV(JJ))).NE.J) GOTO 129
                IC(JJ,J)=IC(JJ,J)-1
                N(JJ)=N(JJ)-1
129             CONTINUE
128           CONTINUE
          goto 10
          ENDIF
          IF (IM.EQ.2) THEN
            XP=0.
            DO 124 J=1,ICAT
              DO 124 JJ=1,ICAT
124             XP=XP+PES(J,JJ)*IY(J)*IY(JJ)
            XP=(XP-FLOAT(IK))/(FLOAT(IK)*(IK-1.))
            XP0S=XP0S+XP
          ENDIF
          WRITE(1,REC=I)0,IK,IR,XP,0.,(NN(J),J=1,IVG)
10      CONTINUE
        IF (NA-NCQ.LE.1) THEN
      WRITE(*,1678)
1678  FORMAT(//' QUEDAN MENOS DE DOS CASOS')
      GOTO 572
      ENDIF
      IF (IM.EQ.2) THEN
          XP0=XP0S/(NA-NCQ)
      ELSE
          XP0=FLOAT(IACU)/(NA-NCQ)
      ENDIF
```

```
C
C  IMPRIME RESULTADOS PARCIALES
C  ----------------------------
       OPEN (8,FILE='KAPMUL.SAL')
       CALL CABECE(8)
       WRITE(*,987)
       WRITE(8,987,ERR=878)
987    FORMAT(/' ANALISIS DE CONCORDANCIA. INDICE KAPPA')
       WRITE(*,50) NOM
       WRITE(8,50,ERR=878) NOM
50     FORMAT(//' NOMBRE DE LOS DATOS: ',A)
       WRITE(*,55)NA,NCQ,IVG,DIS(IDC+1),ICAT
       WRITE(8,55,ERR=878)NA,NCQ,IVG,DIS(IDC+1),ICAT
55     FORMAT(//' CONDICIONES DEL DISEÑO'/' Nº CASOS:',I17/' Nº CASOS'
      1,' ELIMINADOS:',I6/' Nº VARIABLES:',I13,4X,'DISEÑO ',A/
      2' Nº CATEGORIAS:'I12///,10X,'FRECUENCIAS MARGINALES')
       IWW=(ICAT+9)/10
       DO 600 IW=1,IWW
         IJ=MIN0(ICAT,IW*10)
         WRITE(*,56)(J,J=(IW-1)*10+1,IJ)
         WRITE(8,56,ERR=878)(J,J=(IW-1)*10+1,IJ)
56       FORMAT(20X,'CATEGORIA'/' CONDICION',10I7)
         DO 101 I=1,IVG
           DO 102 J=(IW-1)*10+1,IJ
102          PT(J)=PT(J)+FLOAT(IC(I,J))/FLOAT(N(I))
       WRITE(*,57)NOMCV(I),(FLOAT(IC(I,J))/FLOAT(N(I)),J=(IW-1)*10+1,IJ)
101    WRITE(8,57,ERR=878)NOMCV(I),(FLOAT(IC(I,J))/FLOAT(N(I)),J=(IW-1)
      1*10+1,IJ)
57       FORMAT(1X,A,3X,10F7.3)
         WRITE(*,59)(PT(I)/FLOAT(IVG),I=(IW-1)*10+1,IJ)
         WRITE(8,59,ERR=878)(PT(I)/FLOAT(IVG),I=(IW-1)*10+1,IJ)
59       FORMAT(/' PROMEDIO ',10F7.3)
600      CONTINUE
       IF (IGSK.EQ.1.AND.IDC.EQ.1) THEN
C
C  ESTADISTICO GSK
C  ---------------
       WRITE(*,5064)
       WRITE(8,5064,ERR=878)
5064   FORMAT(//' PRUEBA DE HOMOGENEIDAD DE LAS FRECUENCIAS MARGINALES')
       DO 4010 I=1,JPP
         DO 4010 J=1,JPP
           K=(I-1)*JPP+J
           IF (I.EQ.J) THEN
             VP(K)=DBLE(IPP(I))*DBLE(N(1)-IPP(I))/DBLE(N(1))**3.
           ELSE
             VP(K)=-DBLE(IPP(I))*DBLE(IPP(J))/DBLE(N(1))**3.
          ENDIF
4010     CONTINUE
       IFILA=IVG*(ICAT-1)
       II=IFILA*JPP
       DO 3010 I=1,II
3010     AAA(I)=0.
       JJPP=ICAT**(IVG-1)
       IFI=0
       DO 3015 I=1,IVG
         NB=ICAT**(I-1)
         NUNOS=ICAT**(IVG-I)
         LB=NUNOS*ICAT
         DO 3015 J=1,ICAT-1
           IFI=IFI+1
           DO 3015 II=1,NB
             DO 3015 JJ=1,NUNOS
               ICOL=NUNOS*(J-1)+LB*(II-1)+JJ
3015   AAA(IFILA*(ICOL-1)+IFI)=1.
       CALL DGMPRD(AAA,VP,TEM,IFILA,JPP,JPP)
       CALL DTRANS(AAA,VP,IFILA,JPP)
```

```
      CALL DGMPRD(TEM,VP,VF,IFILA,JPP,IFILA)
      IFILC=(IVG-1)*(ICAT-1)
      II=IFILA*IFILC
      DO 3020 I=1,II
3020    CON(I)=0.
      DO 3000 J=1,ICAT-1
        DO 3000 I=1,IVG-1
          CON((IFILC+IVG-1)*(J-1)+I)=1.
          CON(IFILC*(ICAT-1)*I+(IFILC+IVG-1)*(J-1)+I)=-1.
          CB(I+(IVG-1)*(J-1))=DBLE(IC(I+1,J)-IC(1,J))/DBLE(N(I))
3000  CONTINUE
      CALL DGMPRD(CON,VF,TEM,IFILC,IFILA,IFILA)
      CALL DTRANS(CON,VP,IFILC,IFILA)
      CALL DGMPRD(TEM,VP,VF,IFILC,IFILA,IFILC)
      CALL DMINV(VF,IFILC,DET,NM,NMC)
      IF (DET.NE.0.) GOTO 3050
      WRITE(*,5066)
      WRITE(8,5066,ERR=878)
5066  FORMAT(' NO SE PUEDE CALCULAR EL ESTADISTICO GSK')
      GOTO 5068
3050  CALL DGMPRD(CB,VF,TEM,1,IFILC,IFILC)
      CALL DGMPRD(TEM,CB,CON,1,IFILC,1)
      YP=PROB(1,SNGL(CON(1)),IFILC,0)
      WRITE(*,5062)CON(1),IFILC,YP
      WRITE(8,5062,ERR=878)CON(1),IFILC,YP
5062  FORMAT(' ESTADISTICO GSK:',F26.4,'   G.L.:',I4,'   p=',f8.6)
      ENDIF
5068  WRITE(*,588)MAY(IM)
      WRITE(8,588,ERR=878)MAY(IM)
588   FORMAT(//' ACUERDO POR ',A\)
      IF (IM.EQ.1) THEN
        WRITE(*,5888)IOM
        WRITE(8,7888,ERR=878)IOM
5888  FORMAT(' DE ',I3,' OBSERVADORES'\)
7888  FORMAT('+',22x,' DE ',I3,' OBSERVADORES'\)
      ENDIF
      IF (IM.EQ.3) THEN
        WRITE(*,5889) DRAND(IRAND)
        WRITE(8,7889,ERR=878) DRAND(IRAND)
5889    FORMAT(A\)
7889    FORMAT('+',22x,A\)
      ENDIF
      WRITE(*,58) XP0
      WRITE(8,58,ERR=878) XP0
58    FORMAT(//' PROPORCION DE ACUERDO OBSERVADO:',F8.4)
C
C ACUERDOS ESPERADOS
C ------------------
      IF (IDC.EQ.1) THEN
c diseño completo
        IF (IM.EQ.1) THEN
c mayoria
          CALL ESPMAY(1,XMIN,IVG,ICAT,IOM,IC,N,NN,XPE)
        ELSE
        IF (IM.EQ.3) THEN
C RAND
          CALL ESPRAN(1,XMIN,IC,N,IRAND,XPE)
        ELSE
c por pares
          XPE=0.
          AA=FLOAT(N(1))**2.
          DO 2223 J=1,IVG
            DO 2223 K=J+1,IVG
              DO 2223 JJ=1,ICAT
                DO 2223 JJJ=1,ICAT
2223        XPE=XPE+PES(JJ,JJJ)*IC(J,JJ)*IC(K,JJJ)/AA
        XPE=2.*XPE/(FLOAT(IVG)*(IVG-1.))
```

```
            ENDIF
          ENDIF
C  ACUERDOS ESPERADOS JACKKNIFE: DISEÑO COMPLETO
           DO 2500 I=1,NA
             KKK=I+MP
             DO 2202 J=1,IVG
               DO 2202 JJ=1,ICAT
2202         IB(J,JJ)=IC(J,JJ)
             CALL DATA(X,NV,NVR,2,0)
             DO 2204 J=1,IVG
               IB(J,INT(X(ICV(J))))=IC(J,INT(X(ICV(J))))-1
2204         CONTINUE
             IF (IM.EQ.2) THEN
C POR PARES
               XPEJ=0.
               AA=FLOAT(N(1)-1)**2.
               DO 3223 J=1,IVG
                 DO 3223 K=J+1,IVG
                   DO 3223 JJ=1,ICAT
                     DO 3223 JJJ=1,ICAT
3223           XPEJ=XPEJ+PES(JJ,JJJ)*IB(J,JJ)*IB(K,JJJ)/AA
               XPEJ=2.*XPEJ/(FLOAT(IVG)*(IVG-1.))
             ELSE
               DO 2256 J=1,IVG
2256           NNC(J)=N(J)-1
               IF (IM.EQ.1) THEN
C MAYORIA
                 CALL ESPMAY(0,XMIN,IVG,ICAT,IOM,IB,NNC,NN,XPEJ)
               ELSE
C RAND
                 CALL ESPRAN(0,XMIN,IB,NNC,IRAND,XPEJ)
               ENDIF
             XTT=XMIN*(NA-1)
             IF (I.EQ.1.AND.XTT.GT..5) WRITE(*,299) XTT
             ENDIF
             READ(1,REC=I)IA,IK,IR,B
             WRITE(1,REC=I)IA,IK,IR,B,XPEJ*(NA-1-NCQ)
2500       CONTINUE
       ELSE
C diseño incompleto
         XPE=0.
       DO 201 I=1,NA
         IF (I.EQ.1) CALL ITIME(IHA,MINA,ISEA,ICA)
         IF (I.EQ.2) THEN
           CALL ITIME(IHD,MIND,ISED,ICD)
           IH=IHD-IHA
           IF (IH.LT.0) IH=IH+24
           XMIN=(IH*60.+FLOAT(MIND-MINA)+(ISED+ICD/100.-
      1(ISEA+ICA/100.))/60.)*NA
           IF (XMIN.GE..5) WRITE(*,299) XMIN
299    FORMAT(/' ** PACIENCIA **'/' EL TIEMPO ESTIMADO DE CALCULO'
      1,' RESTANTE ES',F9.1,' minutos')
         ENDIF
           READ(1,REC=I)IA,IK,IR,B,C,(NN(J),J=1,IVG)
           IF (IA.EQ.1) GOTO 201
         DO 202 J=1,IVG
           DO 203 JJ=1,ICAT
203          IB(J,JJ)=IC(J,JJ)
           IF (NN(J).EQ.0) THEN
             NM(J)=N(J)
           ELSE
             NM(J)=N(J)-1
           ENDIF
202      CONTINUE
         KKK=I+MP
         CALL DATA(X,NV,NVR,2,0)
         DO 204 J=1,IVG
```

```
          IF (ABS(X(ICV(J))-XNEP).LE.EPS) GOTO 204
          IB(J,INT(X(ICV(J))))=IC(J,INT(X(ICV(J))))-1
204     CONTINUE
      IF (IM.EQ.1.AND.IDC.EQ.0) THEN
C diseño incompleto, por mayoria
          CALL ESPMAY(0,XMIN,IVG,ICAT,IOM,IC,N,NN,XP)
        XPE=XPE+XP
      ENDIF
      IF (IM.EQ.2.AND.IDC.EQ.0) THEN
C  diseño incompleto, por pares
        Q=0.
        IKK=0
        DO 223 J=1,IVG
          IF(NN(J).EQ.0) GOTO 223
          IKK=IKK+1
          IF (IKK.EQ.IK) GOTO 223
          DO 224 K=J+1,IVG
            IF (NN(K).EQ.0) GOTO 224
            AA=FLOAT(N(J))*FLOAT(N(K))
              DO 225 JJ=1,ICAT
              DO 225 JJJ=1,ICAT
225             Q=Q+PES(JJ,JJJ)*IC(J,JJ)*IC(K,JJJ)/AA
224         CONTINUE
223       CONTINUE
        XPE=XPE+2.*Q/(FLOAT(IK)*(IK-1.))
        ENDIF
C  ACUERDOS ESPERADOS JACKKNIFE: DISEÑO INCOMPLETO
        XPEJ=0.
        DO 1108 II=1,NA
          IF (I.EQ.II) GOTO 1108
            READ(1,REC=II)IAC,IKC,IRC,BC,C,(NNC(J),J=1,IVG)
          IF (IAC.EQ.1) GOTO 1108
          IF (IM.EQ.2) THEN
C por pares
          Q=0.
          IKK=0
          DO 1223 J=1,IVG
            IF(NNC(J).EQ.0) GOTO 1223
            IKK=IKK+1
            IF (IKK.EQ.IKC) GOTO 1223
            DO 1224 K=J+1,IVG
              IF (NNC(K).EQ.0) GOTO 1224
              AA=FLOAT(NM(J))*FLOAT(NM(K))
                DO 1225 JJ=1,ICAT
                DO 1225 JJJ=1,ICAT
1225              Q=Q+PES(JJ,JJJ)*IB(J,JJ)*IB(K,JJJ)/AA
1224        CONTINUE
1223      CONTINUE
          XPEJ=XPEJ+2.*Q/(FLOAT(IKC)*(IKC-1.))
          ELSE
c por mayoria
            CALL ESPMAY(0,XMIN,IVG,ICAT,IOM,IB,NM,NNC,Q)
          XPEJ=XPEJ+Q
          ENDIF
1108    CONTINUE
        WRITE(1,REC=I)IA,IK,IR,B,XPEJ,(NN(J),J=1,IVG)
201   CONTINUE
      ENDIF
      IF (IDC.EQ.0) XPE=XPE/(NA-NCQ)
6785  IF (ABS(1.-XPE).GT.EPS) GOTO 7777
      WRITE(*,9999)
9999  FORMAT(//'LA PROPORCION DE ACUERDO ESPERADO ES 1.'
     1,' NO SE PUEDE CALCULAR EL KAPPA')
      GOTO 8888
7777  XKAP=(XP0-XPE)/(1.-XPE)
      WRITE(*,62) XPE,XKAP
      WRITE(8,62,ERR=878) XPE,XKAP
```

```
62     FORMAT(' PROPORCION DE ACUERDO ESPERADO:',F9.4//' INDICE'
      1,' KAPPA:',F27.4)
         SKAC=0.
         MM=NA-NCQ
         YKA=XKAP*MM
         SKA=0.
         DO 1109 II=1,NA
           READ(1,REC=II)IA,JJ,IR,B,C
           IF (IA.EQ.1) GOTO 1109
           IF (ABS(C+1.-MM).LE.EPS) GOTO 2011
           IF (IM.EQ.2) THEN
             SP0=XP0S-B
           ELSE
             SP0=FLOAT(IACU-IR)
           ENDIF
           SXK=(SP0-C)/(MM-1.-C)
           YKAJ=YKA-(FLOAT(MM)-1.)*SXK
           SKA=SKA+YKAJ
           SKAC=SKAC+YKAJ**2.
1109     CONTINUE
         XSKA=SKA/MM
         VARKA=(SKAC-(SKA**2.)/MM)/(MM-1.)
c por errores de precision para muestras grandes, a veces negativo muy pequeño
         IF (VARKA.LE.0) VARKA=0.
         EEKA=SQRT(VARKA/MM)
       CALL PROBI(2,0.05,MM-1,0,T,IE)
       WRITE(*,661)XSKA,SQRT(VARKA),XSKA-T*SQRT(VARKA)
      1,XSKA+T*SQRT(VARKA)
       WRITE(8,661,ERR=878)XSKA,SQRT(VARKA),XSKA-T*SQRT(VARKA)
      1,XSKA+T*SQRT(VARKA)
661    FORMAT(' ESTIMADOR "JACKKNIFE":',F18.4,3X,'ERROR ESTANDAR:',F9.4
      1/' INTERVALO DE CONFIANZA AL 95%:',2X,2F8.4)
2011   IF (IP.EQ.83) THEN
       WRITE(*,662)
       WRITE(8,662,ERR=878)
662    FORMAT(//' MATRIZ DE PESOS USADA')
       DO 660 IW=1,IWW
         IJ=MIN0(ICAT,IW*10)
         WRITE(*,66)(J,J=(IW-1)*10+1,IJ)
         WRITE(8,66,ERR=878)(J,J=(IW-1)*10+1,IJ)
66     FORMAT(/7X,10I7)
       DO 69 I=1,ICAT
       WRITE(*,67)I,(PES(I,J),J=(IW-1)*10+1,IJ)
69     WRITE(8,67,ERR=878)I,(PES(I,J),J=(IW-1)*10+1,IJ)
67     FORMAT(I7,10F7.2)
660    CONTINUE
       ENDIF
8888   CLOSE(1)
       CLOSE(2)
       CLOSE(8)
         IF (IDELET(NOMBRE).NE.0) STOP 6
       WRITE(*,'(//\'' QUIERES COPIA EN PAPEL (S/N) [S] : '')')
       IQ=RUM(66,83.,0,0.,0.)
       IF (IQ.GT.90)IQ=IQ-32
       IF (IQ.EQ.83) CALL COPIA('KAPMUL.sal')
       GOTO 551
1100   WRITE(*,989)I,NOMCV(ICV(J))
989    FORMAT(/' EN EL CASO',I5,' LA VARIABLE ',A,/
      1' NO ES ENTERA O ESTA FUERA DE RANGO')
572    CLOSE(2)
       GOTO 551
878    stop 10
991    stop 9
444    continue
       end

$STORAGE:2
```

```
      SUBROUTINE DTRANS(A,B,NF,NC)
C  TRANSPONE UNA MATRIZ
C  VERSION DOBLE PRECISION
C
C  A  MATRIZ DE ENTRADA
C  B  MATRIZ DE SALIDA TRANSPUESTA
C  NF NUMERO DE FILAS DE A Y COLUMNAS DE B
C  NC   NUMERO DE COLUMNAS DE A Y FILAS DE B
      DOUBLE PRECISION A(1),B(1)
      DO 10 I=1,NC
        DO 10 J=1,NF
10      B((J-1)*NC+I)=A(NF*(I-1)+J)
      RETURN
      END

$storage:2
      SUBROUTINE RAND(X,ICV,IRAND,IR,IER)
C  CALCULA ACUERDOS RAND EN EL ARRAY X
C  X         DATOS
C  ICV       VARIABLES A USAR
C  IRAND     DEFINICION DE ACUERDO 1=A9S, 2=A9R, 3=A7S, 4=A7R, 5=AE
C  IR        0 NO SE CUMPLE, 1 SI
C  IER       0 BIEN, 1 FALTAN OBSERVACIONES
C
C V. ABRAIRA OCT 94
C
      DIMENSION X(1),ICV(1),B(9),IO(9),LI(3),LS(3)
      COMMON ICOD,KKK,XNEP
      DATA EPS/.00001/LI/1,4,7/LS/3,6,9/
      IER=0
      IR=0
      IME=9
      IMA=1
      DO 10 J=1,9
        IF (ABS(X(ICV(J))-XNEP).GT.EPS) GOTO 20
        IER=1
        RETURN
20      II=INT(X(ICV(J)))
        IF (II.GT.IMA) THEN
          IMA=II
          JMA=J
        ENDIF
        IF (II.LT.IME) THEN
          IME=II
          JME=J
        ENDIF
10    CONTINUE
      GOTO (100,200,300,300,500) IRAND
100   IF ((IME.GE.1.AND.IMA.LE.3).OR.(IME.GE.4.AND.IMA.LE.6)
     1.OR.(IME.GE.7.AND.IMA.LE.9)) IR=1
      RETURN
200   IF (IMA-IME.LE.2) IR=1
      RETURN
300   IME=9
      IMA=1
      DO 310 J=1,9
        IF (J.EQ.JMA.OR.J.EQ.JME) GOTO 310
        II=INT(X(ICV(J)))
        IF (II.GT.IMA) THEN
          IMA=II
        ENDIF
        IF (II.LT.IME) THEN
          IME=II
        ENDIF
310   CONTINUE
      GOTO (100,200,999) IRAND-2
500   DO 600 I=1,9
```

```
600       B(I)=X(ICV(I))
        CALL ORDENA(9,B,IO)
        MED=INT(B(IO(5)))
        IMED=(MED-1)/3+1
        IFU=0
        DO 700 I=1,9
          II=INT(X(ICV(I)))
          IF (II.GE.LI(IMED).AND.II.LE.LS(IMED)) GOTO 700
          IFU=IFU+1
          IF (IFU.GE.3) GOTO 800
700     CONTINUE
        IR=1
800     RETURN
999     STOP 'ERROR IMPOSIBLE'
        END

$storage:2
        SUBROUTINE ORDENA(N,X,IO)
C
C  DEJA EN IO LOS INDICES DE LOS N DATOS X EN ORDEN CRECIENTE
C
        DIMENSION X(1),IO(1)
        DO 1 I=1,N
1       IO(I)=I
        IF (N.LE.1) RETURN
        DO 2 J=2,N
        I=J
3       IF (X(IO(I-1)).LE.X(IO(I))) GOTO 2
        IA=IO(I)
        IO(I)=IO(I-1)
        IO(I-1)=IA
        I=I-1
        IF (I.GT.1) GOTO 3
2       CONTINUE
        RETURN
        END

$storage:2
        FUNCTION IREPET(J,L,IVAR)
C  MIRA SI EN EL ARRAY IVAR, DE J ELEMENTOS HAY AL MENOS L REPETIDOS
C
C  DEVUELVE en irepet
C  0 NO LAS HAY
C  1 SI LAS HAY
        DIMENSION IVAR(1)
        DO 10 JJ=1,J
          ICON=1
          DO 10 I=1,J
            IF (I.EQ.JJ) GOTO 10
            IF (IVAR(JJ).NE.IVAR(I)) GOTO 10
            ICON=ICON+1
            IF (ICON.GE.L) GOTO 20
10      CONTINUE
        IREPET=0
        RETURN
20      IREPET=1
        RETURN
        END

$STORAGE:2
        SUBROUTINE ESPMAY(IT,XMIN,J,K,L,IC,N,NN,XPE)
C  CALCULA Y DEVUELVE EN XPE LA PROBABILIDAD DE ACUERDOS ESPERADOS POR MAYORIA
C  IT        0 NO IMPRIME TIEMPO Y LO DEVUELVE EN XMIN, 1 LO IMPRIME
C  J         NUMERO DE OBSERVADORES
C  K         CATEGORIAS
C  L         MAYORIA
C  IC        FRECUENCIAS ABSOLUTAS (OBS,CAT)
```

```
C  N(OBS)    NUMERO DE OBSERVACIONES
C  NN(OBS)   NO CONSIDERA LOS OBSERVADORES EN QUE NN()=0
C  XPE       PROPORCION DE ACUERDOS ESPERADOS
c
      DIMENSION IC(100,30),N(1),NN(1),IVAR(100)
      double precision fa,fb,fmax
      XPE=0.
      IK=0
      DO 100 i=1,J
100     IK=IK+NN(I)
      IF (L.GT.IK) RETURN
      fmax=float(k)**IK-1.
        fa=0.
        CALL ITIME(IHA,MINA,ISEA,ICA)
10      fb=fa
        IF (DABS(FA-1000.).LE..00001) THEN
          CALL ITIME(IHD,MIND,ISED,ICD)
          IH=IHD-IHA
          IF (IH.LT.0) IH=IH+24
          XMIN=(IH*60.+FLOAT(MIND-MINA)+(ISED+ICD/100.-
     1(ISEA+ICA/100.))/60.)*SNGL(fmax/1000.)
          IF (IT.EQ.1.AND.XMIN.GT..5) WRIte(*,199) xmin
199   FORMAT(/' ** PACIENCIA **'/' EL TIEMPO ESTIMADO DE CALCULO DE'
     1,' ACUERDOS ESPERADOS ES',F9.1,' MINUTOS')
        ENDIF
      do 20 jj=IK-1,0,-1
        ivar(IK-jj)=int(fb/float(k)**jj)+1
          fb=dmod(fb,float(k)**jj)
20      continue
      IF (IREPET(IK,L,IVAR).EQ.0) GOTO 300
      DO 200 JJ=1,J-1
        IF (NN(JJ).EQ.1) GOTO 200
        DO 250 LL=J-1,JJ,-1
250       IVAR(LL+1)=IVAR(LL)
200   CONTINUE
      XP=1.
      DO 50 JJ=1,J
        IF (NN(JJ).EQ.0) GOTO 50
        XP=XP*FLOAT(IC(JJ,IVAR(JJ)))/N(JJ)
50      CONTINUE
      XPE=XPE+XP
300   fa=fa+1.
      if (fa.le.fmax) goto 10
      RETURN
      END

$STORAGE:2
        SUBROUTINE ESPRAN(IT,XMIN,IC,N,IRAND,XPE)
C  CALCULA Y DEVUELVE EN XPE LA PROBABILIDAD DE ACUERDOS RAND ESPERADOS
C  IT        0 NO IMPRIME TIEMPO Y LO DEVUELVE EN XMIN, 1 LO IMPRIME
C  IC        FRECUENCIAS ABSOLUTAS (OBS,CAT)
C  N(OBS)    NUMERO DE OBSERVACIONES
C  IRAND     ACUERDO
C  XPE       PROPORCION DE ACUERDOS ESPERADOS
c
      DIMENSION IC(100,30),N(1),VAR(9),icv(9),IVAR(9)
      CHARACTER NOM*9(5),NOMA*16
      DOUBLE PRECISION FMAX,FA,FB,FC
      DATA NOM/'A9S','A9R','A7S','A7R','AE'/
      DO 5 I=1,9
5     ICV(I)=I
      XPE=0.
      KK=MIRA(NOM(IRAND),4,NOMA,0,0)
      OPEN (7,FILE=NOMA,ACCESS='SEQUENTIAL',RECL=18
     1,FORM='UNFORMATTED')
      IF (KK.EQ.1) THEN
        call ITIME(IHA,MINA,ISEA,ICA)
```

```
260      READ(7,END=500)(IVAR(J),J=1,9)
         XP=1.
         DO 250 JJ=1,9
           XP=XP*FLOAT(IC(JJ,IVAR(JJ)))/N(JJ)
250      CONTINUE
         XPE=XPE+XP
         GOTO 260
500      CONTINUE
         CALL ITIME(IHD,MIND,ISED,ICD)
         IH=IHD-IHA
         IF (IH.LT.0) IH=IH+24
         XMIN=IH*60.+FLOAT(MIND-MINA)+(ISED+ICD/100.-
     1(ISEA+ICA/100.))/60.
       ELSE
         WRITE(*,100)
100      format(/' No existe archivo con las permutaciones.'
     1,' Lo voy a crear')
         fmax=9.**9-1.
         fa=0.
         CALL ITIME(IHA,MINA,ISEA,ICA)
10       fb=fa
         IF (DABS(FA-1000.).LE..00001) THEN
           CALL ITIME(IHD,MIND,ISED,ICD)
           IH=IHD-IHA
           IF (IH.LT.0) IH=IH+24
           XMIN=(IH*60.+FLOAT(MIND-MINA)+(ISED+ICD/100.-
     1(ISEA+ICA/100.))/60.)*SNGL(fmax/1000.)
           IF (IT.EQ.1.AND.XMIN.GT..5) write(*,199) xmin
199    FORMAT(/' ** PACIENCIA **'/' EL TIEMPO ESTIMADO DE CALCULO DE'
     1,' ACUERDOS ESPERADOS ES',F9.1,' MINUTOS')
         ENDIF
         do 20 jj=8,0,-1
           FC=9.**JJ
           var(9-JJ)=int(fb/FC)+1.
           fb=Dmod(fb,FC)
20       continue
         CALL RAND(VAR,ICV,IRAND,IR,IER)
         IF (IR.EQ.0) GOTO 300
         WRITE(7)(INT(VAR(J)),J=1,9)
         XP=1.
         DO 50 JJ=1,9
           XP=XP*FLOAT(IC(JJ,INT(VAR(JJ))))/N(JJ)
50       CONTINUE
         XPE=XPE+XP
300      fa=fa+1.
         if (fa.le.fmax) goto 10
       ENDIF
       CLOSE(7)
       RETURN
       END
```

APÉNDICE III

Listado de los datos del estudio del SAT (para su codificación véase texto, el signo ? corresponde a datos no especificados, en este caso, a datos no tomados)

1) Contracturas articulares

```
P R E S T A  PC   V2.2                                13-JUN-1996

LISTADO DE DATOS

NOMBRE DEL ARCHIVO: contrac
NUMERO DE VARIABLES:     7        CON FORMATO Entero

NUMERO DE CASOS DECLARADOS:    10        INTRODUCIDOS:    10

CODIGO DE NO ESPECIFICADO:     -1.00000

LOS NOMBRES DE LAS VARIABLES SON:
  1: MED2     2: MED4     3: MED5     4: MED6     5: MED7
  6: MED8

CASO NO.    1
     1      ?      ?      2      ?      2
CASO NO.    2
     ?      ?      2      2      2      ?
CASO NO.    3
     ?      ?      1      2      ?      2
CASO NO.    4
     1      ?      2      ?      2      ?
CASO NO.    5
     1      ?      ?      ?      1      1
CASO NO.    6
     1      2      1      ?      ?      ?
CASO NO.    7
     ?      2      2      ?      ?      2
CASO NO.    8
     ?      1      ?      ?      1      1
CASO NO.    9
     1      1      ?      1      ?      ?
CASO NO.   10
     ?      2      ?      1      1      ?
```

2) Neuropatía periférica

```
P R E S T A  PC   V2.2                                13-JUN-1996

LISTADO DE DATOS

NOMBRE DEL ARCHIVO: neuro
NUMERO DE VARIABLES:     7        CON FORMATO Entero
```

```
NUMERO DE CASOS DECLARADOS:     10        INTRODUCIDOS:     10

CODIGO DE NO ESPECIFICADO:      -1.00000

LOS NOMBRES DE LAS VARIABLES SON
  1: MED2      2: MED4      3: MED5      4: MED6      5: MED7
  6: MED8


CASO NO.     1
      1      ?      ?      1      ?      1
CASO NO.     2
      ?      ?      1      1      1      ?
CASO NO.     3
      ?      ?      2      2      ?      2
CASO NO.     4
      1      ?      1      ?      1      ?
CASO NO.     5
      2      ?      ?      ?      3      2
CASO NO.     6
      3      2      3      ?      ?      ?
CASO NO.     7
      ?      1      1      ?      ?      1
CASO NO.     8
      ?      2      ?      ?      3      3
CASO NO.     9
      3      1      ?      3      ?      ?
CASO NO.    10
      ?      2      ?      3      2      ?
```

3) Cambios esclerodermiformes en la piel

```
P R E S T A  PC   V2.2                              13-JUN-1996

LISTADO DE DATOS

NOMBRE DEL ARCHIVO: esclero
NUMERO DE VARIABLES:      7        CON FORMATO Entero

NUMERO DE CASOS DECLARADOS:     10        INTRODUCIDOS:     10

CODIGO DE NO ESPECIFICADO:      -1.00000

LOS NOMBRES DE LAS VARIABLES SON
  1: MED2      2: MED4      3: MED5      4: MED6      5: MED7
  6: MED8


CASO NO.     1
      4      ?      ?      4      ?      4
CASO NO.     2
      ?      ?      4      4      2      ?
CASO NO.     3
      ?      ?      3      3      ?      3
CASO NO.     4
      3      ?      3      ?      2      ?
```

```
CASO NO.    5
     1      ?      ?      ?      1      1
CASO NO.    6
     1      1      1      ?      ?      ?
CASO NO.    7
     ?      3      4      ?      ?      3
CASO NO.    8
     ?      1      ?      ?      1      1
CASO NO.    9
     2      2      ?      3      ?      ?
CASO NO.   10
     ?      1      ?      1      2      ?
```

APÉNDICE IV

Listado de los datos del estudio del *nevus flammeus*.

1) Color: Los médicos B, C y D son los que habitualmente tratan a estos pacientes; los A, E y F son otros dermátologos del mismo Servicio. Los colores están codificados como 1: Rosa; 2: Rojo y 3: Violeta.

```
 P R E S T A  PC   V2.2                              6-AGO-1996

 LISTADO DE DATOS

 NOMBRE DEL ARCHIVO: color
 NUMERO DE VARIABLES:     8        CON FORMATO Entero

 NUMERO DE CASOS DECLARADOS:    80        INTRODUCIDOS:    80

 CODIGO DE NO ESPECIFICADO:     -1.00000

 LOS NOMBRES DE LAS VARIABLES SON
   1: A         2: B         3: C         4: D         5: E         6: F


 CASO NO.     1
       1      1      1      1      1      1
 CASO NO.     2
       3      3      3      3      3      3
 CASO NO.     3
       2      2      2      2      2      2
 CASO NO.     4
       1      1      1      1      1      1
 CASO NO.     5
       3      3      3      3      3      3
 CASO NO.     6
       2      2      2      2      2      2
 CASO NO.     7
       1      1      1      1      1      1
 CASO NO.     8
       2      2      2      2      2      2
 CASO NO.     9
       2      2      2      1      2      2
 CASO NO.    10
       2      2      2      2      2      3
 CASO NO.    11
       2      2      2      2      2      2
 CASO NO.    12
       2      2      2      3      3      3
 CASO NO.    13
       2      2      2      2      2      2
 CASO NO.    14
       2      2      2      1      1      2
 CASO NO.    15
```

```
      2      2      2      2      2      2
CASO NO.    16
      1      1      1      1      1      1
CASO NO.    17
      2      2      2      3      2      2
CASO NO.    18
      2      1      2      1      1      2
CASO NO.    19
      3      2      3      2      3      2
CASO NO.    20
      2      2      2      2      2      2
CASO NO.    21
      3      3      3      3      3      3
CASO NO.    22
      1      1      1      1      1      1
CASO NO.    23
      2      2      2      2      2      2
CASO NO.    24
      1      2      2      1      1      1
CASO NO.    25
      2      2      2      2      2      2
CASO NO.    26
      1      1      1      1      1      2
CASO NO.    27
      2      3      2      2      3      3
CASO NO.    28
      2      2      2      2      2      2
CASO NO.    29
      1      1      1      1      1      1
CASO NO.    30
      2      1      2      1      2      2
CASO NO.    31
      2      2      2      2      2      2
CASO NO.    32
      2      2      2      2      2      2
CASO NO.    33
      2      2      2      2      2      2
CASO NO.    34
      2      2      2      2      2      2
CASO NO.    35
      2      2      2      1      1      2
CASO NO.    36
      2      2      2      2      2      2
CASO NO.    37
      3      3      3      2      3      2
CASO NO.    38
      2      1      1      1      1      1
CASO NO.    39
      2      1      1      1      1      2
CASO NO.    40
      1      1      1      1      1      1
CASO NO.    41
      2      1      1      1      2      1
CASO NO.    42
      2      2      2      1      2      2
CASO NO.    43
      1      1      1      1      1      2
CASO NO.    44
      3      3      3      3      3      3
```

CASO NO.	45				
1	1	1	1	1	1
CASO NO.	46				
2	2	2	2	2	2
CASO NO.	47				
2	2	2	2	2	2
CASO NO.	48				
3	3	2	3	3	3
CASO NO.	49				
1	1	1	2	1	1
CASO NO.	50				
2	3	3	3	3	3
CASO NO.	51				
3	3	3	2	3	3
CASO NO.	52				
2	2	2	2	2	3
CASO NO.	53				
2	2	2	2	2	2
CASO NO.	54				
2	2	3	2	2	3
CASO NO.	55				
2	2	2	2	2	3
CASO NO.	56				
3	3	3	3	3	3
CASO NO.	57				
1	1	1	1	1	1
CASO NO.	58				
2	1	2	2	1	1
CASO NO.	59				
2	2	2	2	2	2
CASO NO.	60				
2	2	2	2	2	3
CASO NO.	61				
2	2	2	3	2	2
CASO NO.	62				
1	1	1	1	1	1
CASO NO.	63				
2	1	1	1	2	2
CASO NO.	64				
2	1	1	1	1	2
CASO NO.	65				
3	3	3	2	3	2
CASO NO.	66				
2	2	2	2	2	2
CASO NO.	67				
3	3	3	3	3	3
CASO NO.	68				
2	2	2	2	1	2
CASO NO.	69				
2	2	2	3	2	2
CASO NO.	70				
3	3	3	3	3	3
CASO NO.	71				
2	3	3	2	3	3
CASO NO.	72				
2	2	2	3	2	2
CASO NO.	73				
2	2	2	2	3	3
CASO NO.	74				
2	2	2	3	2	3

```
 CASO NO.    75
       3      3      3      3      3      3
CASO NO.    76
       2      2      2      2      2      2
CASO NO.    77
       2      2      2      1      1      1
CASO NO.    78
       3      3      2      3      3      3
CASO NO.    79
       1      1      2      1      1      1
CASO NO.    80
       2      2      2      2      2      2
```

2) Aclaramiento: Sólo medido por los médicos B, C y D; en las variables de ese nombre figuran el aclaramiento en la escala original de 0 a 100 y en las variables BC, CC y DC el mismo aclaramiento después de categorizar

```
 P R E S T A  PC   V2.2                                  6-AGO-1996

 LISTADO DE DATOS

 NOMBRE DEL ARCHIVO: aclar
 NUMERO DE VARIABLES:      6        CON FORMATO Real

 NUMERO DE CASOS DECLARADOS:     80        INTRODUCIDOS:     80

 CODIGO DE NO ESPECIFICADO:      -1.00000

 LOS NOMBRES DE LAS VARIABLES SON
   1: B        2: C        3: D        4: BC       5: CC       6: DC


 CASO NO.     1
      90      4     90      4     90      4
CASO NO.     2
      70      3     70      3     75      4
CASO NO.     3
      60      3     75      4     50      3
CASO NO.     4
      40      2     30      2     25      2
CASO NO.     5
      80      4     70      3     90      4
CASO NO.     6
      50      3     30      2     30      2
CASO NO.     7
      80      4     85      4     80      4
CASO NO.     8
      40      2     35      2     60      3
CASO NO.     9
      70      3     50      3     70      3
CASO NO.    10
      40      2     20      1     50      3
CASO NO.    11
      85      4     80      4     90      4
CASO NO.    12
      85      4     80      4     80      4
```

CASO NO.	13				
30	2	50	3	30	2
CASO NO.	14				
50	3	60	3	50	3
CASO NO.	15				
60	3	40	2	50	3
CASO NO.	16				
50	3	50	3	40	2
CASO NO.	17				
85	4	85	4	80	4
CASO NO.	18				
100	4	100	4	100	4
CASO NO.	19				
70	3	60	3	90	4
CASO NO.	20				
50	3	50	3	45	2
CASO NO.	21				
100	4	95	4	95	4
CASO NO.	22				
40	2	60	3	60	3
CASO NO.	23				
100	4	90	4	95	4
CASO NO.	24				
90	4	85	4	90	4
CASO NO.	25				
70	3	70	3	70	3
CASO NO.	26				
65	3	65	3	60	3
CASO NO.	27				
75	4	70	3	80	4
CASO NO.	28				
40	2	45	2	50	3
CASO NO.	29				
100	4	95	4	100	4
CASO NO.	30				
100	4	95	4	95	4
CASO NO.	31				
90	4	80	4	90	4
CASO NO.	32				
95	4	85	4	80	4
CASO NO.	33				
100	4	95	4	100	4
CASO NO.	34				
70	3	70	3	60	3
CASO NO.	35				
100	4	100	4	100	4
CASO NO.	36				
50	3	55	3	50	3
CASO NO.	37				
80	4	85	4	85	4
CASO NO.	38				
85	4	60	3	90	4
CASO NO.	39				
100	4	100	4	100	4
CASO NO.	40				
70	3	60	3	70	3
CASO NO.	41				
15	1	20	1	30	2
CASO NO.	42				
40	2	60	3	30	2

CASO NO.	43				
70	3	70	3	40	2
CASO NO.	44				
40	2	50	3	30	2
CASO NO.	45				
85	4	70	3	70	3
CASO NO.	46				
75	4	50	3	70	3
CASO NO.	47				
20	1	20	1	30	2
CASO NO.	48				
75	4	65	3	60	3
CASO NO.	49				
15	1	40	2	40	2
CASO NO.	50				
70	3	70	3	80	4
CASO NO.	51				
50	3	65	3	50	3
CASO NO.	52				
40	2	60	3	70	3
CASO NO.	53				
80	4	95	4	90	4
CASO NO.	54				
75	4	80	4	85	4
CASO NO.	55				
75	4	55	3	80	4
CASO NO.	56				
60	3	50	3	40	2
CASO NO.	57				
75	4	85	4	80	4
CASO NO.	58				
50	3	30	2	50	3
CASO NO.	59				
70	3	50	3	40	2
CASO NO.	60				
50	3	50	3	40	2
CASO NO.	61				
100	4	100	4	100	4
CASO NO.	62				
40	2	40	2	50	3
CASO NO.	63				
30	2	25	2	20	1
CASO NO.	64				
60	3	50	3	50	3
CASO NO.	65				
50	3	45	2	60	3
CASO NO.	66				
40	2	50	3	40	2
CASO NO.	67				
10	1	10	1	15	1
CASO NO.	68				
70	3	75	4	80	4
CASO NO.	69				
75	4	70	3	80	4
CASO NO.	70				
90	4	90	4	90	4
CASO NO.	71				
80	4	75	4	75	4
CASO NO.	72				
60	3	45	2	45	2

```
 CASO NO.    73
      30      2     30      2     40      2
CASO NO.    74
      70      3     70      3     75      4
CASO NO.    75
      80      4     85      4     90      4
CASO NO.    76
      90      4     90      4     95      4
CASO NO.    77
     100      4     95      4    100      4
CASO NO.    78
     100      4     90      4     90      4
CASO NO.    79
      90      4     90      4     80      4
CASO NO.    80
      40      2     35      2     20      1
```

APÉNDICE V

Listado de 18 "circunstancias clínicas" en las que se valora el uso de la ACTP.

CAPÍTULO 1

ASINTOMÁTICOS

A. CON PRUEBA DE ESFUERZO POSITIVA

1. Enfermedad de tronco común izquierdo
   FEVI[6] a) >50%
   b) >30% y ≤ 50%
   c) ≥ 20% y ≤ 30%

2. Enfermedad de tres vasos
   FEVI a) >50%
   b) >30% y ≤ 50%
   c) ≥ 20% y ≤ 30%

3. Enfermedad de dos vasos con afectación de la descendente anterior proximal
   FEVI a) >50%
   b) >30% y ≤ 50%
   c) ≥ 20% y ≤ 30%

4. Enfermedad de dos vasos sin afectación de la descendente anterior proximal
   FEVI a) >50%
   b) >30% y ≤ 50%
   c) ≥ 20% y ≤ 30%

5. Enfermedad de un vaso (descendente anterior proximal)
   FEVI a) >50%
   b) >30% y ≤ 50%
   c) ≥ 20% y ≤ 30%

6. Enfermedad de un vaso (culaquiera que no sea descendente anterior proximal)
   FEVI a) >50%
   b) >30% y ≤ 50%
   c) ≥ 20% y ≤ 30%

[6] FEVI son las siglas de la prueba funcional cardíaca: Fracción de Eyección del Ventrículo Izquierdo.

REFERENCIAS BIBLIOGRAFICAS

[1] Feinstein A.R. *Clinimetrics.* New Haven: Yale University Press; 1987.

[2] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**: 37-46.

[3] Abraira V. y Zaplana J. PRESTA, un paquete de procesamientos estadísticos. *Proceeding de la Conferencia Iberoamericana de Bioingeniería.* Gijón; 1984:100.

[4] Fisher R.A. *Statistical methods for research workers.* New York: Hafner; 1958.

[5] Bartko J.J. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966; **19**: 3-11.

[6] Latour J., Abraira V., Cabello J. y López Sánchez J. Las mediciones clínicas en cardiología: Validez y errores de precisión. *Rev Esp Cardiol* (en prensa).

[7] Rosner B. Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics* 1982; **38**: 105-114.

[8] Donner A. y Donald A. The statistical analysis of multiple binary measurements. *J Clin Epimediol* 1988; **41**: 899-905.

[9] Fleiss J.L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* 1975; **31**: 651-659.

[10] Landis J.R. y Koch G.G. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159-174.

[11] Fleiss J.L. *Statistical methods for rates and proportions.* New York: John Wiley & Sons; 1981.

[12] Elmore J.G., Wells C.K., Lee C.H., Howard D.H. y Feinstein A.R. Variability in radiologist's interpretations of mammograms. *N Engl J Med* 1994; **331**: 1493-1499.

[13] Jelles F., Van Bennekom C.A.M., Lankhorst G.F., Sibbel C.J.P. y Bouter L.M. Inter- and intra-rater agreement of the rehabilitation activities profile. *J Clin Epidemiol* 1995; **48**: 407-416.

[14] Feinstein A.R. y Cicchetti D.V. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; **43**: 543-549.

[15] Cicchetti D.V. y Feinstein A.R. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990; **43**: 551-558.

[16] Guggenmoos-Holzmann I. How reliable are chance-corrected measures of aggreement?. *Stat Med* 1993; **12**: 2191-2205.

[17] Brennan P. y Silman A. Statistical methods for assessing observer variability in clinical measures. *Brit Med J* 1992; **304**: 1491-1494.

[18] Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or parcial credit. *Psychol Bull* 1968; **70**: 213-220.

[19] Maclure M. y Willet W.C. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; **126**: 161-169.

[20] Fleiss J.L. y Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973; **33**: 613-619.

[21] Graham P. y Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. *J Clin Epidemiol* 1993; **46**: 1055-1062.

[22] Dunn G. *Design and analysis of reliability studies. The statistical evaluation of measurement errors.* New York: Oxford University Press; 1989.

[23] Landis J.R. y Koch G.G. A one-way components of variance model for categorical data. *Biometrics* 1977; **33**: 159-174.

[24] Fleiss J.L. y Cuzick J. The reliability of dichotomous judgments: Unequal numbres of judges per subject. *Appl Psychol Meas* 1979; **3**: 537-542.

[25] Fleiss J.L., Nee J.C.M. y Landis J.R. The large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979; **86**: 974-977.

[26] Landis J.R. y Koch G.G. An application of hierarchical kappa-type statistics in the assesment of majority agreement among multiple observers. *Biometrics* 1977; **33**: 363-374.

[27] Davies M. y Fleiss J.L. Measuring agreement for multinomial data. *Biometrics* 1982; **38**: 1047-1051.

[28] Schouten H.J.A. Nominal scale agreement among observers. *Psychometrika* 1986; **51**: 453-466.

[29] Gross S.T. The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* 1986; **42**: 883-893.

[30] Koch G.G., Imrey P.B. y Reinfurt D.W. Linear model analysis of categorical data with incomplete response vectors. *Biometrics* 1972; **28**: 663-692.

[31] Lázaro P. y Fitch K. From universalism to selectivity: is 'appropriateness'" the answer?. *Health Policy* 1996; **36**: 261-272.

[32] Brook R.H., Chassin M.R., Fink A., Salomon D.H., Kosecoff J. y Park R.E. A method for the detailed assessment of the appropiatness of medical technologies. *Int J Technol Ass Health Care* 1986; **2**: 53-68.

[33] Hubert L. Kappa revisited. *Psychol Bull* 1977; **84**: 289-297.

[34] Fleiss J.L., Cohen J., y Everitt B.S. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969; **72**: 323-327.

[35] Quenouille M.H. Notes on bias in estimation. *Biometrika* 1956; **43**: 353-360.

[36] Tukey J.W. Bias and confidence in not-quitte large samples. *Ann Math Stat* 1958; **29**:614.

[37] Kraemer H.C. Extension of the kappa coefficient. *Biometrics* 1980; **36**: 207-216.

[38] Fleiss J.L. y Davies M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *Am J Epidemiol* 1982; **115**: 841-845.

[39] Efron B. y Gong G. A leisurely look at the boostrap, the jackknife and cross-validation. *Am Stat* 1983; **37**: 36-48.

[40] Arvesen J.N. Jackkifing U-statistics. *Ann Math Stat* 1969; **40**: 2076-2100.

[41] Arvesen J.N. y Schmitz T.H. Robust procedures for variance component problems using the jackknife. *Biometrics* 1970; **26**: 677-686.

[42] Parr W.C. y Tolley H.D. Jackknifing in categorical data analysis. *Aust J Stat* 1982; **24**: 67-79.

[43] McNemar Q. Note on the sampling error of the difference between correlated proportions or porcentages. *Psychometrika* 1947; **12**: 153-157.

[44] Fleiss J.L. y Everitt B.S. Comparing the marginal totals of square contingency tables. *Brit J Math Stat Psy* 1971; **24**: 117-123.

[45] Cochran W.G. The comparison of porcentages in matched samples. *Biometrika* 1950; **37**: 256-266.

[46] Fleiss J.L. A note on Cochran´s Q test. *Biometrics* 1965; **21**: 1008-1010.

[47] Koch G.G., Landis J.R., Freeman J.L., Freeman Jr. D.H. y Lehnen R.G. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 1977; **33**: 133-158.

[48] Grizzle J.E., Starmer C.F. y Koch G.G. Analysis of categorical data by linear models. *Biometrics* 1969; **25**: 489-504.

[49] Grandjean P. y Tarkowski S. *Toxic oil syndrome. Mass food poisoning in Spain.* Copenhagen: World Health Organization. Regional Office for Europe; 1983.

[50] Nadal J. y Tarkowski S. *Toxic oil syndrome. Current knowledge and future perspectives.* Copenhagen: World Health Organization. Regional Publications European Series No. 42; 1992.

[51] Fleiss J.L. *The design and analysis of clinical experiments.* New York: John Wiley & Sons; 1986.

[52] Enjolras O. y Mulliken J.B. The current management of vascular birthmarks. *Pediatr Dermatol* 1993; **10**: 311-333.

[53] Geronemus R.G. Pulsed dye laser treatment of vascular lesions in children. *J Dermatol Surg Onc* 1993; **19**: 303-310.

[54] Noe J.M., Barsky S.H., Geer D.E. y Rosen S. Port wine stains and the response to argon laser therapy: succesful treatment and the predictive role of color, age and biopsy. *Plast Reconstr Surg* 1980; **65**: 130-136.

[55] Pérez B., Abraira V., Nuñez M., Boixeda P., Pérez Corral F. y Ledo A. Evaluation of agreement between dermatologists in the assessment of the color of Nevus Flammeus and its clearance after treatment with the Flaslamp-pumped Dye Laser. *Dermatology* (en prensa).

[56] Lázaro P., Abraira V., Gómez de la Cámara A., Kahan. J.P. y Rodríguez Artalejo F. Uso apropiado de angioplastia transluminal percutánea (ACTP) en España. Proyecto FIS 95/1956.

[57] Cicchetti D.V. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. *Appl Psychol Meas* 1981; **5**: 101-104.

[58] System/360 Scientific Subroutine Package. Version III. IBM. 1970.