Universidad de Murcia

Facultad de Psicología

Departamento de Psicología Básica y Metodología

TESIS DOCTORAL

MEDIDAS DE ACUERDO Y DE SESGO ENTRE JUECES

Autora: Ana Pilar Benavente Reche

Director: Dr. Manuel Ato García

Murcia, Julio de 2009

Agradecimientos

Recuerdo aún con cariño y una sonrisa el día 28 de Diciembre del 2001, fecha en la que me enteré que era beneficiaria de una beca predoctoral FPU. Fue un momento único, una mezcla de sentimientos solo comparable a la que disfruto en este momento al finalizar la redacción de este de proyecto de tesis doctoral, fruto de mucho esfuerzo, trabajo y dedicación de un magnífico equipo humano con el que he tenido la suerte de participar y aprender , tanto en el plano personal como profesional, en pocas palabras " ha sido y es una experiencia inefable e inolvidable". Unos meses atrás veía tan lejano este momento..., y conforme se acercaba pensaba entre otras muchas cosas escribir con antelación los agradecimientos. Quien bien me conoce sabe que, pese a mi "apariencia alocada", me gusta planificar las cosas, pero a veces llego justa de tiempo, por las "sempiternas leyes de Murphy" y otras que en un ataque de sinceridad, he de confesar son fruto de mi forma de ser, "algo nerviosilla". Nunca un año se presentó con tantas pruebas y obstáculos, con seguridad puedo decir que los aprendizajes obtenidos en este proceso marcarán mi camino de hoy en adelante.

Hoy hago memoria y deseo hacer mención a todos los que que ha estado y estarán conmigo, que tanto quiero y me han ayudado. Es justo por ello brindarles un pequeño homenaje dedicándoles unas palabras.

En primer lugar debo y quiero mencionar a **Manuel Ato García**. Por él siento desde hace años una profunda admiración, y esa admiración ha ido en aumento conforme he tenido la oportunidad de tratarlo a nivel personal y profesional. No he conocido a nadie que muestre un entusiasmo similar en su trabajo, en la labor incansable de la investigación y que consiga transmitir a todos

los afortunados que le tenemos cerca. Mi sueño es ser algún día como él, o al menos acercarme un poquito....Es y ha sido un compañero de trabajo envidiable en todos los sentidos, y pese a ser mi director de tesis y supuesto "jefe", me ha tratado siempre como a uno más del resto de compañeros. Su continuo apoyo personal, emocional y profesional ha ido mucho más allá de lo imaginable, sin él jamás habría sido capaz de terminar este proyecto, los dos lo sabemos. Además, con permiso de su familia, ha sido para mí en cierto modo como el padre que Dios se llevó hace unos años y que a buen seguro donde quiera que se encuentre, le estará dando las gracias por cuidar así de su niña. Gracias Manolo por confiar en mí y por todo.

A mis compañeros y amigos del Área de Metodología de las Ciencias del Comportamiento: Lola Hidalgo, Fulgencio Marín, Antonio Velandrino, Julio Sánchez Meca, Jose Antonio López Pina, Jose Antonio López López, Rafael Rabadán y Juanjo López, en especial a estos dos últimos, darles las gracias por su ayuda en sus aportaciones a esta tesis.

A mis otros compañeros y amigos del Área de Psicología Básica, en especial a Damián Amaro por aclararme y ayudarme con los "engorrosos" trámites de tipo burocrático.

No puedo dejar de mostrar mi agradecimiento a mis amigos, ellos saben quienes son, sobre todo por su sinceridad, honestidad, apoyo en los momentos difíciles, y por qué no decirlo también, en los momentos de alegría. Gracias por estar ahí y que sea para toda la vida.

A **Marga, Inma, Carolina y Pulpita**. Gracias por brindarme vuestra hermosa amistad entre otras cosas, por los buenos e inolvidables momentos compartidos y simplemente por estar ahí cuando siempre os necesitaba. Nunca os olvidaré. Os quiero.

A **Raúl,** por su apoyo y cariño durante estos años en este proceso a largo

plazo, a mis desplantes, cambios de humor y a su apoyo incondicional y "logístico" en mis momentos de debilidad. Gracias.

A **mi abuela María Dolores**, que siempre estará en mi memoria con su sonrisa y del grato recuerdo de pasar horas a mi lado junto al pc, y sobre todo por su cariñosa frase : "nieta, ¿cuando vas a dejar de estudiar?", a la que yo contestaba siempre "si Dios quiere, nunca abuela".

A **mis hermanos Leandro y María Dolores**, a los que adoro y ellos lo saben. Por su continuo apoyo, en los momentos en los que flaqueaba y por sacarme a flote. Pese a que somos diferentes al expresar nuestras sentimientos, sé que se han preocupado y han sufrido para que saliera adelante este proyecto. Gracias hermanos por ser como sois y estar ahí .

A toda mi familia, paterna y materna, con la pregunta eterna , ¿cuándo terminas la tesis?

A mi prima **Toñi** y de nuevo a mi hermana **María Dolores**, por releer esta tesis y tratar de corregir errores, pese a reconocerme que no se enteraban de nada.

A la música que me ha acompañado en tantas horas de soledad y trabajo.

Y por último, a **mi madre María,** que como diría Antonio Vega " es una mujer hecha de algodón, de seda y de hierro puro". No tengo palabras de agradecimiento para ella, sin la cual este proyecto jamás habría sido posible. No he conocido a alguien con tanta fuerza y coraje para sacar adelante todo, siempre dando todo sin esperar nada a cambio, de invertir toda su vida sin medir la rentabilidad que le aporte su inversión, sufriendo cuando sufro, alegrándose de mis alegrías y siendo mi mejor confidente. Gracia mamá, te quiero mucho.

A todos ellos y muchos más que seguro se me han olvidado, espero me disculpen, muchas **gracias por todo**

A mis padres:

Juan Miguel y María.

Gracias por darme la vida y buscar siempre lo mejor para mí

Índice General

Introducción

En la investigación en las ciencias biológicas y sociales muchas medidas que se registran sobre los individuos se basan en las observaciones subjetivas de expertos o en la utilización de instrumentos diseñados para la medida de complejos constructos. Puesto que las medidas subjetivas y los instrumentos de medida son propensos a error, y la replicabilidad es un ingrediente esencial del conocimiento científico, una temática que se ha desarrollado preferentemente por los científicos que trabajan en el amplio espectro de disciplinas que cubren todas las ciencias biológicas y sociales es el **análisis de la fiabilidad** y la investigación acerca de las características y consecuencias de los **errores de medida** (Cochran, 1968; Crocker y Algina, 1986; Dunn, 1989, 2004).

Uno de los procedimientos estadísticos empleados para determinar la fiabilidad de un instrumento de medida es la **fiabilidad entre jueces**

(*reliability inter-rater*)**,** que es la correlación de las medidas aportadas por dos o más jueces u observadores que valoran un mismo conjunto de ítemes o instrumentos de medida. Un procedimiento alternativo que originalmente surgió en el ámbito de la metodología observacional (Suen y Ary, 1989) es el **acuerdo entre jueces** (*inter-rater agreement*). Aunque ambos términos se han empleado de forma similar en el pasado, en teoría reflejan conceptos diferentes. Los coeficientes de fiabilidad expresan la habilidad para diferenciar entre objetos y se definen en términos de una razón de varianzas -según la teoría psicométrica clásica, es la varianza atribuida a las diferencias entre sujetos dividida por la varianza total-. Por el contrario, los coeficientes de acuerdo expresan la concordancia o el consenso entre jueces u observadores. De ahí que un término afín al de acuerdo sea el de **concordancia** (Johnston y Bolstad, 1973), que se utiliza cuando se comparan dos jueces y se persigue describir el grado de acuerdo entre ambos. En cambio, la fiabilidad es un término más restrictivo que se refiere a la concordancia respecto a lo que se considera como "verdadero". Por esta razón, cuando se comparan dos evaluadores independientes solo puede hablarse de concordancia o acuerdo, pero cuando uno de ellos se compara con un estándar que se asume "verdadero" se habla de fiabilidad. Los parámetros de acuerdo determinan si el mismo valor se obtiene cuando una medida se registra dos veces, ya sea por el mismo observador o por observadores diferentes. Además, como trataremos aquí, los coeficientes utilizados para evaluar la fiabilidad y el acuerdo difieren según la escala de medida utilizada.

El interés de este trabajo se enmarca dentro de este contexto y se centra en una descripción crítica de los procedimientos estadísticos para evaluar el acuerdo entre las valoraciones de jueces en aquellas características

psicológicas o sociales que interesan al investigador aplicado, tanto con escalas de medida cuantitativas como con escalas de medida categóricas.

Siguiendo a Bakeman y Guttman (1987:131-132) hay al menos tres razones esenciales que justifican la utilización del análisis del acuerdo entre evaluadores. La primera porque permite asegurar que los observadores son precisos y los procedimientos utilizados replicables. La segunda porque persigue la calibración de observadores múltiples entre sí o con relación a algún estándar, en particular cuando la tarea a realizar es compleja. La tercera razón es porque aporta un feedback necesario al investigador acerca de la eficacia del entrenamiento cuando los jueces han sido previamente entrenados.

Algunos conceptos básicos

En el contexto de un estudio típico de acuerdo, denotamos como **ítem, sujeto** (*subject*), u **objeto** (*target*) a la entidad que es objeto de medida –para variables cuantitativas– o clasificación –para variables categóricas– y como **juez, observador, evaluador** (en términos genéricos*, rater*) o, en general, **instrumento de medida,** a la entidad que realiza la medida o clasificación de los ítemes. En este proceso analítico se utiliza por tanto un conjunto de ítemes que es evaluado por un conjunto de jueces o instrumentos de medida. Asumimos que los ítemes se extraen al azar de una población, que llamaremos **población de ítemes**. Del mismo modo se asume que existe una población de jueces o instrumentos de medida, que llamaremos **población de jueces**, de

donde se extraen los jueces, bien sea de forma arbitraria –modelo de efectos fijos– o bien al azar por algún procedimiento al uso –modelo de efectos aleatorios–. En el caso más usual, la población de jueces es fija mientras que la población de ítemes es aleatoria.

El resultado del proceso es la **medida o clasificación** (*rating*) de un ítem por parte de dos o más jueces que actúan independientemente. Este resultado se asume asimismo que procede de una población, la **población de medidas** o **clasificaciones** de un ítem realizadas por el conjunto de los jueces. Puesto que cabe esperar que algunos ítemes sean más fáciles de medir o de clasificar que otros, el resultado de este proceso puede reflejar en algunos casos el verdadero acuerdo que existe entre los observadores sobre la medida o clasificación de un ítem, pero en otros casos también puede incluir algún componente de error. Por esta razón, para un ítem determinado es conveniente distinguir (Gwett, 2001:28) entre una medida o **clasificación sistemática** (*deterministic rating*), cuando todos los jueces son capaces de identificar las características del ítem y medirlo o clasificarlo sistemáticamente de la misma forma, o **aleatoria** (*random rating*), cuando los evaluadores no son capaces de identificar las características distintivas del ítem y lo clasifican o miden dependiendo de criterios subjetivos cambiantes.

Es conveniente distinguir (Shoukri, 2004:5) dos componentes del error de medida que resultan complementarios: el **error aleatorio** o **imprecisión** (*random error, imprecision*), que es la cantidad de variabilidad inherente en los jueces o instrumentos de medida, y el **error sistemático** o **sesgo** (*systematic error, bias*), que es la desviación del valor observado respecto al valor verdadero. El error sistemático es aquel que se presenta siempre de la misma

forma, es decir, sistemáticamente. Por ejemplo, si tres personas cuyos pesos reales exactos son de 50, 55 y 60 Kg. se pesan en una báscula que arroja respectivamente pesos de 53, 58 y 63 Kg., las magnitudes de peso que reporta esta báscula defectuosa estarían afectadas de error sistemático, en este caso un valor constante positivo de +3 Kg. Es decir, la suma de los pesos reales es de 165, la de los pesos de la báscula de 174, y la diferencia media (9/3 = 3) es una estimación del **sesgo**. La varianza de las diferencias es una estimación de la **precisión** del instrumento, que en este caso es 9, siendo su error típico también igual a 3. En otros casos, el error sistemático no es constante, sino proporcional al valor real; por ejemplo, cuando los pesos de la báscula fueran 51, 56.100 y 61.200 Kg., vendrían afectados por un error proporcional del +2%. Ambas situaciones pueden darse conjuntamente, o lo que es lo mismo, no son excluyentes. Pero en general, mientras que el error aleatorio es impredecible, y por tanto no es en teoría susceptible de corrección, el error sistemático sigue un patrón conocido y sistemático que puede ser objeto de corrección.

Un modelo estadístico muy simple para analizar estos datos es el modelo clásico de la teoría de tests (Crocker y Algina, 1986): $X_{ij} = \tau_i + e_j$, donde X_{ij} es la *j*-ésima medida del *i*-esimo objeto, τ_i es el valor verdadero del ítem u objeto *i* y e_j es el error de medida asociado con X_{ij}. Asumiendo que los errores de medida son independientes entre sí y que la media de todos los errores es cero, entonces la varianza total de las medidas es igual a la suma de la varianza de los valores verdaderos y de los errores de medida: $\sigma_X^2 = \sigma_\tau^2 + \sigma_e^2$. Cuando todas las condiciones de medida son constantes, la precisión del proceso de medida viene representada por la raíz cuadrada de la varianza de los errores (en términos de un modelo ANOVA, la raíz cuadrada de

la MC Error), que se denomina alternativamente **desviación típica de repetibilidad** (*repetibility standard deviation*) en las ciencias experimentales y **error típico de medida** (*measurement standard error*) en las ciencias sociales y del comportamiento. No obstante ambos conceptos son en su esencia equivalentes (Dunn, 2004: 4).

El *International Standard Organization* ISO 5725 (ISO 1994a-1998) ha definido las condiciones experimentales recomendadas para la determinación de la precisión y del sesgo, enumerando entre los principales factores que contribuyen a la variabilidad de los resultados de un determinado procedimiento de medida los siguientes: el operador, el equipo utilizado, la calibración del equipo, el ambiente (temperatura, humedad, polución del aire, etc.) y el tiempo transcurrido entre las medidas. El ISO 5725 utiliza la desviación típica de repetibilidad como índice de precisión de un procedimiento de medida bajo unas condiciones específicas que denomina genéricamente **condiciones de repetibilidad** (*repetibility conditions*), que son las condiciones bajo las cuales se obtienen resultados de tests independientes con el mismo método en ítemes idénticos, en el mismo laboratorio, con el mismo operador y usando el mismo equipo dentro de intervalos pequeños de tiempo. Cuando tales condiciones se mantienen constantes, la desviación típica de repetibilidad se toma entonces como índice de precisión. En contraste, el ISO 5725 define las **condiciones de reproductibilidad** (*reproductibility conditions*) como aquellas condiciones donde los mismos resultados se obtienen con el mismo método, pero en diferentes laboratorios, con diferentes operadores y usando equipos diferentes (Bartko, 1991). La reproductibilidad es un aspecto de la precisión en ciencias experimentales de un sistema de medida (de Mast, 2007). En tales condiciones se suele emplear como índice de

precisión la **desviación típica de reproductibilidad** (*reproductibility standard deviation*).

Problemas de medida similares se encuentran también en las ciencias sociales y del comportamiento, y no solo en las ciencias experimentales. Sin embargo, las disciplinas que conforman las ciencias sociales y del comportamiento utilizan una terminología diferente. En lugar del concepto de reproductibilidad se emplea el concepto de **generalizabilidad** (*generalizability*), y todas las discusiones acerca de esta temática tienen hoy en día lugar en el contexto de la denominada **teoría de la generalizabilidad** (Cronbach, Rajaratnam y Gleser, 1963; Gleser, Cronbach y Rajaratnam, 1965; Shavelson y Web, 1991; Brennan, 2001a). Así, mientras que las áreas de las ciencias experimentales centran su preocupación hacia los estudios de precisión entre laboratorios, en las áreas de las ciencias sociales y del comportamiento la preocupación es el diseño y el análisis de estudios de generalizabilidad. El objeto es de naturaleza similar, porque tanto en un **estudio de generalizabilidad** (*generalizability study*) como en un **estudio de reproductibilidad** (*reproductibility study*) se investigan de forma sistemática las fuentes de variación de las medidas sociales y psicológicas, y más concretamente, la estimación eficiente de los componentes de la varianza que explican cada una de los efectos del modelo (pero no interesa la significación estadística de los efectos). Tales componentes pueden ser después combinados para producir coeficientes de fiabilidad que son el objeto final de interés del investigador aplicado.

Además de los estudios de reproductibilidad o de generalizabilidad, la evaluación estadística de errores de medida utiliza asimismo un segundo tipo

de estudios que Dunn (2004:18) denomina **estudios de comparación de métodos** (*method comparison studies*), donde el objetivo es investigar la covariación de las medidas producidas por dos (o más) métodos de medida, estimar sus parámetros y comparar su precisión. Es también posible en este contexto estimar coeficientes de fiabilidad a partir de los resultados de este tipo de estudios, que pueden emplear tanto medidas cuantitativas como medidas categóricas. Una forma específica de estudios de comparación de métodos, que es particularmente popular en disciplinas de tipo clínico, son los **estudios de evaluación de la concordancia o el acuerdo** entre dos métodos, jueces u observadores. Es aquí donde tienen su cabida los procedimientos estadísticos que son el objeto de este trabajo.

En consecuencia, dado un conjunto de ítemes, el análisis del grado de acuerdo global existente entre dos o más evaluadores se asume que refleja un componente de naturaleza sistemática (que se produce cuando todos los jueces utilizan el mismo criterio para interpretar las características de algunos ítemes y si se repitiera el estudio en las mismas condiciones se produciría con gran probabilidad el mismo resultado) y un componente de naturaleza aleatoria (cuando los evaluadores difieren en la interpretación de las características de algunos ítemes y, excepto por azar, no producen el mismo resultado en una repetición del estudio). En un estudio típico de acuerdo se asume además que existe un cierto equilibrio entre ambos componentes, tanto del efecto sistemático (acuerdo puro) como del efecto aleatorio (acuerdo debido al azar), en una proporción desconocida.

La presencia de error de medida en la evaluación de un ítem por un conjunto de observadores es la razón por la que el acuerdo entre sus

respectivas medidas o clasificaciones no es usualmente perfecto. En general, cuando mayor es el error de medida, menor es el acuerdo. La Tabla 1.1, tomada de la tesis doctoral de Carrasco (2004:11) presenta una clasificación de los estudios utilizados para la evaluación de la calidad de los instrumentos de medida.

Tabla 1.1.
Clasificación de los estudios para evaluar la calidad de los instrumentos de medida

Objetivos básicos del estudio	*Valores utilizados para la comparación*	*Denominación del estudio*
Evaluar la independencia de los errores Estimar la magnitud del error aleatorio	Valores obtenidos con el mismo procedimiento o instrumento de medida	Fiabilidad Consistencia Repetibilidad Reproductibilidad
Decidir si un instrumento puede reemplazar a otro Evaluar si distintos métodos o instrumentos son de hecho intercambiables	Valores obtenidos con un método/juez alternativo	Concordancia Consenso Acuerdo
Cuantificar el error de medida Estimar los parámetros que corrigen el error de medida	Valores reales de la variable	Calibración

Fiabilidad y acuerdo entre jueces

Aunque en muchos círculos científicos se emplean todavía en la actualidad como sinónimos términos tales como **fiabilidad entre jueces** *(inter-rater reliability*) y **acuerdo entre jueces** *(inter-rater agreement*), que se definen genéricamente como el grado de homogeneidad que existe entre diferentes jueces o instrumentos de medida en la valoración de un conjunto de ítemes (Wikipedia, entrada *inter-rater reliability*), no es en absoluto apropiada esta confusión en la terminología, ya que en teoría ambos conceptos son netamente diferentes y representan dos atributos importantes de una escala de medida, como refleja la Tabla 1.1.

Tinsley y Weiss (1975, 2000), Kozlowski y Hattrup (1992), Fleenor, Fleenor y Grossnickle (1996), Stemler (2004), Vangeneugden *y otros* (2005), De Vet (1998, 2006) y LeBreton y Senter (2008), entre otros muchos, han delimitado con claridad las diferencias entre ambos conceptos.

La fiabilidad entre jueces se refiere a la consistencia relativa en la medida o clasificación por parte de los jueces de un conjunto de ítemes (Bliese, 2000), y se utiliza para conocer si los jueces ordenan los ítemes de forma que sea relativamente consistente con otros jueces. Nótese que el interés no es la equivalencia puntual de las valoraciones, sino más bien la equivalencia relativa (o la consistencia) de la medida o clasificación.

En contraste, el acuerdo entre jueces se refiere a la equivalencia absoluta en las valoraciones emitidas por los jueces de un conjunto de ítemes y se utiliza para conocer si las valoraciones son intercambiables o equivalentes en

términos de sus valores absolutos. Es por tanto una cuestión de consenso (y no de consistencia) entre las valoraciones de los observadores.

La Tabla 1.2 presenta datos ficticios (tomados de un ejemplo utilizado por Tinsley y Weiss, 2000:99) para distinguir entre ambos conceptos, donde 3 evaluadores valoran un conjunto de 10 ítemes en una escala de valoración de 0 a 10 puntos. El caso I identifica un estudio con alta fiabilidad y alto grado de acuerdo entre jueces. Nótese que las valoraciones de los 3 jueces son exactamente iguales y en tal caso fiabilidad y acuerdo obtienen su valor máximo, que es igual a 1. El caso II identifica un estudio con alta fiabilidad pero bajo acuerdo entre jueces. Nótese que en este caso las valoraciones son proporcionales entre jueces, pero no son exactamente iguales. Y finalmente, el caso III identifica un estudio con relativamente baja fiabilidad y moderadamente alto acuerdo entre observadores.

Tabla 1.2. Tres casos ficticios para comparar fiabilidad y acuerdo

Ítemes	*Caso I*			*Caso II*			*Caso III*		
	Juez 1	*Juez 2*	*Juez 3*	*Juez 1*	*Juez 2*	*Juez 3*	*Juez 1*	*Juez 2*	*Juez 3*
1	1	1	1	1	3	6	5	6	5
2	2	2	2	1	3	6	5	4	4
3	3	3	3	2	4	7	6	4	6
4	4	4	4	2	4	7	4	5	6
5	5	5	5	3	5	8	5	4	4
6	6	6	6	3	5	8	6	6	5
7	7	7	7	4	6	9	4	4	5
8	8	8	8	4	6	9	5	5	4
9	9	9	9	5	7	10	4	5	3
10	10	10	10	5	7	10	6	6	6
Fiabilidad	1.000			1.000			.380		
Acuerdo	1.000			.000			.660		

En consecuencia, una alta fiabilidad no es en absoluto indicación de que los jueces acuerden en sentido absoluto en el grado en el que los objetos valorados posean la característica que se juzga (caso II), ni tampoco indica necesariamente que los observadores estén en desacuerdo (caso III). El acuerdo es alto siempre y cuando los evaluadores concuerdan en su respuesta, y disminuye en la medida en que disminuye la concordancia.

Objetivos de este trabajo

Un detenido repaso a la literatura de investigación aplicada demuestra que se emplean muchos y muy variados procedimientos para evaluar el acuerdo entre jueces u observadores, dependiendo de la disciplina de que se trate y del nivel de medida de las variables utilizadas. Sin ninguna duda, el más popular es el coeficiente *kappa*, originalmente propuesto por Cohen (1960) y la razón que justifica su popularidad es la simplicidad computacional que se requiere para obtenerlo y su facilidad de acceso y de cálculo mediante alguno de los principales paquetes estadísticos (SAS, SPSS, STATA, GENSTAT, SYSTAT), aunque el coeficiente solo se aplica a datos nominales u ordinales. Sin embargo, el coeficiente *kappa* ha sido severamente criticado por presentar problemas de prevalencia y de sesgo, que tratamos detalladamente más adelante. Por otra parte, con datos numéricos los procedimientos utilizados en una disciplina científica son en la práctica desconocidos en otras áreas diferentes. Así, el coeficiente de correlación de contingencia (Lin, 1989) se emplea usualmente en disciplinas clínicas, pero que sepamos nunca se ha aplicado en las ciencias sociales, mientras que el coeficiente r_{WG} (James, Demaree y Wolf, 1984) se emplea en determinadas áreas de la psicología social , pero es totalmente desconocido en las ciencias biológicas y otras disciplinas de la psicología y las ciencias sociales. Por su parte, un coeficiente que por su generalidad presenta gran interés, aunque más complejo de obtener desde un punto de vista computacional, es el coeficiente *iota* (Janson y Olson, 2001), que se basa en la aplicación de los métodos de permutación (Berry y Mielke, 2007) y es prácticamente desconocido en todos los campos científicos, a pesar

de ser un coeficiente que se aplica a datos tanto numéricos como categóricos e incluso puede ser fácilmente extendido al caso multivariante.

Dada esta anómala situación de la investigación actual sobre el acuerdo entre jueces u observadores, nos proponemos realizar como ***primer objetivo*** de este trabajo un tratamiento formal y pormenorizado de las medidas de acuerdo más importantes que se han propuesto en la literatura (con atención especial a las medidas para datos categóricos, que representan más del 90% de las aplicaciones de la psicología y ciencias afines), de sus condiciones de aplicabilidad y de sus respectivas ventajas e inconvenientes, que sirva de apoyo útil al investigador aplicado para tomar una decisión apropiada acerca de qué medida de acuerdo es más conveniente aplicar en cada situación de investigación concreta. Aunque hay varios textos que han tratado con carácter monográfico esta temática (véase por ejemplo Shoukri, 2004 y von Eye y Mun, 2005), no se observa en ninguno de los dos el fin integrador que se pretende en este trabajo. A este respecto hemos considerado dos grandes bloques de medidas de acuerdo discernibles en la literatura:

1) Medidas descriptivas
2) Medidas basadas en modelos estadísticos

Un problema fundamental que concierne a muchas de las medidas de acuerdo es la dificultad computacional para su obtención en la práctica; y aunque algunas de ellas pueden encontrarse en programas estadísticos muy especializados (por ejemplo, la biblioteca de programas de R o macros

específicos de SAS o SPSS), no son procedimientos en general sencillos de comprender por el investigador aplicado. Por esta razón, como ***segundo objetivo*** de este trabajo se ha detallado cuando ha sido necesario (con medidas descriptivas sobre todo, pero también con medidas basadas en modelos loglineales y mixtura) la forma de calcular los coeficientes de acuerdo con el apoyo de plantillas de computación práctica y otros recursos didácticos, muchos de las cuales son de elaboración exclusiva para este trabajo y no han sido hasta ahora previamente publicados.

El ***tercer objetivo*** que este trabajo persigue es la demostración práctica de que la totalidad de los procedimientos estadísticos descriptivos para evaluar el acuerdo entre jueces utilizados con datos numéricos y categóricos tienen un origen común y pueden ser formalizados mediante un modelo ANOVA en dos sentidos, donde los ítems u objetos actúan como efecto aleatorio y los jueces u observadores como efecto fijo (o en muy raras ocasiones, aleatorio). El correspondiente modelo mixto para datos numéricos es bien conocido (véase Ato y Vallejo, 2007; Milliken y Johnson, 2009), pero para datos categóricos puede ser o bien formalizado mediante un modelo mixto lineal generalizado, en la línea apuntada por Nelson y Edwards (2008), o mediante una adaptación del modelo ANOVA clásico para datos categóricos (CATANOVA), en la línea de los clásicos trabajos de Light y Margolin (1971) y Margolin y Light (1974) y los más recientes trabajos de Onukogu (1985a,b), Cox (1997) y Singh (1993, 1996). Esta es la línea que hemos perseguido en este trabajo.

Uno de los procedimientos de mayor interés para datos categóricos, porque encajan en el marco del modelado estadístico que permite superar muchos de los problemas que presentan los procedimientos basados en el enfoque

descriptivo, son las medidas de acuerdo basadas en modelos mixtura, que no solo contemplan el ajuste de modelos a datos empíricos, sino también la posibilidad de incluir medidas complementarias a la evaluación del acuerdo, abriendo nuevas e interesantes perspectivas de investigación futuras. Como ***cuarto objetivo*** de este trabajo se considera una de estas medidas complementarias el **sesgo** entre jueces u observadores y se propone realizar un análisis pormenorizado del sesgo comparando, mediante simulación MonteCarlo, una medida clásica del sesgo entre jueces (Ludbrook, 2002, 2004) con una medida obtenida con la versión ampliada de los modelos mixtura recientemente propuesta por Ato, López y Benavente (2008).

Capítulo 1

Datos cuantitativos

1.1. Medidas de acuerdo para variables numéricas

Para definir una medida del acuerdo existente entre jueces, que es uno de los objetivos esenciales de esta tesis doctoral, en el pasado se ha considerado deseable (véase Dunn, 1989: 12) que el efecto del acuerdo debido al azar sea controlado para que la medida refleje en su pureza el verdadero grado de acuerdo. Son muchas las formas de acuerdo que pueden definirse en este contexto, dependiendo de la naturaleza de la medida y de su nivel de agregación. En este capítulo se abordan algunas de las medidas de acuerdo más relevantes para la investigación aplicada, y particularmente para las ciencias

sociales y del comportamiento, que se han propuesto cuando el nivel de medida es **numérico.**

Los capítulos siguientes examinarán con detalle las medidas más populares que se han propuesto cuando el nivel de medida es **categórico.** Aunque el grado de atención que se ha prestado a los datos categóricos para la evaluación del acuerdo ha sido bastante mayor que el dedicado a datos numéricos en la literatura de investigación psicológica, en general las medidas de acuerdo con datos categóricos han resultado considerablemente más controvertidas. Se advierte una considerable proliferación en la cantidad de las medidas propuestas, en algunos casos insuficientemente justificadas desde un punto de vista estadístico, y muchas de ellas presentan problemas de cálculo por lo que no resultan suficientemente comprensibles. Quizás por esta razón se ha popularizado entre los investigadores aplicados uno de los procedimientos descriptivos más sencillos de cálculo, el coeficiente *kappa* de Cohen (1960).

En cualquiera de los dos casos, las medidas de acuerdo se pueden también dividir entre **técnicas agregadas**, que utilizan un único coeficiente global para valorar el acuerdo, sin considerar sus potenciales componentes, y **técnicas desagregadas** que, además de un coeficiente global, permiten evaluar por separado los diferentes componentes de error sistemático y error aleatorio tratados anteriormente. Obviamente, las técnicas desagregadas permiten analizar con mayor grado de detalle las posibles causas de la falta de acuerdo y son en general preferibles a las técnicas agregadas.

Cuando el nivel de medida es numérico, los procedimientos más comunes para la evaluación del acuerdo entre observadores en la investigación aplicada son el **coeficiente de correlación intraclase** (*Intraclass Correlation*

Coefficient, ICC), el **coeficiente de correlación de concordancia** (*Concordance Correlation Coefficient,* CCC) y el **índice de acuerdo** r_{WG}, que tratamos detenidamente en este mismo capítulo.

Es importante, sin embargo, destacar las diferentes raíces teóricas y científicas de los procedimientos citados. Mientras que la correlación intraclase se propuso en su origen como medida básica de fiabilidad y continúa siendo un pilar básico de la medida del comportamiento en psicometría con el advenimiento de la teoría de la generalizabilidad (Cronbach y otros, 1972; Shavelson y Webb, 1991; Brennan, 2001a), el coeficiente de correlación de concordancia se propuso más recientemente (Lin, 1989, aunque tiene antecedentes históricos que se remontan a los trabajos de Krippendorf, 1970) en el contexto de la investigación biológica, y el índice de acuerdo r_{WG} se propuso también recientemente (James y otros, 1984) en el contexto de la investigación psicosocial. Esta dispersión da una idea del problema que intentamos abordar en este trabajo, el que muchos procedimientos estadísticos para evaluar el acuerdo tienen orígenes diferentes, asociados en particular con áreas concretas de conocimiento, a veces sin ninguna conexión entre sí, pero que tienen una notable similaridad. Gran parte de este trabajo persigue demostrar las bases estadísticas comunes que tienen los estudios para la evaluación del acuerdo.

1.2. El coeficiente de correlación intraclase

El coeficiente de correlación intraclase (*Intraclass Correlation Coefficient*, ICC) fue propuesto por Sir Francis Galton en 1887 para definir la relación entre medidas de individuos de una misma familia, y en concreto, para estimar la correlación entre todas las parejas (*clusters*) de hermanos posibles. Pearson (1901) propuso el estimador del ICC en base al producto de los momentos. Más adelante, Fisher (1925) formuló el estimador del ICC utilizando los componentes de la varianza y definiéndolo como la razón entre la variabilidad de los *clusters* en relación con la variabilidad total. Obviamente, esta definición depende del diseño de recogida de datos y del modelo de medida subyacente utilizados. En consecuencia, el ICC no tendrá la misma expresión para medir la fiabilidad de un método de medida que para evaluar el acuerdo o concordancia entre instrumentos de medida.

El uso del coeficiente de correlación intraclase como un índice de la fiabilidad de las medidas está bien documentado en la investigación psicológica (e.g. Ebel, 1951; Haggard, 1958; Bartko, 1966; Cronbach y otros., 1972). Dado un conjunto de ítemes y un conjunto de medidas realizadas por un conjunto de jueces, como se presenta en la Tabla 1.2 para N ítemes y J jueces, Shrout y Fleiss (1979) y McGraw y Wong (1996) describieron cinco modelos –tres modelos básicos propuestos por los primeros y dos modelos complementarios propuestos por los últimos– para un estudio típico de fiabilidad entre jueces, cada uno de los cuales requiere un modelo matemático específico para describir los resultados. Todos los modelos considerados pueden contener efectos para el j-ésimo juez, para el i-ésimo ítem y para la interacción entre juez e ítem, además de un nivel constante (μ) de las

medidas registradas y un componente de error aleatorio. Dependiendo de la forma con que el estudio se diseñe, existen diferentes modelos ANOVA resultantes, que se tratarán de forma separada a continuación.

La notación general que emplearemos en esta sección se resume en la Tabla 1.1.

Tabla 1.1. Notación general

Ítemes /Objetos (I)	*Jueces / Métodos (J)*				
	1	...	*j*	...	*J*
1	y_{1l}	...	y_{1j}	...	y_{1J}
...	...	...	...	...	...
i	y_{il}	...	y_{ij}	...	y_{iJ}
...	...	...	...	...	...
N	y_{Nl}	...	Y_{Nj}	...	Y_{NJ}

1.2.1 Modelo I

Cada uno de los *N* ítemes es evaluado por un conjunto diferente de *J* jueces aleatoriamente seleccionados de una población de jueces. En tal caso, las respuestas de los diferentes observador constituyen réplicas que se anidan dentro de las condiciones del factor ítemes. Siendo y_{ij} la respuesta que

corresponde al *j*-ésimo juez/método para del *i*-ésimo ítem/objeto, el modelo que se aplica es un ANOVA sencillo con efectos aleatorios,

$$y_{ij} = \mu + a_i + e_{ij} \qquad \text{(Ec. 1.1)}$$

donde μ es una constante (la media de la población de todas las medidas), a_i es la desviación del *i*-ésimo ítem respecto de la media μ, y e_{ij} es un componente residual que incluye efectos inseparables debidos a los jueces y a la interacción ítemes $\times$ jueces, que consta igualmente de un componente de error aleatorio. El componente a_i se asume que varía normalmente con media 0 y varianza σ_I^2 y es independiente de todos los demás componentes del modelo. También se asume que el componente e_{ij} se distribuye normal e independientemente con media 0 y varianza σ_e^2. El modelo ANOVA resultante se muestra en la Tabla 1.2.

Tabla 1.2. ANOVA para el Modelo I

Fuentes	*gl*	*MC*	*E(MC)*
Ítemes (*I*)	$N-1$	MC_I	$\sigma_e^2 + J\,\sigma_I^2$
Error	$N(J-1)$	$MC_{J(I)}$	σ_e^2
Total	$NJ-1$		

El cálculo de la correlación intraclase para el Modelo I puede obtenerse utilizando las medias cuadráticas o bien las varianzas estimadas a partir de la

estimación de las esperanzas de las medias cuadráticas (véase Ato y Vallejo, 2007) mediante

$$\tilde{\rho}_I = \frac{MC_I - MC_{J(I)}}{MC_I + (N-1)MC_{J(I)}} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_e^2} \qquad \text{(Ec. 1.2)}$$

El coeficiente de correlación intraclase para el Modelo I representa el grado de acuerdo absoluto entre las medidas realizadas sobre ítemes aleatoriamente seleccionados.

1.5.2. Modelo II

Cada uno de los *N* ítemes seleccionados al azar de una población es evaluado por el mismo conjunto de *J* observadores, que se asume igualmente extraído de manera aleatoria de una población. En este caso, el efecto debido a los Jueces y el efecto de la interacción de Ítemes $\times$ Jueces son perfectamente separables del componente de error. Siendo y_{ij} la medida resultante del *i*-ésimo ítem por el *j*-ésimo juez, el modelo que se aplica es un ANOVA en dos sentidos con efectos aleatorios,

$$y_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij} \qquad \text{(Ec. 1.3)}$$

donde μ es una constante (la media de la población de todas las medidas), a_i es la desviación del *i*-ésimo ítem respecto de la media μ, b_j es la desviación del *j*-ésimo juez respecto de la media μ, (ab_{ij}) es el efecto de la interacción ítem $\times$ juez y e_{ij} es un componente de error aleatorio. Del mismo modo que con el Modelo I, el componente a_i se asume que varía normalmente con media 0 y varianza σ_I^2 y es independiente de todos los demás componentes del modelo. Se asume que el componente b_j varía normalmente con media 0 y varianza σ_J^2. De igual modo se asume que el componente e_{ij} se distribuye normal e independientemente con media 0 y varianza σ_e^2. Sin embargo, puesto que no hay medidas repetidas de cada observador en la valoración de cada objeto, los componentes (ab_{ij}) y e_{ij} no pueden ser estimados por separado. Cuando la interacción está presente la tabla ANOVA del modelo presenta la forma que resume la Tabla 1.3.

Tabla 1.3. ANOVA para el Modelo II (interacción presente)

Fuentes	*gl*	*MC*	*E(MC)*
Ítemes (*I*)	$N-1$	MC_I	$\sigma_e^2+J\,\sigma_I^2+\sigma_{IJ}^2$
Jueces (*J*)	$J-1$	MC_J	$\sigma_e^2+I\,\sigma_J^2+\sigma_{IJ}^2$
Residual	$(N-1)(J-1)$	MC_R	$\sigma_e^2+\sigma_{IJ}^2$
Total	$NJ-1$		

Este es el modelo contemplado en el trabajo original de Shrout y Fleiss (1979), que asume la existencia de interacción entre ítemes y jueces. Mientras

que hay solamente un coeficiente de correlación intraclase cuando los datos siguen una representación mediante un modelo ANOVA sencillo, hay dos coeficientes de correlación intraclase cuando los datos siguen una representación mediante un modelo ANOVA en dos sentidos (véase Berk, 1979; Suen y Ary, 1989 y Shavelson y Webb, 1991): el primero mide la correlación intraclase usando una definición de consistencia (**análisis de la consistencia)**; y el segundo mide la correlación intraclase usando una definición de desacuerdo absoluto (**análisis del acuerdo**). Usualmente la varianza de los jueces σ_J^2 se excluye del denominador en el análisis de la consistencia, ya que se asume que es una fuente irrelevante de varianza, pero se utiliza como una fuente esencial para el análisis del acuerdo.

El cálculo del coeficiente de correlación intraclase para el Modelo II puede obtenerse utilizando las medias cuadráticas y sus respectivas esperanzas de las medias cuadráticas (Ato y Vallejo, 2007).

Para el análisis de la consistencia, el coeficiente de correlación intraclase para el Modelo II con interacción presente resulta

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R} = \frac{\sigma_I^2}{\sigma_I^2 + (\sigma_{IJ}^2 + \sigma_e^2)} \qquad \text{(Ec. 1.4)}$$

y se interpreta como el grado de consistencia entre medidas. Se conoce también como **fiabilidad referenciada a la norma**. En la teoría de la generalizabilidad la Ecuación 1.4 estima la correlación al cuadrado entre las medidas individuales y las puntuaciones del universo.

Un modelo alternativo al Modelo II, que McGraw y Wong (1996) llamaron Modelo IIA, asume que la interacción ítemes × jueces no está presente. En tal caso, la tabla ANOVA presenta la forma que se muestra en la Tabla 1.4, mucho más simple que la propuesta en la Tabla 1.3.

Tabla 1.4. ANOVA para el Modelo IIA (interacción ausente)

Fuentes	*gl*	*MC*	*E(MC)*
Ítemes (*I*)	$N-1$	MC_I	$\sigma_e^2 + J\,\sigma_I^2$
Jueces (*J*)	$J-1$	MC_J	$\sigma_e^2 + I\,\sigma_J^2$
Residual	$(N-1)(J-1)$	MC_R	σ_e^2
Total	$NJ-1$		

Para el análisis de la consistencia, el coeficiente de correlación intraclase con el Modelo IIA se define como

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_e^2} \qquad \text{(Ec. 1.5)}$$

Para el análisis del acuerdo, el coeficiente de correlación intraclase para el Modelo II se define mediante

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R + \frac{J}{N}(MC_J - MC_R)} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_J^2 + (\sigma_{IJ}^2 + \sigma_e^2)} \qquad \text{(Ec. 1.6)}$$

y se interpreta como el acuerdo absoluto entre las medidas. En contraste con la Ecuación 1.5, esta formulación se conoce también como **fiabilidad referenciada al criterio** (McGraw y Wong, 1996).

Por su parte, para el Modelo IIA, el análisis del acuerdo se define del siguiente modo:

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R + \frac{J}{N}(MC_J - MC_R)} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_J^2 + \sigma_e^2} \qquad \text{(Ec. 1.7)}$$

1.2.3. Modelo III

Cada uno de los *N* ítemes, seleccionados al azar de una población, es evaluado por el mismo conjunto de *J* jueces, cuyo conjunto representa toda la población de jueces. En este caso, el efecto debido a los Jueces y el efecto de la interacción de Ítemes × Jueces son perfectamente separables del componente de error. Siendo y_{ij} la medida resultante del *i*-ésimo ítem por el *j*-ésimo juez, el modelo que se aplica es un ANOVA en dos sentidos con efectos mixtos,

$$y_{ij} = \mu + a_i + \beta_j + (a\beta)_{ij} + e_{ij} \qquad \text{(Ec. 1.8)}$$

donde μ es una constante (la media de la población de todas las medidas), a_i es la desviación del *i*-ésimo ítem respecto de la media μ, β_j es la desviación del *j*-ésimo juez respecto de la media μ, $(a\beta)_{ij}$ es el efecto de la interacción Ítem $\times$ Juez y e_{ij} es un componente de error aleatorio. El componente a_i es un efecto aleatorio que varía normalmente con media 0 y varianza σ_I^2 y es independiente de todos los demás componentes del modelo. Sin embargo, el componente β_j es un efecto fijo, que se asume sometido a la restricción $\sum \beta_j^2 = 0$. Asimismo, se asume que el componente e_{ij} se distribuye normal e independientemente con media 0 y varianza σ_e^2. No obstante, puesto que no hay medidas repetidas de cada juez en la valoración de cada ítem, los componentes $(a\beta)_{ij}$ y e_{ij} no pueden ser estimados por separado.

Cuando la interacción está presente (Modelo III), la tabla ANOVA del modelo adopta la forma que presenta la Tabla 1.5, y se corresponde con el Modelo III original de Shrout y Fleiss (1979).

Tabla 1.5. ANOVA para el Modelo III (interacción presente)

Fuentes	*gl*	*MC*	*E(MC)*
Ítemes (*I*)	$I-1$	MC_I	$\sigma_e^2+J\,\sigma_I^2$
Jueces (*J*)	$J-1$	MC_J	$\sigma_e^2+I\,\theta_J^2+\frac{J}{(J-1)}\sigma_{IJ}^2$
Residual	$(N-1)(J-1)$	MC_R	$\sigma_e^2+\frac{J}{(J-1)}\sigma_{IJ}^2$
Total	$NJ-1$		

Un modelo complementario fue formulado por McGraw y Wong (1996), que llamaron Modelo IIIA, cuando la interacción se asume ausente. El Modelo IIIA resulta más simple que el modelo original y su tabla ANOVA presenta la forma que muestra la Tabla 1.6 siguiente.

Tabla 1.6. ANOVA para el Modelo IIIA(interacción ausente)

Fuentes	*gl*	*MC*	*E(MC)*
Ítemes (*I*)	$N-1$	MC_I	$\sigma_e^2+J\,\sigma_I^2$
Jueces (*J*)	$J-1$	MC_J	$\sigma_e^2+N\,\sigma_J^2$
Residual	$(N-1)(J-1)$	MC_R	σ_e^2
Total	$NJ-1$		

Para un análisis de la consistencia, el coeficiente de correlación intraclase se define para el Modelo III mediante

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R} = \frac{\sigma_I^2 - \sigma_{IJ}^2/(J-1)}{\sigma_I^2 + (\sigma_{IJ}^2 + \sigma_e^2)} \qquad \text{(Ec. 1.9)}$$

y para el Modelo IIIA

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_e^2} \qquad \text{(Ec. 1.10)}$$

En cualquiera de los dos casos, el coeficiente evalúa el grado de consistencia entre las medidas realizadas bajo los niveles fijos del factor de jueces, aunque subestima la fiabilidad en el caso del Modelo III.

Para un análisis del acuerdo, el coeficiente de correlación intraclase se define para el Modelo III como

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R + \frac{J}{N}(MC_J - MC_R)} = \frac{\sigma_I^2 - \sigma_{IJ}^2/(J-1)}{\sigma_I^2 + \theta_J^2(\sigma_{IJ}^2 + \sigma_e^2)} \qquad \text{(Ec. 1.11)}$$

y para el Modelo IIIA como

$$\tilde{\rho}_I = \frac{MC_I - MC_R}{MC_I + (J-1)MC_R + \frac{J}{N}(MC_B - MC_R)} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_J'^2 + \sigma_e^2} \qquad \text{(Ec. 1.12)}$$

donde σ'^2_J es un estimador especial de la varianza de los jueces que se tratará más adelante y se interpreta como el acuerdo absoluto entre las medidas realizadas asumiendo niveles fijos para el factor jueces.

1.2.4. Correlación intraclase y teoría de la generalizabilidad

En la teoría clásica de tests, una observación de cualquier característica sobre un individuo (X) se asume que es una combinación de la puntuación verdadera de tal individuo (T) más error de medida aleatorio (E). Hoy por hoy resulta obvio que el supuesto de que toda la varianza presente en las valoraciones de los jueces se compone de varianza verdadera y varianza de error es demasiado simplista (Marcoulides, 2000). Además, hay muchos enfoques para estimar la fiabilidad de las puntuaciones, cada uno de los cuales genera un coeficiente diferente (consistencia interna, fiabilidad test-retest, fiabilidad de formas equivalentes y fiabilidad entre jueces), lo que conduce a diferentes estimaciones de las puntuaciones verdaderas para cada estudio, sin ningún medio lógico de combinarlas. Pero además de la puntuación verdadera de un individuo, existen múltiples fuentes potenciales de error. El objetivo es obtener la estimación más precisa de la puntuación que una persona obtendría si no hubiera fuentes de error que contaminan los resultados. Cada una de las formas de fiabilidad citadas identifica solo una fuente de error a un tiempo.

Se precisa por tanto alguna forma de combinar todas las fuentes de variabilidad en un estudio único, utilizando todos los datos disponibles para

estimar la varianza entre sujetos y los diferentes componentes de error. Este innovador enfoque fue originalmente desarrollado por Cronbach, Rajaratnam y Gleser (1963) y es conocido como **teoría de la generalizabilidad** (Cronbach, Glesser y Rajaratnam, 1972; Shavelson y Webb, 1991; Marcoulides, 2000; Brennan, 2001a). La teoría de la generalizabilidad representa hoy por hoy el marco teórico donde se fundamentan modelos y estimaciones más complejos de correlación intraclase que hemos tratado aquí. La esencia de esta teoría es el reconocimiento de que en cualquier situación de medida hay múltiples fuentes de varianza de error y el objetivo esencial es identificar, estimar y posiblemente encontrar estrategias para reducir la influencia de tales fuentes sobre la medida en cuestión.

1.3. El coeficiente de correlación de concordancia

Siguiendo a Carrasco (2004), si una característica cuantitativa –e.g. el tiempo de respuesta– se midiera con dos procedimientos distintos –a saber, dos tests paralelos *A* y *B*– en un conjunto de *N* ítemes tomados al azar de una población y los pares de medidas resultantes se representaran sobre un diagrama de dispersión, una simple inspección de los puntos del diagrama permitiría comprobar la fiabilidad, acuerdo o concordancia entre las medidas de ambos procedimientos examinando la aproximación de los puntos a la bisectriz (esto es, la línea que marca la igualdad $A = B$ o línea con ángulo de 45º respecto del origen). Si todos los puntos se sitúan encima de la bisectriz, la concordancia

será perfecta. En cambio, si los puntos se desvían de la bisectriz, la concordancia será tanto menor cuando mayor sea la distancia de los puntos respecto de la bisectriz. En líneas generales, la media de las distancias de cada uno de los puntos es proporcional a la desviación cuadrática media (*Mean Squared Deviation,* MSD).

$$MSD = \frac{1}{N}\sum_{i=1}^{N}(A_i - B_i)^2 \qquad \text{(Ec. 1.13)}$$

Siendo μ_A y μ_B las respectivas medias de cada test, σ_A y σ_B sus respectivas desviaciones típicas y ρ_{AB} su correlación, la expresión de la desviación cuadrática media puede reformularse como

$$MSD = (\mu_A - \mu_B)^2 + (\sigma_A - \sigma_B)^2 + 2(1 - \rho_{AB})\sigma_A\sigma_B \qquad \text{(Ec. 1.14)}$$

La concordancia o acuerdo entre ambos tests será perfecta si *MSD* = 0, y solamente será posible si los tres términos de la ecuación son cero, o análogamente, que no exista diferencia entre medias (es decir, que no exista error sistemático constante), que no exista diferencia entre las desviaciones típicas (dicho de otro modo, que no exista error sistemático proporcional), y que la correlación sea perfecta (por lo tanto, ausencia de error aleatorio). En otras palabras, debe cumplirse simultáneamente que $(\mu_A - \mu_B) = 0$, $(\sigma_A - \sigma_B) = 0$ y $\rho_{AB} = 1$. Procedimientos aplicados en el pasado para

analizar la concordancia, tales como la igualdad de medias o el coeficiente de correlación de Pearson no se consideran hoy aceptables, ya que puede haber medias iguales con variables que no concuerdan (debido a diferencias en variabilidad), y de manera análoga, la relación lineal entre variables puede ser perfecta, pero la recta de ajuste puede no ser la bisectriz (debido a diferencias entre medias y/o variabilidad).

Un procedimiento que actualmente está acaparando mucho interés por parte de los investigadores para evaluar la concordancia/acuerdo entre medidas fue derivado por Lin (1989, 1990, 2002) reescalando la MSD para adoptar valores comprendidos entre -1 y $+1$, y desde entonces se conoce como **coeficiente de correlación de concordancia** (*Concordance Correlation Coefficient*, CCC), que se formula de la forma siguiente:

$$\rho_C = 1 - \frac{E(A_i - B_i)^2}{E(A_i - B_i)^2 | A, B \text{ independientes}} = \frac{2\sigma_{AB}}{\sigma_A^2 + \sigma_B^2 + (\mu_A - \mu_B)^2} \qquad \text{(Ec. 1.15)}$$

donde σ_{AB} es la covarianza entre A y B. La medida CCC asume que tanto A como B se distribuyen de forma normal. Cuando la concordancia es perfecta, CCC adopta el valor 1, y en el caso de independencia entre las medidas adopta el valor 0. Aunque en teoría CCC puede tomar valores negativos, no resultan de interés sustantivo.

En su formulación original (Lin, 1989, aunque algunos consideran que fue inicialmente presentado por Krippendorff, 1979; véase por ejemplo Nickerson, 1987; Shoukri, 2002), el coeficiente de correlación de concordancia se definió como el producto de dos componentes: un componente de **precisión**

(*precision*) y otro de **exactitud** (*accuracy*). Como consecuencia de esta distinción, a la que probablemente debe su popularidad, CCC no evalúa únicamente en qué medida cada observación se desvía de la línea de ajuste a los datos (el componente de precisión), sino además en qué medida la recta de ajuste se desvía de la línea de 45º a través del origen (el componente de exactitud). La concordancia o acuerdo entre 2 observadores requiere que sus medidas tengan una correlación lineal perfecta (precisión) y presenten ausencia de sesgo sistemático (exactitud). Ambos componentes se obtienen a partir de la fórmula general del coeficiente de correlación de concordancia, siendo su covarianza $\sigma_{AB}=\rho_{AB}\sigma_A\sigma_B$, mediante

$$\rho_C=\frac{2\rho_{AB}\sigma_A\sigma_B}{\sigma_A^2+\sigma_B^2+(\mu_A-\mu_B)^2}=\rho_{AB}\left(\frac{2\sigma_A\sigma_B}{\sigma_A^2+\sigma_B^2+(\mu_A-\mu_B)^2}\right)=\rho_{AB}\chi_{AB} \quad \text{(Ec. 1.16)}$$

donde ρ_{AB} es el coeficiente de correlación producto-momento de Pearson y representa el componente de precisión, entretanto que χ_{AB} representa el componente de exactitud. Así, CCC se reduce por tanto al coeficiente de correlación de Pearson cuando $\mu_A=\mu_B$ y $\sigma_A^2=\sigma_B^2$, esto es, en el caso de homogeneidad de medias y varianzas. El coeficiente de correlación de concordancia se define por consiguiente como el producto del componente del precisión por el componente de exactitud, permitiendo en este sentidoí un análisis desagregado de ambos componentes del CCC (Lin, 2008). La relación entre el coeficiente de correlación de Pearson y el coeficiente de correlación de concordancia es patente en la ecuación anterior.

1.4. Procedimientos de estimación del coeficiente de correlación de concordancia

Se han propuesto varios procedimientos para estimar el coeficiente de correlación de concordancia. En sus primeros trabajos, Lin (1989, 1992) propuso estimar el CCC por el **método de los momentos**, es decir, sustituyendo directamente las varianzas, covarianzas y medias paramétricas por sus correspondientes estimadores muestrales de una muestra bivariante independiente en la fórmula general para obtener un estimador de ρ_C, demostrando la normalidad asintótica de $\tilde{\rho}_C$ y proponiendo que la aproximación normal puede ser sensiblemente mejorada utilizando la transformación *Z* de Fisher,

$$Z = \tanh^{-1}(\tilde{\rho}_C) = 1/2 \ln\left(\frac{1+\tilde{\rho}_C}{1-\tilde{\rho}_C}\right) \qquad \text{(Ec. 1.17)}$$

Un enfoque alternativo la presentan Carrasco y Jover (2003), quienes proponen estimar el OCCC (*Overall Concordance Correlation Coefficient*, OCCC) mediante el **método del coeficiente de correlación intraclase** con el Modelo IIIA de la correlación intraclase, utilizando para ello un modelo ANOVA factorial mixto con procedemiento de estimación mediante máxima verosimilitud (*Maximum Likelihood Estimation*, MLE) o máxima verosimilitud restringida (*Restricted Maximum Likelihood Estimation*, REML). El modelo estructural que se propone es el siguiente:

$$Y_{ij} = \mu + a_i + \beta_j + e_{ij} \qquad \text{(Ec. 1.18)}$$

donde β_j es el efecto de los Jueces , que en general representan efectos fijos (en raras ocasiones pueden considerarse también efectos aleatorios), a_i es el efecto de los Ítemes –que son siempre efectos aleatorios– y e_{ij} representan los efectos residuales –que por su propia naturaleza son también efectos aleatorios–.

Como se mostró anteriormente, dependiendo de la naturaleza, fija o aleatoria, del efecto de los jueces, es posible seleccionar entre una de dos expresiones para el coeficiente de correlación intraclase,

$$\tilde{\rho}_{I(1)} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_J^2 + \sigma_e^2} \qquad \text{(Ec. 1.19)}$$

donde el efecto de los Jueces se considera aleatorio y se distribuye de forma normal, $b_j \sim N(0, \sigma_J^2)$, o bien,

$$\tilde{\rho}_{I(2)} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_e^2} \qquad \text{(Ec. 1.20)}$$

donde el que efecto de los Jueces se considera como un efecto fijo. Es obvio

que el primero es el único procedimiento válido para abordar un análisis del acuerdo, ya que es el único que tiene en cuenta las diferencias entre observadores. Por consiguiente, para medir el acuerdo entre observadores es preciso utilizar el coeficiente de correlación intraclase $\tilde{\rho}_{I(1)}$, aún cuando el efecto del Observador sea fijo (lo que sucede con mucha mayor frecuencia que cuando es aleatorio). Y, en este caso, la estimación de la varianza de los jueces no es en realidad una varianza, sino una media cuadrática que puede definirse mediante las diferencias entre medias,

$$\sigma_J^2 = \frac{1}{J(J-1)} \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} (\mu_j - \mu_{j'})^2 \qquad \text{(Ec. 1.21)}$$

o bien a través de los efectos de tratamiento,

$$\sigma_J^2 = \frac{1}{J-1} \sum_{j=1}^{J} \beta_j^2 \qquad \text{(Ec. 1.22)}$$

Si los componentes de la varianza se expresan en términos de las varianzas, medias y covarianzas de las medidas de un conjunto de J jueces u observadores, se obtienen las siguientes expresiones, donde σ_I^2 y σ_e^2 son componentes de varianza aleatorios

$$\sigma_I^2 = \frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sigma_{jj'} \qquad \text{(Ec. 1.23)}$$

$$\sigma_e^2 = \frac{1}{J}\sum_{i=1}^{k}\sigma_i^2 - \frac{2}{J(J-1)}\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sigma_{jj'} \qquad \text{(Ec. 1.24)}$$

Y a partir del desarrollo de las esperanzas de las medidas cuadráticas del correspondiente análisis de varianza mixto, es sencillo estimar los tres componentes de la varianza mediante

$$\hat{\sigma}_I^2 = \frac{MC_I - MC_R}{J} \qquad \text{(Ec. 1.25)}$$

$$\hat{\sigma}_J^{'2} = \frac{MC_J}{N-1} \qquad \text{(Ec. 1.26)}$$

$$\hat{\sigma}_e^2 = MC_R \qquad \text{(Ec. 1.27)}$$

ya que

$$\hat{\rho}_C = \frac{\hat{\sigma}_I^2}{\hat{\sigma}_I^2 + \hat{\sigma}_J^2 + \hat{\sigma}_e^2} = \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\hat{\sigma}_{jj'}}{(J-1)\sum_{j=1}^{J}\hat{\sigma}_j + \sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\mu_j - \mu_{j'})^2} \qquad \text{(Ec. 1.28)}$$

Adviértase en particular la anómala estimación del componente de varianza

σ_J^2, cuya fórmula no contiene ningún término sustractivo que haga referencia al denominador de la razón *F* y asimismo emplea grados de libertad (en lugar de niveles de tratamiento) como denominador en la fórmula. Como explican Carrasco y Jover (2003: 850; véase también Fleiss, 1986) esta anomalía se debe a la necesidad de postular como modelo teórico el Modelo IIA (con el efecto de los Jueces aleatorio) con el requisito de estimar los parámetros mediante el modelo IIIA (con el efecto de los Jueces fijo) para interpretar el resultado como una medida de acuerdo.

Un tercer enfoque fue propuesto por King y Chinchilli (2001), que propusieron estimar el OCCC mediante el **estadístico U** y utilizar sus propiedades para la inferencia estadística (particularmente, errores típicos e intervalos de confianza). Mientras que los dos primeros enfoques asumen una distribución normal multivariante para las medidas del juez, el enfoque del estadístico U no requiere el supuesto de normalidad multivariante.

Similarmente, un enfoque que se basa en la utilización de **modelos de ecuaciones de estimación generalizada** (*Generalized Estimating Equations,* EEG) fue propuesto por Barnhart y Williamson (2001) y Barnhart y otros (2002, 2005). Las ventajas de este método son varias (Lin, 2007:630) entre las que merece la pena destacar por su interés estadístico las siguientes:

- el acuerdo entre los *J* jueces puede ser modelado con las covariantes ajustadas;

- el enfoque no requiere el conocimiento completo de la distribución de

los datos (es decir, es un procedimiento semiparamétrico). En contraste, el enfoque con el estadístico U es no paramétrico mientras que el enfoque de los momentos y el de los componentes de la varianza es paramétrico;tanto las estimaciones como las inferencias para las estimaciones se pueden obtener de forma simultánea.

1.5. Generalización del coeficiente de correlación de concordancia

Hay varios sentidos en que se ha postulado la generalización del coeficiente de correlación de concordancia (CCC), que fue originalmente definido para medidas cuantitativas y solamente dos jueces, y que revela el gran esfuerzo de investigación que se está realizando en este área científica. Un número monográfico del Journal of Pharmaceutical Statistics (2007) ha sido especialmente dedicado a los resultados esenciales y a la interesante prospectiva que se vislumbra en esta temática. La primera extensión es la generalización a más de dos jueces. La segunda es la generalización a medidas categóricas (nominales u ordinales). La tercera es la ampliación para incluir también medidas repetidas y datos longitudinales. Todas estas formas de generalización se tratan con más detalle a continuación.

1.5.1. Generalización a más de dos jueces

El coeficiente de correlación de concordancia se definió originalmente para 2 jueces u observadores (*A* y *B*). Utilizando *j* y *j′* para denotar a dos jueces diferentes de un conjunto de *J* jueces, CCC puede ser fácilmente generalizado al caso de más de dos jueces, como han demostrado Barnhart, Haber y Song (2002) y Carrasco y Jover (2003) mediante esta adaptación de la fórmula original (Ecuación 1.15):

$$\rho_C = \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sigma_{jj'}}{(J-1)\sum_{j=1}^{J}\sigma_j^2 + \sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\mu_j - \mu_{j'})^2} \qquad \text{(Ec. 1.29)}$$

Es conveniente notar que, en muchas ocasiones, el resultado final es similar al obtenido calculando el promedio de los coeficientes calculados para cada una de las posibles parejas de observadores del conjunto. Este resultado se conoce también con el acrónimo OCCC (*Overall Contingency Correlation Coefficient*).

1.5.2. Generalización a medidas categóricas

Asumiendo que las observaciones registradas por dos evaluadores *A* y *B* son (A_i, B_i), para todo *i* = 1, 2,…, *N*, y se seleccionan independientemente de una población bivariante con función de distribución conjunta, F_{AB} y siendo F_A y F_B las distribuciones marginales de *A* y *B* respectivamente, King y Chinchilli (2001a,b) mostraron cómo construir versiones robustas del coeficiente de correlación de concordancia entre *A* y *B* a partir de una función convexa de la distancia, $g(.)$. Para ello definieron una función integrable $g(A-B)$ respecto de F_{AB} cuyo valor esperado $E[g(A-B)]$ puede ser utilizado para obtener el acuerdo entre las variables *A* y *B*. El coeficiente de correlación de concordancia se define en este contexto en los términos siguientes:

$$\rho_C = \frac{[E_{F_A F_B} g(A-B) - E_{F_A F_B} g(A+B)] - E_{F_{AB}} g(A-B) - E_{F_{AB}} g(A+B)}{[E_{F_A F_B} g(A-B) - E_{F_A F_B} g(A+B)] + \frac{1}{2} E_{F_{AB}} [g(2\mathrm{A}) - g(2\mathrm{B})]}$$

(Ec. 1.30)

que se reduce al coeficiente de correlación de concordancia definido por Lin (1989) cuando g es la función cuadrática euclidiana de la distancia, o sea, cuando $g(z) = z^2$. El estimador del coeficiente de correlación de contingencia generalizado con una muestra de *N* ítemes, dado un par de jueces u observadores *A* y *B*, $(A_i, B_i), \ldots, (A_N, B_N)$ es asintóticamente normal con

un tamaño muestral moderado y resulta igual a

$$\tilde{\rho}_C = \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}[g(A_i - B_i) - g(A_i + B_i)] - \sum_{i}^{N}[g(A_i - B_i) - g(A_i + B_i)]}{\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}[g(A_i - B_i) - g(A_i + B_i)] + \frac{1}{2}\sum_{i=1}^{N}[g(2\mathrm{A}_i) - g(2\mathrm{B}_i)]}$$

(Ec. 1.31)

Cuando A y B siguen una distribución multinomial y ambas variables representan niveles de respuesta nominales, el coeficiente de correlación de concordancia generalizado puede reproducir el coeficiente *kappa* de Cohen (1968) definiendo como función de distancia

$$g(z) = \begin{bmatrix} 0 & , si \quad z = 0 \\ 1 & , si \quad z \neq 0 \end{bmatrix} \qquad \text{(Ec. 1.32)}$$

La aplicación de esta función de distancia simplifica la ecuación general del coeficiente de correlación de concordancia formulada por King y Chinchilli (2001) produciendo

$$\rho_g = \frac{P_1(A \neq B) - P_D(A \neq B)}{P_1(A \neq B)} \qquad \text{(Ec. 1.33)}$$

donde P_I se refiere a la probabilidad bajo independencia (bajo la cual

$F_{AB} \neq F_A F_B$) y P_D a la probabilidad bajo no independencia (y en consecuencia, $F_{AB} = F_A F_B$). Como puede observarse en la Ecuación 1.21, esta formulación tiene solo en cuenta la información referida al desacuerdo, pero es esencialmente equivalente a la formulación que utiliza solo el acuerdo, tal y como fue definida originalmente por Cohen (1968), que se tratará en el capítulo siguiente.

1.5.3. Generalización a datos longitudinales

El coeficiente de correlación de concordancia también ha sido recientemente extendido para incluir medidas repetidas (King, Chinchilli y Carrasco, 2007; Carrasco, King y Chinchilli, 2009). En primer lugar, es fácil contemplar un aumento en el número de réplicas. Partiendo de la ecuación básica (Ec. 1.18), asumamos que hay *m* réplicas para cada ítem y juez. El modelo de la Ecuación 1.18 puede formularse entonces mediante

$$Y_{ijl} = \mu + \alpha_i + \beta_j + e_{ijl} \qquad \text{(Ec. 1.34)}$$

y, aunque la expresión para el coeficiente CCC es la misma, el proceso de estimación mejora sensiblemente porque los componentes de la varianza se estiman con mayor eficiencia al ser la varianza del estimador del coeficiente más pequeña (Carrasco y Jover, 2003).

Sin embargo, si la estructura de los datos indica que depende de una covariante (por ejemplo, son medidas repetidas registradas a lo largo del tiempo), las réplicas ya no pueden considerarse como tales y la estimación del coeficiente de correlación de contingencia a partir de la Ecuación 1.34 sería sesgada.

En este caso, siendo *m* el efecto (fijo) del tiempo (para $m=1,...,M$), Carrasco, King y Chinchilli (2009) han desarrollado una especificación del coeficiente de correlación intraclase a partir del siguiente modelo lineal mixto,

$$Y_{ijm}=\mu+\alpha_i+\beta_j+\gamma_m+(\alpha\beta)_{ij}+(\alpha\gamma)_{\Im}+(\beta\gamma)_{jm}+e_{ijm} \qquad \text{(Ec. 1.35)}$$

donde μ es la media global, α_i es el efecto aleatorio del ítem u objeto que se asume que se distribuye según $\alpha_i \sim N(0,\sigma_\alpha)$, β_j es el efecto fijo del juez u observador, γ_m es el efecto fijo del tiempo, $(\alpha\beta)_{ij}$ es el efecto aleatorio de la interacción ítem-juez que se asume distribuida según $(\alpha\beta)_{ij} \sim N(0,\sigma_{\alpha\beta})$, $(\alpha\gamma)_{\Im}$ es el efecto aleatorio de la interacción ítem-tiempo que se asume distribuida según $(\alpha\gamma)_{\Im} \sim N(0,\sigma_{\alpha\gamma})$, $(\beta\gamma)_{jm}$ es el efecto aleatorio de la interacción juez-tiempo que se asume distribuida según $(\beta\gamma)_{jm} \sim N(0,\sigma_{\beta\gamma})$ y e_{ijm} es error aleatorio. El modelo de la Ecuación 1.35 puede formularse también (Verbeke y Mollenberghs (2000) mediante

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{b}+\boldsymbol{e} \qquad \text{(Ec. 1.36)}$$

donde **Y** es el vector de respuesta, **X** es la matriz de diseño de los efectos fijos, que contiene los efectos de juez, tiempo e interacción juez x tiempo y **Z** es la matriz de diseño de los efectos aleatorios, que contiene los efectos de ítem, ítem x juez e ítem x tiempo. Los vectores aleatorios **b** y **e** se asumen independientes y distribuídos según $\boldsymbol{b} \sim MVN(\mathbf{0}, \boldsymbol{G})$ y $\boldsymbol{e} \sim MVN(\mathbf{0}, \sigma_e \boldsymbol{R})$, siendo R una matriz $m \times m$ que explica la correlación longitudinal entre residuales.

El coeficiente de correlación de contingencia para medidas repetidas se estima, siguiendo a Carrasco, King y Chinchilli (2009:94) mediante el coeficiente de correlación intraclase apropiado

$$CCC = \frac{\sigma_{\alpha}^2 + \sigma_{\alpha\gamma}^2}{\sigma_{\alpha}^2 + \sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta}^2 + \sigma_{\beta\gamma}^2 + \sigma_e^2} \qquad \text{(Ec. 1.37)}$$

Esta versión del coeficiente de correlación intraclase asume efectos fijos para jueces y tiempo y efectos aleatorios para ítem. Otras versiones del coeficiente de correlación intraclase se han tratado en Vangeneugden y otros (2005).

1.6. El coeficiente r_{WG} y sus alternativas

En el ámbito de la psicología social y de las organizaciones es bastante común evaluar un ítem por parte de un conjunto de evaluadores sobre una escala de Likert. Es usual utilizar un coeficiente de acuerdo de carácter básico (denominado r_{wg}), o alguna de sus alternativas, que prácticamente resultan desconocidas en el resto de las áreas de la psicología. De hecho, resulta llamativo saber que ni siquiera se cita este coeficiente en textos específicos sobre análisis de acuerdo para las ciencias biológicas y sociales tales como Shoukri (2004) o von Eye y Mun (2005), pero puede constatarse que el coeficiente es bastante citado (casi 1000 referencias hemos encontronado en revistas referenciadas del ISI desde 1984) en campos tan dispares como la dirección estratégica y la enfermería.

El coeficiente r_{WG} fue inicialmente formulado por Finn (1970) como un estimador de la fiabilidad con la que un grupo de jueces clasifican estímulos en categorías utilizando la ecuación general:

$$r_{WG} = 1 - \frac{V(observada)}{V(esperada)} \qquad \text{(Ec. 1.38)}$$

Por ejemplo, supongamos que 10 observadores deben clasificar un determinado ítem sobre una escala de 5 categorías, y de ellos 4 lo clasifican en la primera, 4 en la segunda y 2 en la tercera. El acuerdo perfecto entre los jueces se produciría si hubiera varianza cero entre las clasificaciones, mientras

que sería igual a 0 si las clasificaciones fueran totalmente aleatorias, en cuyo caso la distribución de las clasificaciones sería rectangular con una frecuencia de 2 en cada categoría. En tal caso, la varianza de las clasificaciones esperadas sería en realidad una constante (2), mientras que su varianza observada seria . 622 y por tanto el coeficiente resultaría igual a $r_{WG} = 1 - (.622/2) = .689$.

James, Demaree y Wolf (1984) redefinieron el coeficiente propuesto por Finn (1970) para el caso cuantitativo, y propusieron utilizar una de las dos formas alternativas que se describen a continuación:

1) La primera forma utiliza un conjunto de jueces que valora un único ítem en una única variable sobre una escala de intervalo,

$$r_{WG} = 1 - \frac{S_X^2}{\sigma_E^2} \qquad \text{(Ec. 1.39)}$$

donde S_X^2 es la varianza observada sobre la variable *X* tomada sobre *J* diferentes observadores y σ_E^2 es la varianza esperada cuando existe una completa falta de acuerdo entre los jueces. Básicamente, la varianza esperada es la que se obtendría si todos los jueces respondieran al azar cuando evalúan el ítem y por ello se trata de una distribución teórica (esto es, no está determinada empíricamente) y condicional (es decir, asume respuesta aleatoria). La determinación de la forma de la distribución esperada es uno de los factores que más complica la utilización del coeficiente r_{WG}. En general, en presencia

de falta de acuerdo entre los observadores, la distribución que se asume es la distribución rectangular o uniforme, para la que

$$\sigma_E^2 = \frac{K^2 - 1}{12} \qquad \text{(Ec. 1.40)}$$

donde *K* es el número de categorías de respuesta (por ejemplo, sobre una escala de Likert). Sin embargo, James, Demaree y Wolf (1984) estimularon el empleo de otras distribuciones alternativas, tales como las causadas por sesgo de respuesta (por ejemplo, sesgo de tolerancia o sesgo de centralidad) o cualesquier otras distribuciones nulas.

2) La segunda forma es una generalización de la primera donde un conjunto de J jueces valoran un conjunto de I ítemes paralelos de una única variable sobre una escala de intervalo,

$$r_{WG(I)} = \frac{I\left[(1 - \bar{S}_{X_I})/\sigma_E^2\right]}{I\left[(1 - \bar{S}_{X_I})/\sigma_E^2\right] + \bar{S}_{X_I}/\sigma_E^2} \qquad \text{(Ec. 1.41)}$$

donde $\bar{S}_{X_I}$ es la media de las varianzas observadas para los I ítemes esencialmente paralelos y el resto de la fórmula es similar.

El ejemplo numérico que se presenta en la Tabla 1.7 (tomado de James, Demaree y Wolf, 1984: 88) ilustra los cálculos requeridos para la valoración de tres ítemes por 10 observadores utilizando diferentes alternativas de respuesta.

Aplicando las ecuaciones 1.35 y 1.36, en el primer caso el número de alternativas es K = 5 y el coeficiente resulta igual a $r_{WG(1)}$=.130, en el segundo caso el número de alternativas es K = 7 y el coeficiente es $r_{WG(2)}$=.940 y en el tercer caso el número de alternativas es K = 9 y el coeficiente es $r_{WG(3)}$=.920.

Tabla 1.7.
Juicios emitidos por 10 jueces para (1) ítemes con diferentes alternativas de respuesta y (2) 3 ítemes con 7 alternativas de respuesta

Jueces	*K = 5*	*K = 7*	*K = 9*	*K = 7*
1	5.000	6.000	4.000	5.000
2	2.000	6.000	4.000	4.000
3	3.000	7.000	4.000	4.670
4	5.000	7.000	5.000	5.670
5	2.000	7.000	5.000	4.670
6	3.000	7.000	5.000	5.000
7	1.000	7.000	5.000	4.330
8	4.000	6.000	5.000	5.000
9	3.000	7.000	6.000	5.330
10	4.000	7.000	6.000	5.670
Media $(\bar{X})$	3.200	6.700	4.900	
Varianza observada $(S^2_{X_I})$	1.730	.230	.540	.830
Varianza esperada $(\sigma_E\ 2)$	2.000	4.000	6.670	4.000
$r_{WG(I)}$	.130	.940	.920	.970

Los mismos datos pueden también servir para ilustrar la aplicación de la Ecuación 1.37, asumiendo ahora que 10 jueces (filas) valoran 3 ítemes

(columnas) sobre una escala común de 7 alternativas de respuesta tal y como se muestra en la última columna. Los resultados que interesan son obviamente el promedio de las varianzas observadas $(\bar{S}^2_{X_I} = (1.730 + .230 + .540) = .830)$ y la varianza esperada, que asumiendo una distribución uniforme con 7 alternativas de respuesta es igual a 2. En consecuencia, el coeficiente de acuerdo para el conjunto de los datos es $r_{WG(I)} = .970$.

Schmidt y Hunter (1989) criticaron el uso de los índices r_{WG} y $r_{WG(I)}$ que se hizo en un trabajo de Kozlowski y Hults (1987), fundamentalmente debido a la confusión semántica que surgió como consecuencia de denominarles "coeficientes de fiabilidad" en lugar de la más apropiada denominación de "coeficientes de acuerdo", tal y como señalamos en la Introducción de este trabajo, ya que el coeficiente no tenía cabida en la teoría clásica de la fiabilidad al valorar solamente un ítem. El tema se solventó poco después cuando Kozlowski y Hattrup (1992) y James, Demaree y Wolf (1993) reconocieron que ambos coeficientes eran de hecho medidas de acuerdo y no medidas de fiabilidad.

En el ejemplo utilizado, los coeficientes r_{WG} y $r_{WG(I)}$ se utilizaron con escalas de Likert de 5 o más categorías, pero son perfectamente aplicables con escalas numéricas (en particular, escalas de intervalo). Más recientemente, Burke, Finkelstein y Dusig (1999) propusieron una nueva medida de acuerdo de la misma familia desarrollada para ser utilizada con múltiples observadores que valoran un ítem sobre una variable usando una escala de intervalo. Su interés era estimar el acuerdo en la métrica de la escala original del ítem. Dado un conjunto de jueces que valoran un único ítem, su propuesta presentaba dos formas posibles según se estime con la media $(AD_{M(I)})$ o preferentemente

con la mediana $(AD_{MDN(I)})$

$$AD_{M(I)} = \frac{\sum_{j=1}^{J} |X_{ij} - \bar{X}_i|}{J} \qquad \text{(Ec. 1.42)}$$

$$AD_{MDN(I)} = \frac{\sum_{j=1}^{J} |X_{ij} - MDN_i|}{J} \qquad \text{(Ec. 1.43)}$$

El resultado puede ser con facilidad generalizado al caso de *I* ítemes paralelos valorados por *J* evaluadores mediante

$$AD_{M(I)} = \frac{\sum_{i=1}^{I} AD_{M(i)}}{I} \qquad \text{(Ec. 1.44)}$$

$$AD_{MDN(I)} = \frac{\sum_{i=1}^{I} AD_{MDN(i)}}{I} \qquad \text{(Ec. 1.45)}$$

Finalmente, la más reciente propuesta de la familia es el coeficiente a_{WG}, que fue sugerido por Brown y Hauenstein (2005) para superar las limitaciones encontradas en la familia de coeficientes r_{WG} y más concretamente las siguientes:

- los coeficientes basados en r_{WG} son dependientes de la escala y por tanto pueden ser distintos dependiendo de si se emplea una escala Likert de 5, 7 o 9 puntos;

- el tamaño muestral, por ejemplo el número de jueces, influye de forma determinante en los valores del coeficiente r_{WG} que en consecuencia influye sobre la interpretabilidad de los resultados;

- los investigadores asumen que la distribución uniforme es válida para todos los casos, pero este supuesto es a todas luces incorrecto.

El índice propuesto de Brown y Hauenstein (2005) sigue la lógica del clásico *kappa* de Cohen, que ellos extendieron a la situación estándar de múltiples jueces que valoran un único ítem utilizando una escala de intervalo, mediante

$$a_{WG} = 1 - \frac{2 * S_X^2}{\left[(H+L) * \bar{X} - \bar{X}^2 - (H * L)\right] * \left[J/(J-1)\right]} \qquad \text{Ec. 1.46)}$$

donde $\bar{X}$ es la media de la valoración del conjunto de los jueces, H es el valor máximo posible de la escala, L es el mínimo posible, J es el número de jueces y S_X^2 es la varianza observada sobre X. Y similarmente a lo que sucede con los índices $r_{WG(I)}$ y $\mathrm{AD}_{\mathrm{M(I)}}$ y $AD_{MDN(I)}$ existe una versión multi-ítem para cuando I ítemes paralelos son valorados por J evaluadores

mediante

$$a_{WG}(I) = \frac{\sum_{i=1}^{I} a_{WG(I)}}{I} \qquad \text{(Ec. 1.47)}$$

Un interesante tutorial de LeBreton y Senter (2008) resume las principales cuestiones que plantea el uso de esta familia de índices y su reciente popularidad en ciertas áreas de la psicología social debido, en parte, al papel creciente que están jugando las técnicas de modelado multinivel (en concreto, los modelos lineales mixtos y los modelos de ecuaciones estructurales multinivel) en la psicología de las organizaciones, donde los índices de acuerdo se utilizan con frecuencia para justificar la agregación de datos de nivel inferior en compuestos más complejos de nivel superior (por ejemplo, medidas de afecto a nivel individual pueden utilizarse para medir el tono afectivo del grupo).

Capítulo 2

Datos nominales: dos jueces, dos categorías

2.1. Introducción y notación

La mayor parte de la investigación que ha tratado sobre el tema del acuerdo entre jueces u observadores (*rater agreement*) versa sobre la derivación de alguna medida descriptiva de acuerdo. Con datos categóricos, el caso más simple consiste en evaluar el acuerdo entre 2 observadores, que llamaremos juez *A* y juez *B*, respectivamente, sobre una escala binaria (nominal). Por ejemplo, supongamos que se pide a $J = 2$ jueces que clasifiquen cada uno de un conjunto de N ítemes/objetos o sujetos sobre una escala de $K = 2$ categorías; para la que se requiere una respuesta "Sí" ($k = 1$) ó "No" ($k = 2$). El resultado

de las respuestas de los 2 jueces puede representarse en una tabla de contingencia 2×2, conocida asimismo en el contexto de estudios de fiabilidad como **tabla de acuerdo** (e.g. Fleiss, Levin y Paik, 2003), de la forma siguiente:

Tabla 2.1

		Juez B		
		1: Sí	*2: No*	
Juez A	*1: Sí*	n_{11}	n_{12}	n_{1+}
	2: No	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	$n_{++}= N$

donde n_{11}, ... , n_{22} representan el número de ítemes que corresponde a cada una de las posibles combinaciones de observadores por categorías de respuesta (frecuencias conjuntas), n_{1+} , n_{2+} son el número de ítemes que corresponde a cada una de las dos categorías para el juez *A* (sus frecuencias marginales), n_{+1} , n_{+2} son el número de ítemes que corresponde a cada una de las dos categorías para el juez *B* (sus frecuencias marginales) y $n_{++} = N$ es el número total de ítemes –tamaño muestral– utilizado. Emplearemos como notación para tablas de acuerdo el número de ítemes (n) distinguiendo como subíndices la categoría seleccionada por el juez *A* (que llamaremos k) de la categoría seleccionada por el juez *B* (que llamaremos k').

Siendo $P(A=k\,,B=k')=\pi_{kk'}$ (para k = 1, 2 y k' = 1, 2), la probabilidad

de cada uno de los posibles patrones de respuesta conjunta entre ambos observadores, $P(A=k)=\pi_k$ (para k = 1, 2), la probabilidad de respuesta a cada categoría para el juez A y $P(B=k')=\pi_{k'}$ (para k' = 1, 2), la probabilidad de respuesta a cada categoría para el juez B, entonces (A, B) asume una distribución binomial bivariante para las frecuencias conjuntas de la tabla de acuerdo resultante y los estimadores máximo-verosímiles son, respectivamente,

$$\hat{\pi}_{kk'}=\frac{n_{kk'}}{N}$$
$$\hat{\pi}_{k}=\frac{n_{k+}}{N}$$
$$\hat{\pi}_{k'}=\frac{n_{+k'}}{N}$$

De forma similar, también puede visualizarse una tabla de acuerdo 2×2 utilizando proporciones. En concreto, transformando las frecuencias $n_{kk'}$ en proporciones, la Tabla 2.1 sería

Tabla 2.2

		Juez B		
		1: Sí	*2: No*	
Juez A	*1: Sí*	p_{11} (a)	p_{12} (b)	p_{1+} (p_1)
	2: No	p_{21} (c)	p_{22} (d)	p_{2+} (q_1)
		p_{+1} (p_2)	p_{+2} (q_2)	$p_{++} = 1$

donde los valores entre paréntesis (a, b, c y d) representan probabilidades conjuntas de la tabla de acuerdo, p_1, q_1 representan probabilidades marginales de cada respuesta para el juez A y p_2, q_2 representan probabilidades marginales para cada respuesta del juez B, que se usan con el objeto de facilitar la interpretación en ciertas situaciones.

En cualquier caso, es pertinente utilizar $p_o = p_{11} + p_{22} = a + d$, o dicho de otro modo, la suma de las probabilidades diagonales de la tabla de acuerdo, que representa la proporción total del acuerdo existente entre ambos evaluadores, como medida descriptiva de acuerdo, en la que p_o se refiere a la **probabilidad de acuerdo observada** (*Raw Observed Agreement*, ROA), que varía entre los límites 0 –cuando existe desacuerdo absoluto entre ambos jueces– y 1 – cuando el acuerdo entre observadores es perfecto–. En el pasado, la ROA se ha utilizado como medida bruta del acuerdo entre jueces, y en ocasiones se ha denominado también **coeficiente de igualación simple** (*simple matching*; véase Gwett, 2002:51; Shoukri, 2004:24). La varianza de la

probabilidad de acuerdo observada es $V(p_o)=p_o(1-p_o), 0\leq p_o\leq 1.$ Una atenta mirada a una tabla de acuerdo 2×2 conduce a la conclusión de que una condición necesaria, aunque no suficiente, para que el acuerdo sea perfecto es que $p_{1+}=p_{+1}$ –y por tanto, $p_{2+}=p_{+2}$ –, i.e., que las probabilidades marginales de fila y columna sean iguales, y como consecuencia que *A* y *B* tengan la misma distribución marginal.

Lo que resulta inadecuado en el uso de la probabilidad de acuerdo observada ROA, como apuntaron investigadores pioneros en el tema (e.g. Fleiss, Levin y Paik, 2003, Hsu y Feld, 2003) es que cabe esperar que cierta proporción del acuerdo entre los jueces se produzca por azar. La solución más lógica consiste en determinar –y consecuentemente corregir– el "supuesto" grado de acuerdo producido por azar mediante la introducción de un nuevo parámetro: la probabilidad de acuerdo esperada por azar (*Proportion of Chance or Expected Agrement*, p_e).

2. 2. Medidas de acuerdo corregidas del azar

Dada una tabla de acuerdo entre 2 (o más) observadores, una **medida general de acuerdo corregido del azar** (*Random Corrected Agreement,* RCA), que distingue la probabilidad de acuerdo observada (p_o) de la probabilidad de acuerdo esperada por azar (p_e), se define con la formulación siguiente:

$$RCA = \frac{p_o - p_e}{1 - p_e} \qquad \text{(Ec. 2.1)}$$

Ésta es la solución propuesta en todas de las medidas descriptivas que se analizan en este capítulo, pero en cada una de ellas se utiliza una definición diferente de *Random Corrected Agreement* (RCA) para obtener el parámetro p_e, en la cual se especifica alguna fórmula de corrección del azar para definir dicho parámetro; p_o es la probabilidad de acuerdo observada y $p_o - p_e$ es el exceso de probabilidad de acuerdo observada que se asume que incluye p_o. La diferencia debe ser ponderada con $1 - p_o$, el valor máximo posible de la probabilidad de acuerdo esperada por azar. Los valores de RCA deben caer dentro del intervalo $\left[\frac{-\sum p_e}{(1-\sum p_e)}; +\frac{\sum p_o}{(1-\sum p_e)}\right]$, donde $RCA = \frac{-\sum p_e}{(1-\sum p_e)} > -1$ es el límite inferior asociado con el desacuerdo perfecto, de modo que cuando RCA = 0 implica que la probabilidad de acuerdo observado es igual a la probabilidad de acuerdo esperada por azar y $RCA = \frac{\sum p_o}{(1-\sum p_e)} \leq 1$ es el límite superior asociado con el acuerdo perfecto, en cuyo caso la probabilidad de acuerdo esperado por azar es cero.

Hay otras alternativas para obtener medidas descriptivas de acuerdo, pero puesto que no persiguen un objetivo estricto de valoración del acuerdo entre jueces, no se abordarán en este trabajo (véase e.g. Blackman y Koval, 1993;

Goodman y Kruskal, 1954; Fleiss, Levin y Paik, 2003; Shoukri, 2004).

Las cuatro medidas descriptivas más comunes tratadas en la literatura sobre acuerdo entre observadores se basan en la fórmula general RCA (Ecuación 2.1) y asumen independencia entre los evaluadores en el proceso de clasificación. Las diferencias entre ellos se deben a la definición específica de acuerdo esperado por azar p_e. Tales medidas son las siguientes:

1) el coeficiente de acuerdo σ propuesto por Bennett y otros (1954),
2) el coeficiente de acuerdo π propuesto por Scott (1955),
3) el coeficiente de acuerdo κ propuesto por Cohen (1960),
4) el coeficiente de acuerdo γ propuesto por Gwett (2008).

Estos coeficientes se computan a partir de una muestra aleatoria de *N* ítemes seleccionados de una población supuestamente infinita de ítemes; por tanto, se asume variabilidad de muestreo. La variabilidad muestral se desprende del hecho de que dos muestras seleccionadas de la población de interés probablemente produzcan diferentes estimadores de los coeficientes. En consecuencia, es necesario tomar la variabilidad muestral en cuenta cuando se realiza algún tipo de inferencia estadística con los coeficientes de acuerdo.

En este capítulo utilizaremos un ejemplo ilustrativo para el cálculo de los cuatro coeficientes que se muestra en la tabla de acuerdo siguiente, que llamaremos Ejemplo 2.1:

Tabla 2.3: Ejemplo 2.1

		Juez B		
		1: Sí	*2: No*	
Juez A	*1: Sí*	24 (.240)	11 (.110)	35 (.350)
	2: No	3 (.030)	62 (.620)	65 (.650)
		27 (.270)	73 (.73)0	100 (1)

Puesto que N = 100, la probabilidad de acuerdo observado es $p_o = .240 + .620 = .860$ (es decir, el porcentaje total de acuerdo observado corresponde al 86% de los ítemes).

2.2.1. El coeficiente σ de Bennett y otros (1954)

Una primera solución fue originalmente propuesta por Bennett, Alpert y Goldstein (1954:307) utilizando como definición de acuerdo esperado por azar un valor fijo, la inversa del número de categorías (K). Zwick (1988) ha apuntado que esta solución ha sido varias veces rebautizada con nombres diferentes pero bajo la misma definición. Como consecuencia, el coeficiente σ se ha convertido en una medida de acuerdo consolidada después de haber sido profundamente tratada en otros trabajos (e.g. Holley y Guilford (1964),

Maxwell y Pilliner (1968), Janson y Vegelius (1979), Maxwell (1977) y Brennan y Prediger (1981)). Así, Bennett, Alpert y Goldstein (1954) lo llamaron "puntuación S"; Holley y Guilford (1964), que fueron los primeros en considerarlo como un coeficiente para evaluar la fiabilidad entre jueces, lo llamaron "coeficiente G"; Maxwell y Pilliner (1968) y Maxwell (1977) lo denominaron "coeficiente de error aleatorio"; Janson y Vegelius (1979) lo llamaron "coeficiente C", mientras que Brennan y Prediger (1981) lo denominaron κ_n (véase en Zwick, 1988, y en Hsu y Feld, 2003, una explicación más detallada de esta polémica). Ante tanta diversidad, hemos optado por referirnos a esta solución como coeficiente σ en honor a la propuesta original de Bennett y otros. En todos los casos, la probabilidad de acuerdo corregido del azar, siendo $K = 2$, es $p_e^{\sigma} = 1/2 = .500$.

Esta forma de corrección del azar asume que los evaluadores clasifican uniformemente los ítemes en categorías y por tanto se basa en una distribución uniforme de los ítemes. Para $K = 2$ categorías, siendo $p_e^{\sigma} = .500$ se simplifica el cálculo del coeficiente mediante

$$\hat{\sigma} = \frac{p_o - p_e^{\sigma}}{1 - p_e^{\sigma}} = 2(a+d) - 1 = 2\,p_o - 1 \qquad \text{(Ec. 2.2)}$$

Para los datos empíricos de la tabla de acuerdo del Ejemplo 2.1 que se muestra en la Tabla 2.3, $\hat{\sigma} = (.860 - .500)/.500 = .720$, según la primera parte de la fórmula, o también $\hat{\sigma} = (2)(.240 + .620) - 1 = .720$, según la segunda y la tercera parte de la fórmula. Un estimador de la varianza muestral del coeficiente (Gwett, 2001b:69) puede obtenerse aplicando

$$V(\hat{\sigma}) = 4 \left(\frac{1-f}{N-1}\right) p_o(1-p_o) \qquad \text{(Ec. 2.3)}$$

donde f se refiere a la fracción de muestreo de los objetos de la población supuestamente infinita de objetos. En general, la fracción de muestreo se considera prácticamente insignificante y no afecta en la práctica al resultado.

Para los datos del Ejemplo 2.1, $V(\hat{\sigma}) = (4)(1/99)(.860)(.140) = .005$. Una prueba apropiada para evaluar la significación estadística del coeficiente σ y establecer intervalos de confianza sobre el mismo (véase Gwett, 2001b:69) es la siguiente:

$$z_{\hat{\sigma}} = \frac{\hat{\sigma} - .500/N}{\sqrt{V(\hat{\sigma})}} \qquad \text{(Ec. 2.4)}$$

que se distribuye normalmente y debe por tanto ser comparada con los percentiles de una distribución normal. Para los datos de la tabla de acuerdo del Ejemplo 2.1, $z_{\hat{\sigma}} = [.720 - (.5/100)]/.070 = 10.210; P < .0100;$ y por tanto, el coeficiente resulta estadísticamente significativo.

2.2.2. El coeficiente π de Scott (1955)

Una segunda solución fue propuesta por Scott (1955), en respuesta a la planteada anteriormente por Bennett y otros (1954), utilizando como probabilidad de acuerdo esperado el cuadrado de la media de las probabilidades marginales de fila y columna. Con K = 2 categorías, la probabilidad de acuerdo esperado por azar sería

$$p_e^{\pi} = \left(\frac{p_1 + p_2}{2} \right)^2 + \left(\frac{q_1 + q_2}{2} \right)^2 \qquad \text{(Ec. 2.5)}$$

Esta formulación asume que los observadores clasifican los objetos utilizando una distribución homogénea común. Scott (1955) supuso que la probabilidad de acuerdo por azar dependería de las probabilidades de clasificación –marginales– de los jueces, pero desgraciadamente en ciertos casos representa una inadecuada aproximación a la probabilidad de acuerdo por azar. Para los datos del Ejemplo 2.1, la aplicación de tal fórmula arroja como probabilidad esperada por azar

$$p_e^{\pi} = \left(\frac{.350 + .270}{2} \right)^2 + \left(\frac{.650 + .730}{2} \right)^2 = .057$$

El estimador del grado de acuerdo resultante se define entonces, para los datos de la Tabla 2.3, mediante

$$\hat{\pi} = \frac{p_o - p_e^{\pi}}{1 - p_e^{\pi}} = \frac{.860 - .572}{1 - .572} = .670$$

Si el número de ítemes es alto, entonces la varianza del coeficiente puede ser aproximada mediante

$$V(\hat{\pi}) = \left(\frac{N}{N-1}\right)\left(\frac{1-f}{N}\right)\left(\frac{p_o(1-p_o)}{1-p_e^{\pi}}\right) \qquad \text{(Ec. 2.6)}$$

donde f representa la fracción de muestreo de los ítemes utilizados respecto del total (N) supuestamente infinito de ítemes. De hecho, en la práctica la fracción de muestreo es irrelevante y la aproximación es equivalente a la propuesta originalmente en el trabajo de Scott (1955),

$$V(\hat{\pi}) = \frac{p_o(1-p_o)}{(N-1)(1-p_e^{\pi})^2} \qquad \text{(Ec. 2.7)}$$

Para los datos de la tabla de acuerdo del Ejemplo 2.1,

$$\mathrm{V}(\hat{\pi}) = \frac{(.860)(.140)}{(99)(1-.570)^2} = .007$$

Para probar la significación estadística y establecer intervalos de confianza sobre el coeficiente π puede utilizarse

$$z_{\hat{\pi}} = \frac{\hat{\pi}}{\sqrt{V(\hat{\pi})}} \qquad \text{(Ec. 2.8)}$$

que se distribuye también de forma normal y resulta estadísticamente significativo: $z_{\hat{\pi}} = .670/.080 = 8.400$ *;* $P < .0100$.

Una característica controvertida de p_e^{π} es que su valor mínimo es .500 y su valor máximo es 1. Cuando p_e^{π} alcanza el valor máximo –que ocurre cuando todos los ítemes o sujetos se clasifican en la diagonal principal de la tabla de acuerdo– el coeficiente es indefinido, obteniendo un valor de 0 cuando en realidad debería ser 1, ya que todas las posibles valoraciones de ambos evaluadores serán en realidad acuerdo debido al azar; característica que ha generado mucha polémica (e.g. Zwick, 1988; Hsu y Feld, 2003).

En consecuencia, el coeficiente π solamente es adecuado cuando el nivel de acuerdo entre ambos evaluadores es bajo. Esto se debe al hecho de que la expresión para calcular p_e^{π} aproxima la probabilidad de acuerdo bajo el poco realista supuesto de valoración aleatoria por parte de los evaluadores. Sin embargo, el contexto metodológico donde se utiliza este coeficiente es óptimo para probar su significación estadística (Gwett, 2001b:63).

2.2.3. El coeficiente κ de Cohen (1960)

Una tercera solución fue propuesta por Cohen (1960), quien criticó la propuesta de Scott (1955) en lo relativo al supuesto de homogeneidad de las probabilidades de clasificación entre los observadores. Utilizando un enfoque distinto al problema, Cohen sugirió un estimador del grado de acuerdo que llamó κ (*kappa*), que presenta la misma forma que π y define como fórmula de corrección del azar (véase Tabla 2.2)

$$p_e^{\kappa} = p_1 p_2 + q_1 q_2 \qquad \text{(Ec. 2.9)}$$

y por tanto asume que cada observador clasifica los sujetos usando su propia distribución.

Para los datos empíricos de la Tabla 2.3, la probabilidad de acuerdo estimada es $p_e^{\kappa} = (.350)(.270) + (.650)(.730) = .570$. El estimador κ se define para los mismos datos mediante

$$\hat{\kappa} = \frac{p_o - p_e^{\kappa}}{1 - p_e^{\kappa}} = \frac{.860 - .570}{1 - .570} = .680$$

Desde su introducción en 1960, el coeficiente κ (*kappa*) se ha convertido en la medida por excelencia para la evaluación del acuerdo entre jueces. Durante casi medio siglo, la práctica totalidad de la investigación aplicada de

las ciencias biológicas y sociales reporta como medida genérica de acuerdo el coeficiente $\kappa)$. En un trabajo de revisión, Hsu y Feld (2003:206) comprobaron que *kappa* había sido citado en artículos de la *American Psychological Association* y del *Social Scisearch* más de 2000 veces en el periodo comprendido desde 1989 a 2002.

Se han propuesto varias fórmulas para calcular la varianza del coeficiente $\kappa)$. Para el caso $2\times2,$ Gwett (2002c) propuso la siguiente:

$$V(\hat{\kappa})=\left(\frac{1-f}{N-1}\right)\left(\frac{p_o(1-p_o)}{(1-p_e^{\kappa})^2}\right) \qquad \text{(Ec. 2.10)}$$

donde f representa la fracción de muestreo, que en la práctica resulta prácticamente cero. Para los datos del Ejemplo 2.1,

$$\mathrm{V}(\hat{\kappa})=\left(\frac{1}{99}\right)\left[\frac{(.860)(.140)}{(1-.570)^2}\right]=.0065$$

Es posible también evaluar la significación estadística del coeficiente, utilizando una distribución normal, mediante

$$\mathrm{z}_{\hat{\kappa}}=\frac{\hat{\kappa}}{\sqrt{\mathrm{V}(\hat{\kappa})}} \qquad \text{(Ec. 2.11)}$$

Para los datos del Ejemplo 2.1, $z_{\hat{\kappa}} = .680/.080 = 8.500; p < .0100$ y por tanto resulta también estadísticamente significativo.

En general, el coeficiente κ presenta propiedades estadísticas óptimas como medida de acuerdo:

- En primer lugar, cuando el acuerdo observado (p_o) es igual al acuerdo esperado por azar (p_e^{κ}) entonces $\kappa = 0$.

- En segundo lugar, tomará su valor máximo de 1 si y sólo si el acuerdo es perfecto (esto es, $p_o = 1$ y $p_e^{\kappa} = 0$).

- Finalmente, κ nunca puede ser menor de -1. Sin embargo, sus límites superior e inferior son función de las probabilidades marginales. Así, κ tomará el valor +1 si y solo si las probabilidades marginales son exactamente iguales y todas las casillas no diagonales son cero. Si $p_e^{\kappa} = .500$ entonces el valor mínimo es -1; en cualquier otro caso, el valor mínimo se encuentra entre -1 y Landis y Koch (1977) han caracterizado diferentes rangos de valores para κ con respecto al grado de acuerdo con el que se asocian.

En general, valores mayores de .750 representan un acuerdo excelente una vez corregido el efecto del azar, mientras que valores por debajo de .400 representan un acuerdo inaceptable y valores superiores a .400 e inferiores a .

750 representan un acuerdo entre moderado y aceptable, una vez corregido el efecto del azar.

2.2.4. El coeficiente γ de Gwett (2008)

Otra solución ha sido propuesta recientemente por Gwett (2002c, 2008), quien define la probabilidad de acuerdo por azar –tomando el promedio de la primera probabilidad marginal de fila y columna $\hat{\pi}_1 = (p_{1+} + p_{+1})/2$ – como

$$p_e^{\gamma} = 2\hat{\pi}_1(1-\hat{\pi}_1) \qquad \text{(Ec. 2. 12)}$$

y por tanto se asume que, en su respuesta a los ítemes, los jueces utilizan también una distribución homogénea. Para los datos del Ejemplo 2.1, $\hat{\pi}_1 = (.350+.270)/2 = .310$ y por tanto $p_e^{\gamma} = (2)(.310)(.690) = .430$. En consecuencia, aplicando la ecuación general RCA, el coeficiente de acuerdo γ resulta igual a

$$\gamma = \frac{p_o - p_e^{\gamma}}{1 - p_e^{\gamma}} = \frac{.860 - .430}{1 - .430} = .750$$

Un estimador apropiado de la varianza de γ fue asimismo propuesto recientemente por Gwett (2002, 2008),

$$V(\hat{\gamma}) = \left(\frac{1-f}{N-1}\right)\left[\frac{p_o(1-p_o)}{(1-p_e^{\gamma})^2}\right] \qquad \text{(Ec. 2.13)}$$

donde *f* se refiere a la fracción de muestreo de *N* ítemes de una población . Asumiendo la población infinita, la fracción de muestreo es en la práctica irrelevante. Para los datos del Ejemplo 2.1, la varianza de γ es

$$V(\hat{\gamma}) = \left(\frac{1}{99}\right)\left[\frac{(.860)(.140)}{(1-.430)^2}\right] = .0037$$

Y de modo similar a los coeficientes anteriores, es posible también someter a prueba la significación estadística del coeficiente γ mediante

$$z_{\hat{\gamma}} = \frac{\hat{\gamma}}{\sqrt{V(\hat{\gamma})}} \qquad \text{(Ec. 2.14)}$$

Para los datos del Ejemplo 2.1, $z_{\hat{\gamma}} = .760/.060 = 12.670\,; P < .0100$ que por lo demás también resulta estadísticamente diferente de cero.

2.3. Contexto de acuerdo y contexto de asociación

Bloch y Kraemer (1989) alertaron contra el uso indiscriminado de las medidas de acuerdo y la interpretación derivada de tal uso en muchos trabajos de investigación. Para aclarar posiciones, distinguieron dos contextos en los que la interpretación de una medida de acuerdo debe fundamentarse: el contexto de acuerdo y el contexto de asociación. Los datos ficticios registrados por 4 jueces (*A, B, C, D*) a una misma cuestión se muestran en la Tabla siguiente:

Tabla 2.4: Datos ficticios

A	*B*	*C*	*D*
1	1	2	2
2	2	3	4
3	3	4	6
4	4	5	8
5	5	6	10

Asumiendo que las respuestas se miden en la misma escala, cabría concluir que existe asociación perfecta entre todas las medidas, pero también que *A* y *B* están en perfecto acuerdo porque son iguales, mientras que *A* y *C* no están en perfecto acuerdo porque se observa un sesgo en la respuesta del segundo; *A* y *D* tampoco están en perfecto acuerdo ya que hay un cambio de escala en la respuesta del segundo. En conclusión, dos medidas pueden tener una correlación perfecta, pero sólo pueden considerarse en perfecto acuerdo cuando son exactamente iguales (o lo que es lo mismo, cuando los observadores son

intercambiables y los resultados son reproducibles).

2.3.1. El contexto de acuerdo: *kappa* intraclase

En el contexto teórico donde se interpreta el acuerdo entre evaluadores, la utilización de un índice de acuerdo sólo tiene sentido cuando las respuestas que *J* jueces emiten sobre un objeto son potencialmente intercambiables (es decir, tienen una distribución invariante a todas las permutaciones de los índices). En consecuencia, cualquier discrepancia entre las respuestas de los *J* observadores se consideran errores. En este contexto, Kraemer (1979) propuso un modelo de acuerdo –que se detalla en la Tabla 2.5– donde en el conjunto de las observaciones, para cada objeto *i*, la probabilidad de una respuesta positiva (e.g. "Sí") es $P(Sí)=p_i$ y su complemento es $P(No)=p_i'=1-p_i$, y en la población de objetos o sujetos, $E(p_i)=\pi$, $E(p_i')=1-\pi$, su varianza es $V(p_i)=V(p_i')=\pi(1-\pi)$ y su covarianza es $C(p_i,p_i')=\rho\,\pi(1-\pi)$, donde ρ es el coeficiente de correlación intraclase definido como la razón entre la covarianza y la raíz cuadrada del producto de las varianzas:

$$\rho=\frac{C(p_i,p_i')}{\sqrt{V(p_i)V(p_i')}} \qquad \text{(Ec. 2.15)}$$

En el marco de una tabla de acuerdo, el modelo teórico resultante se representa en la Tabla 2.5.

Tabla 2.5: Modelo teórico para el contexto de acuerdo

		Juez B		
		+ (*Sí*)	- (*No*)	*Total*
Juez A	+ (*Sí*)	$E(p_i^2)$	$E(p_i p_i')$	P
	- (*No*)	$E(p_i' p_i)$	$E(p_i'^2)$	P'
	Total	P	P'	1

Nótese que el modelo asume homogeneidad marginal. En consecuencia, en la población la probabilidad de que exista acuerdo entre evaluadores para el objeto i es $\pi^2(1-\pi)^2$. El acuerdo entre observadores es sin embargo aleatorio cuando la probabilidad de acuerdo es $\pi^2+(1-\pi)^2$. Y el índice *kappa*, siendo las probabilidades de acuerdo observado y esperado $P_o=\pi^2+(1-\pi)^2+2\rho\pi(1-\pi)$ y $P_e=\pi^2+(1-\pi)^2$, sería igual al coeficiente de correlación intraclase

$$\kappa_I=\frac{P_o-P_e}{1-P_e}=\frac{2\rho\pi(1-\pi)}{1-\pi^2-(1-\pi)^2}=\rho \qquad \text{(Ec. 2.16)}$$

Debido a esta equivalencia, *kappa* se conoce también en el contexto de acuerdo como ***kappa* intraclase**. Los estimadores máximo-verosimímiles de los parámetros π y κ_I son, respectivamente,

$$\hat{\pi} = \frac{2a + b + c}{2N} \qquad \text{(Ec. 2.17)}$$

$$\hat{\kappa}_I = \frac{4(ad - bc) - (b - c)^2}{(2a + b + c)(2d + b + c)} \qquad \text{(Ec. 2.18)}$$

siendo su varianza asintótica

$$V(\hat{\kappa}_I) = \sqrt{V(\hat{\rho})} \qquad \text{(Ec. 2.19)}$$

Si la fórmula para el coeficiente de correlación intraclase para datos numéricos bajo el modelo de efectos aleatorios en un sentido (véase Capítulo 1, p.20) se aplica a datos binarios, entonces el resultado es *kappa* intraclase (Winer, Brown y Michels, 1991).

Es importante remarcar que el índice *kappa* intraclase es algebraicamente equivalente al coeficiente π de Scott,asumiendo que los dos observadores son homogéneos en su respuesta.

2.3.2. Contexto de asociación: *kappa* ponderado

En el contexto de acuerdo, las valoraciones que los evaluadores hacen para cada uno de los sujetos se supone que son intercambiables. Cuando hay dos valoraciones independientes, pero no intercambiables, para el sujeto *i*, entonces $P(Sí|A)=p_i$ y $P(No|A)=p_i'$ para el primer juez (*A*) y $P(Sí|B)=q_i$ y $P(No|B)=q_i'$ para el segundo juez (*B*). En la población se cumple que $E(p_i)=P$ y $E(q_i)=Q$. .La probabilidad de que ambos observadores den una respuesta positiva es por tanto $E(p_i q_i)$. Este modelo, que se resume en la Tabla 2.6, no asume homogeneidad de la respuesta entre ambos evaluadores.

Tabla 2.5: Modelo teórico para el contexto de acuerdo

		Juez B		
		+ (*Sí*)	- (*No*)	*Total*
Juez A	+ (*Sí*)	$E(p_i^2)$	$E(p_i p_i')$	P
	- (*No*)	$E(p_i' p_i)$	$E(p_i'^2)$	P'
	Total	P	P'	1

Hay muchas medidas de asociación para este contexto que se han propuesto en la literatura estadística. Una de las más populares se fundamenta en la incorporación de un peso arbitrario *w* (para $0 \leqslant w \leqslant 1$) para ponderar las casillas de la tabla de acuerdo. Este fue el enfoque seguido por Spitzer y otros

(1967) primero y por Cohen (1968) después para llegar a la familia de estimadores de *kappa* ponderado, como se tratará más adelante.

2.4. Dos paradojas asociadas con el coeficiente *kappa*

La utilización de *kappa* como medida de acuerdo ha recibido muchas críticas. Básicamente se han detectado dos paradojas que pueden explicarse en términos de los efectos de **prevalencia** y **sesgo**. El efecto de prevalencia ocurre en presencia de una proporción global extrema de resultados para una categoría. El efecto de sesgo de un observador respecto de otro ocurre, en cambio, cuando sus probabilidades marginales son diferentes; el sesgo es mayor cuanto más discrepantes son sus respectivas probabilidades marginales y menor cuanto más similares son. Ambos efectos se han demostrado en los trabajos de Spitznagel y Hazer (1985), Cichetti yFeinstein (1990), Feinstein y Cicchetti (1990), Byrt, Bishop y Carlin (1993), Agresti, Ghosh y Bini (1995), Lantz y Nebenzahl (1996) y Hoehler (2000), entre otros. Un conjunto de ocho casos paradójicos para tablas 2×2 se muestra en la Tabla 2.7.

La primera (prevalencia) se describe en los términos siguientes: "si p_e es grande, el proceso de corrección del azar puede convertir un alto valor de p_o en un valor relativamente bajo de κ; por eso, con diferentes valores de p_e, el coeficiente κ para valores idénticos de p_o puede ser más de dos veces superior en un caso que en otro" (Feinstein y Cichetti, 1990; Byrt, Bishop y Carlin, 1993). Para ilustrar esta paradoja, en los dos primeros casos

de la Tabla 2.7 siendo p_o=.850, las proporciones de casos discrepantes prácticamente idénticas y las distribuciones marginales muy similares, κ_2 es menos de la mitad de κ_1 y lo mismo sucede con π, pero no con σ, ni en cierta medida tampoco con γ. Este efecto diferencial se atribuye a que la prevalencia de casos positivos (p_{11}) es en el segundo caso de .800 mientras que en el primer caso es de .400.

Tabla 2.7

		Juez B			
Caso 1		Sí	No	Total	σ_1=.700
Juez A	Sí	*.400*	*.090*	.490	π_1=.700
	No	*.060*	*.450*	.510	κ_1=.710
	Total	.460	.540	1.00	γ_1=.700
Caso 2		Sí	No	Total	σ_2=.700
Juez A	Sí	*.800*	*.100*	.900	π_2=.310
	No	*.050*	*.050*	.100	κ_2=.320
	Total	.850	.150	1.000	γ_2=.810
Caso 3		Sí	No	Total	σ_3=.200
Juez B	Sí	*.450*	*.150*	.600	π_3=.120
	No	*.250*	*.150*	.400	κ_3=.130
	Total	.700	.300	1.00	γ_3=.270
Caso 4		Sí	No	Total	σ_4=.200
Juez A	Sí	*.250*	*.350*	.600	π_4=.190
	No	*.050*	*.350*	.400	κ_4=.260
	Total	.300	.700	1.00	γ_4=.210

Tabla 2.7 (continuación)

		Juez B			
Caso 5		Sí	No	Total	σ_5=.200
Juez A	Sí	*.400*	*.200*	.600	
	No	*.200*	*.200*	.400	κ_5=.170
	Total	.600	.400	1.00	γ_5=.230
Caso 6		Sí	No	Total	
Juez A	Sí	*.400*	*.350*	.750	
	No	*.050*	*.200*	.250	κ_6=.240
	Total	.450	.550	1.00	
Caso 7		Sí	No	Total	
Juez A	Sí	*.400*	*.100*	.500	
	No	*.100*	*.400*	.500	κ_7=.600
	Total	.500	.500	1.00	
Caso 8		Sí	No	Total	σ_8=.600
	Sí	*.700*	*.100*	.800	π_8=.375
Juez A	No	*.100*	*.100*	.200	κ_8=.375
	Total	.800	.200	1.00	γ_8=.710

El coeficiente κ es además afectado por la prevalencia de las categorías "Sí" y "No" Byrt y otros (1993:425) denominan **índice de prevalencia** (*Prevalence Index*, PI) a la diferencia entre las probabilidades marginales de las categorías "Sí"y "No". Siendo $p_{1+}=a+b$ y $p_{+1}=a+c$ para la categoría "Sí" y $p_{2+}=c+d$ y $p_{+2}=b+d$ para la categoría "No", PI = $a-d$, que

puede tomar cualquier valor en el rango -1 (cuando $a=0$ y $d=1$) y +1 (cuando $a=1$ y $d=0$), siendo igual a 0 cuando las categorías "Sí" y "No" son equiprobables (prevalencia de .500). El efecto de prevalencia puede inferirse de los casos 7 y 8 de la Tabla 2.7. En ambos casos la probabilidad se acuerdo es la misma (o sea, $p_o=.800$) pero en el caso 7 el índice de prevalencia es PI = 0 y $\kappa=.600$ mientras que en el caso 8 PI = .600 y $\kappa=.375$. Esta notable diferencia en los valores del coeficiente *kappa* se debe al efecto de prevalencia: cuanto mayor es de *pi*, mayor es p_e y mayor es κ.

La segunda paradoja -sesgo- se describe en los términos siguientes: "las tablas de acuerdo con probabilidades marginales heterogéneas entre los dos jueces producen valores superiores de κ que las tablas con probabilidades marginales homogéneas" (Maclure y Willett, 1987; Feinstein y Cichetti, 1990; Byrt, Bishop y Carlin, 1993; Nelson y Pepe, 2000). En los casos 3 y 4 de la Tabla 2.7, siendo $p_o=.600$, las probabilidades marginales del juez *A* iguales pero las probabilidades marginales del juez *B* discrepante entre ambos casos, κ_4 es el doble de κ_3. Este efecto diferencial indeseable se atribuye a la discrepancia que existe entre las probabilidades marginales del juez *B*. Los mismos problemas afectan también al coeficiente π, pero no a σ ni a γ.

Byrt y otros (1993:424) definen el **índice de sesgo** (*Bias Index*, BI) como la diferencia en las proporciones de la respuesta "Sí" para los dos observadores y la estiman mediante $(a+b)-(a+c)=(b-c)$. El valor absoluto de BI tiene un valor mínimo de 0 (cuando $b = c$) y un valor máximo de 1 (cuando b ó c equivalen a N). El valor mínimo sólo puede obtenerse si las proporciones

marginales son iguales. El efecto de sesgo puede inferirse comparando los casos 5 y 6 de la Tabla 2.7. En ambos casos, la probabilidad de acuerdo esperado es $p_o = .600$, pero en el caso 5 las probabilidades marginales son homogéneas entre ambos evaluadores y el valor de κ es .170 mientras que en el caso 6 son heterogéneas y el valor de κ se incrementa hasta .240. En el caso 5, BI = 0 mientras que en el caso 6, BI = .300. La causa de esta paradoja se debe al sesgo existente entre observadores: conforme se incrementa el sesgo, p_e disminuye y κ aumenta.

En la misma línea de la propuesta para obtener un coeficiente corregido del sesgo entre jueces, Byrt y otros (1993:425) proponen un índice de acuerdo entre jueces para el caso 2×2 que ajusta el valor de *kappa* para las diferencias en la prevalencia de las categorías "Sí" y "No" y para el sesgo entre jueces. Lo denominan ***kappa* con prevalencia y sesgo corregidos** (*prevalence-adjusted bias-adjusted kappa*, PABAK), produciendo un valor de *kappa* que resulta de reemplazar tanto *b* y *c* como *a* y *d* por sus respectivos promedios. El índice de *kappa* resultante es similar al índice σ de Bennett. PABAK reescala de hecho la probabilidad de acuerdo observada p_o de forma que toma valores entre -1 (cuando $a = d = 0$) y +1 (cuando $b = c = 0$), siendo igual a 0 cuando la probabilidad de acuerdo esperado es igual a .500 (Looney y Hagan, 2008). Su fórmula general para el caso 2×2 es la siguiente:

$$\text{PABAK} = 2\,p_o - 1 \qquad \text{(Ec. 2.20)}$$

en cuya expresión revela que está linealmente relacionado con la probabilidad de acuerdo esperado. Obsérvese que, siendo $p_o = a + d$, la Ecuación 2.12 coincide exactamente con el valor del estimador del coeficiente σ. Para los datos del Ejemplo 2.1, PABAK= σ=.700.

Por su parte, κ se relaciona con PABAK mediante ésta fórmula:

$$\kappa = \frac{\text{PABAK} - \text{PI}^2 + \text{BI}^2}{1 - \text{PI}^2 + \text{BI}^2} \qquad \text{(Ec. 2.21)}$$

Resulta sencillo demostrar que, a menos que PABAK = 1, cuanto mayor sea el valor absoluto de BI, mayor será κ –dado PI constante –, y cuanto mayor sea el valor absoluto de PI, más pequeño será el valor de κ –dado BI constante–. Si ambos efectos, sesgo y prevalencia, están presentes, entonces el resultado puede deparar que κ sea mayor o menor que PABAK, dependiendo del tamaño relativo de BI y de PI.

2.5. La varianza de las medidas descriptivas de acuerdo mediante *jackknife*

Las expresiones para obtener las varianzas para cada uno de los coeficientes de acuerdo asumen que existe un diseño experimental equilibrado donde el mismo número de evaluadores valora cada uno de los objetos, y por tanto no son aplicables a diseños no equilibrados. Para computar la varianza de un coeficiente de acuerdo en un contexto no equilibrado se requiere la derivación de una expresión específica de tal contexto particular. Esta dificultad puede superarse hoy utilizando técnicas estadísticas apropiadas para estimar varianzas sin tener que derivar sus expresiones matemáticas. Entre las técnicas estadísticas más conocidas se encuentran los procedimientos de remuestreo denominados *jackknife* y *bootstrap*.

Quenouille (1949) introdujo el procedimiento *jackknife* originalmente como un método para reducir el sesgo de algunos estimadores. Posteriormente, Tukey (1958) propuso su utilización para la estimación de la varianza. Como tal técnica de estimación el procedimiento *jackknife* pertenece a la clase más general de procedimientos de remuestreo, que estiman la varianza de un estadístico computando el mismo estadístico varias veces a base de usar en cada ocasión una submuestra diferente de la muestra original y promediando las diferencias cuadráticas respecto de las estimaciones.

Para cualquiera de los coeficientes de acuerdo tratados en secciones anteriores, la forma concreta de obtener un estimador mediante *jackknife* sigue un conjunto de pasos que se especifican más abajo. Utilizaremos a título de ejemplo el coeficiente κ para ilustrar el proceso de computación.

1) Sea $\hat{\kappa}$ el estimador de κ, basado en el conjunto de datos de una muestra de tamaño N y sea $\hat{\kappa}_{(-i)}$ el estimador de κ cuando la i-ésima observación se elimina de la muestra, obteniendo así una muestra de tamaño $N-1$. Todas las observaciones deben ser una a una eliminadas de la muestra.

2) Se aplica la fórmula RCA general para obtener cada uno de los N estimadores $\hat{\kappa}_{(-i)}$, mediante

$$\hat{\kappa}_{(-i)} = \frac{p_{o(-i)} - p_{e(-i)}}{1 - p_{e(-i)}} \qquad \text{(Ec. 2.22)}$$

y reemplazando después la observación eliminada.

3) Una vez calculados los N estimadores $\hat{\kappa}_{(-i)}$, la varianza del estimador se obtiene finalmente aplicando

$$V(\hat{\kappa}) = \left(\frac{N-1}{N} \right) \sum_{i=1}^{N} (\hat{\kappa}_{(-i)} - \bar{\kappa}_{(-i)})^2 \qquad \text{(Ec. 2.23)}$$

donde

$$\bar{\kappa}_{(-i)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\kappa}_{(-i)} \qquad \text{(Ec. 2.2)}$$

Para una tabla de acuerdo 2×2 este proceso se simplifica notablemente, porque solo se precisan realizar cuatro operaciones diferentes, una para cada casilla de la tabla.

4) El programa MEVACO (López y Ato, 2008) implementa este proceso para tablas de acuerdo con 2 jueces y cualquier número de categorías. Los resultados son similares a las varianzas estándar de cada estimador, con aproximación a la centésima (véase ejemplo en Tabla 2.8).

Tabla 2.8 Resumen de resultados

Coeficiente	*Estimador*	*Desviación típica (estándar)*	*Desviación típica (jackknife)*
σ	.720	.070	.070
π	.670	.080	.080
κ	.680	.080	.080
γ	.760	.060	.060

Capítulo 3

Datos nominales y ordinales:

dos jueces, más de dos categorías

3.1. Introducción

En el capítulo anterior se presentaron un conjunto de procedimientos estadísticos para estimar el grado de acuerdo entre $J = 2$ jueces que debían clasificar un conjunto de N ítemes en una de $K = 2$ categorías posibles. En aquel contexto sólo era posible considerar categorías de respuesta nominales. Sin embargo, muchos estudios de acuerdo utilizan una escala de medida con

más de 2 categorías. El objetivo de este capítulo consiste en extender y generalizar los resultados del capítulo anterior al caso de $J = 2$ jueces que clasifican los ítemes en una escala con $K > 2$ categorías. En este contexto es posible considerar no solamente categorías nominales, como en el capítulo anterior, sino también categorías ordinales.

Este capítulo no presenta mayores dificultades técnicas que las discutidas en el capítulo anterior. No obstante, la descripción de los procedimientos de computación implican una mayor complejidad en la nomenclatura, lo que requiere un tratamiento pormenorizado y puede facilitar la transición al capítulo siguiente, que generalizará este contexto a un número mayor de jueces.

3.2. Categorías nominales

3.2.1. El caso *K* x *K*

En el capítulo anterior se examinó la evaluación del acuerdo entre observadores para el caso 2×2, que implica 2 jueces –*A* y *B*, por ejemplo– y 2 categorías de respuesta. La generalización a más de 2 categorías de respuesta requiere un tratamiento diferenciado debido a algunas circunstancias que se tratarán en el capítulo que nos ocupa.

Supongamos que se pide a 2 evaluadores que clasifiquen de forma independiente un conjunto de *N* ítemes u objetos sobre una escala de $K=3$

categorías de respuesta. Las categorías de respuesta pueden ser nominales (e.g. correspondiente a una clasificación simple de los ítemes) u ordinales (por caso, ajustada a una ordenación de los ítemes). El resultado de las respuestas de los evaluadores puede representarse en una tabla de contingencia 3×3. Por lo común, la generalización a cualquier número de categorías produce una tabla de contingencia que, utilizando las frecuencias observadas para cada casilla de la tabla, para el juez *A* (k = 1,..., K) y para el juez *B* (k' = 1,..., K) adopta la forma siguiente:

Tabla 3.1. Frecuencias de clasificación para el caso $K\times K$

		Juez B				
Juez A	*Categorías*	*1*	*2*	...	*K*	Marginales
	1	n_{11}	n_{12}	...	n_{1K}	n_{1+}
	2	n_{21}	n_{22}	...	n_{2K}	n_{2+}
	⋮	⋮	⋮	$n_{kk'}$	⋮	⋮
	K	n_{K1}	n_{K2}	...		n_{K+}
	Marginales	n_{+1}		...		*N*

Conviene distinguir en este lugar entre varios tipos de muestreo, dependiendo de si se fija el total muestral N, los marginales de fila n_{k+} o conjuntamente los marginales de fila n_{k+} y de columna $n_{+k'}$ (véase Brennan y Prediger, 1981:690-693; Martín y Femia, 2004:4-5). En el **muestreo tipo I** (o muestreo multinomial) se prefija el total muestral (*N*) y los marginales de fila y columna se asumen aleatorios. En el **muestreo tipo II** (o muestreo multinomial de producto) por el contrario se prefijan los marginales de fila (o de columna). En tal caso hay *K* distribuciones multinomiales (tantas como

marginales de fila o de columna de la Tabla 3.1), mientras que en el **muestreo tipo I** únicamente hay una distribución multinomial (que corresponde a todas las casillas de la Tabla 3.1). Asumiendo que se fijan los marginales de fila, el juez *A* se considera el experto –*gold standard*– con cuya ejecución se pretende comparar el juez *B*. En el **muestreo tipo III** se fijan conjuntamente los marginales de fila y columna y los observadores *A* y *B* se asumen con un grado de experiencia similar. Conviene advertir que es en este último contexto –muestreo tipo III– donde se contemplan todas las medidas de acuerdo corregidas del azar que serán discutidas posteriormente.

Utilizando proporciones, la Tabla 3.2 siguiente permite simplificar notablemente el proceso de cálculo, aunque sin perder de perspectiva la eficiencia computacional que puede deteriorar el uso de proporciones.

Tabla 3.2: Proporciones de clasificación para el caso $K \times K$

		Juez B				
Juez A	*Categorías*	*1*	*2*	...	*K*	Marginales
	1	p_{11}	p_{12}	...	p_{1K}	p_{1+}
	2	p_{21}	p_{22}	...	p_{2K}	p_{2+}
	⋮	⋮	⋮		⋮	⋮
	K	p_{K1}	p_{K2}	...	p_{KK}	p_{K+}
	Marginales			...		1

3.2.2. Ejemplo 3.1

Cuando las categorías se definen según una escala nominal, las fórmulas que se manejan para el cálculo de las medidas descriptivas de acuerdo son una generalización de las expuestas para el caso de 2 jueces y utilizan la misma notación y representación mediante tablas de acuerdo $K \times K$. Como ejemplo ilustrativo (en lo sucesivo, Ejemplo 3.1) utilizaremos una tabla de acuerdo 3×3 tomada del texto de Fleiss, Levin y Paik (2003:606; véase también Gwett, 2001:82), en el que se estudió el acuerdo obtenido en el diagnóstico emitido en una institución de salud mental por dos psiquiatras mediante la clasificación en una de tres posibles categorías diagnósticas ("psicosis", "neurosis" y "trastorno orgánico") para una muestra de 100 pacientes. Los datos se reproducen en la Tabla 3.3.

Tabla 3.3. Ejemplo 3.1

Diagnóstico psiquiátrico de $N = 100$ pacientes				
Categorías	*psicosis*	*neurosis*	*orgánico*	Marginales
psicosis	**75** (.750)	**1** (.010)	**4** (.040)	80 (.800)
neurosis	**5** (.050)	**4** (.040)	**1** (.010)	10 (.100)
orgánico	**0** (.000)	**0** (.000)	**10** (.100)	10 (.100)
Marginales	80 (.800)	5 (.050)	15 (.150)	100 (1.000)

Nota: Los valores en negrita son frecuencias; los valores entre paréntesis son proporciones

En todos los casos se utiliza la **proporción de acuerdo observado,** p_o, que se define– para toda categoría k de A y k' de B y siendo $k = k'$, como

$$p_o = \sum_{k=1}^{K} p_{kk'} \qquad \text{(Ec. 3.1)}$$

mientras que la definición de la **proporción de acuerdo esperado por azar** difiere para cada medida de acuerdo concreta.

Para los datos empíricos del Ejemplo 3.1 de la Tabla 3.3, la proporción de acuerdo observada es constante para todas las medidas descriptivas que se utilizan en este capítulo, siendo en concreto, $p_o = .750 + .040 + .100 = .890$.

En el caso del coeficiente σ (Bennett y otros, 1950; Maxwell, 1977; Brennan y Prediger, 1981), la proporción de acuerdo esperado por azar es sencillamente el inverso del número de categorías,

$$p_e^{\sigma} = 1/K \qquad \text{(Ec. 3.2)}$$

Para los datos del Ejemplo 3.1, $p_e^{\sigma} = 1/3 = .333$. El cálculo del coeficiente se obtiene aplicando la fórmula RCA

$$\hat{\sigma} = \frac{p_o - p_e^{\sigma}}{1 - p_e^{\sigma}} = \frac{K}{K-1}(p_o - 1) \qquad \text{(Ec. 3.3)}$$

que para los datos del Ejemplo 3.1 es $\hat{\sigma} = [(3)(.890) - 1]/2 = .835$. La varianza puede obtenerse aplicando una expresión desarrollada por Gwett

(2001:91)

$$V(\hat{\sigma}) = \left(\frac{1-f}{N}\right)\left(\frac{N}{N-1}\right)\left(\frac{K}{K-1}\right)^2 p_o(1-p_o) \qquad \text{(Ec. 3.4)}$$

donde f es la fracción de muestreo, en la práctica irrelevante si se asume que el número de ítemes del universo de donde se ha extraído es infinito. Aplicando la Ecuación 3.4 a los datos del Ejemplo 3.1,

$$V(\hat{\sigma}) = \left(\frac{1}{99}\right)\left(\frac{3}{2}\right)^2 (.890)(.110) = .002$$

En el caso del coeficiente π (Scott, 1955) la proporción de acuerdo esperado por azar es la suma de los cuadrados de los promedios de las probabilidades marginales simétricas para cada categoría, i.e.

$$p_e^{\pi} = \sum_{i=1}^{K} \left(\frac{p_{i+} + p_{+i}}{2}\right)^2 = \left(\frac{p_{1+} + p_{+1}}{2}\right)^2 + \left(\frac{p_{2+} + p_{+2}}{2}\right)^2 + \ldots + \left(\frac{p_{K+} + p_{+K}}{2}\right)^2$$

(Ec. 3.5)

Para los datos del Ejemplo 3.1,

$$p_e^{\pi} = [(.800+.800)/2]^2 + [(.100+.050)/2]^2 + [(.100+.150)/2]^2 = .661$$

y aplicando la fórmula RCA resulta

$$\hat{\pi} = \frac{p_o - p_e^{\pi}}{1 - p_e^{\pi}} = \frac{.89 - .6612}{1 - .6612} = .675$$

El cálculo de la varianza puede obtenerse utilizando una solución propuesta asimismo por Gwett (2001:89),

$$V(\hat{\pi}) = \left(\frac{1-f}{N-1}\right)\left[\frac{p_o(1-p_o)}{(1-p_e^{\pi})^2}\right] \qquad \text{(Ec. 3.6)}$$

que, prescindiendo de la fracción de muestreo que se asume en la práctica igual a 0, es similar a la fórmula originalmente propuesta por Scott (1955: 325). Para los datos empíricos del Ejemplo 3.1 la varianza sería, $V(\hat{\pi}) = .009$.

En el caso del coeficiente κ (Cohen, 1960) la proporción de acuerdo esperado por azar es igual a la suma de los productos de las probabilidades marginales simétricas para cada categoría,

$$p_e^{\kappa} = \sum_{K=1}^{K} p_{k+}\, p_{+k'} = p_{1+}\, p_{+1} + p_{2+}\, p_{+2} + ... + p_{K+}\, p_{+K} \quad \text{(Ec. 3.7)}$$

Para los datos empíricos del Ejemplo 3.1,

$$p_e^{\kappa} = (.800)(.800) + (.100)(.050) + (.100)(.150) = .660$$

y aplicando la fórmula RCA se obtiene

$$\hat{\kappa} = \frac{p_o - p_e^{\kappa}}{1 - p_e^{\kappa}} = \frac{.890 - .660}{.340} = .676$$

La varianza puede obtenerse aplicando una expresión desarrollada también por Gwett (2002:78),

$$V(\hat{\kappa}) = \left[\frac{p_o(1 - p_o)}{N(1 - p_e^{\kappa})^2} \right] \quad \text{(Ec. 3.8)}$$

Para los datos del Ejemplo 3.1, $V(\hat{\kappa}) = .0085$.

Y finalmente, respecto al coeficiente γ, la probabilidad de acuerdo esperada por azar se define como

$$p_e^{\gamma} = \frac{1}{(K-1)} \sum_{k=1}^{K} \pi_K (1 - \pi_K) \quad \text{(Ec. 3.9)}$$

y para los datos empíricos del Ejemplo 3.1 (Tabla 3.1), siendo

$$\pi_1 = (.800 + .800)/2 = .800$$

$$\pi_2 = (.100 + .050)/2 = .075$$

$$\pi_3 = (.100 + .150)/2 = .125$$

resulta igual a

$$p_e^{\gamma} = \frac{1}{2}\ \left[(.800)(.200) + (.075)(.925) + (.125)(.875)\right] = .1694$$

Empleando nuevamente la fórmula RCA, se obtiene

$$\hat{\gamma} = \frac{p_o - p_e^{\gamma}}{1 - p_e^{\gamma}} = \frac{.890 - .1694}{.8306} = .868$$

La varianza se calcula aplicando una fórmula desarrollada asimismo por Gwett (2001:82), que en esencia resulta muy similar a las anteriores:

$$V(\hat{\gamma}) = \left(\frac{1-f}{N-1}\right)\left(\frac{p_o(1-p_o)}{(1-p_e^{\pi})^2}\right) \qquad \text{(Ec. 3.10)}$$

En consecuencia, para los datos empíricos del Ejemplo 3.1, $V(\hat{y})=.0014$.

Nótese la complejidad que representa el cálculo de la varianza para todas las medidas descriptivas estudiadas en esta sección, cada una de las cuales utiliza una expresión particular, aunque las fórmulas son bastante similares entre sí.

Una alternativa muy interesante, que resulta imprescindible en aquellas ocasiones donde no se contemplan los requisitos necesarios para su utilización –e.g. en el caso de distribuciones no conocidas o con diseños no equilibrados–, es el procedimiento *jackknife* (Quenouille, 1949; Tukey, 1959), que se expuso en el capítulo anterior, aunque requiere un proceso de computación intensiva a través del ordenador. El programa MEVACO (López y Ato, 2008) calcula todas las medidas descriptivas que se han discutido hasta este capítulo utilizando el procedimiento *jackknife* para obtener los errores típicos de los coeficientes. Adviértase en la Tabla 3.4, que resume los principales resultados tratados hasta ahora, la notable aproximación del procedimiento *jackknife* para obtener las varianzas de los coeficientes de acuerdo con relación a los valores obtenidos aplicando las fórmulas específicas para el cálculo de las varianzas. Los estudios de Brennan (2001; ver también Gwett, 2001 y Dunn, 2004) han apuntado que, en general, con tablas de acuerdo el procedimiento *jackknife* ofrece la mejor aproximación para el cálculo de las varianzas en comparación con otros procedimientos alternativos de computación intensiva mediante ordenador (por ejemplo, alguna de las diferentes versiones del procedimiento *bootstrap*).

La Tabla 3.4 expone los estimadores de los coeficientes de acuerdo, junto con sus errores típicos estándar (obtenidos mediante las fórmulas apuntadas

más arriba, Ecuaciones 3.4, 3.6, 3.8 y 3.10) y sus errores típicos mediante el procedimiento *jackknife*, las razones críticas y sus correspondientes valores de probabilidad.

Tabla 3.4. Estimación y contraste de hipótesis para los coeficientes de acuerdo

Coeficiente	*Estimador*	*Error típico estándar*	*Error típico jackknife*	*Z**	*P > z*
σ	.835	.047	.048	17.400	<.001
π	.675	.093	.092	7.340	<.001
κ	.676	.092	.091	7.430	<.001
γ	.868	.037	.040	21700	<.001

Nota. Z^* asume como error típico el error típico *jackknife*

Una visión alternativa consiste en representar cada uno de los coeficientes de acuerdo en el contexto de una tabla 2×2 para comparar cada una de las categorías con las restantes. Para ello es necesario colapsar los valores de la Tabla 3.3 para cada categoría. De este modo, la tabla 2×2 resultante de colapsar la primera categoría ("psicosis") frente a todas las demás ("otras") sería:

Tabla 3.5. Ejemplo 3.1 colapsando la 1ª categoría

	psicosis	*otras*	Marginal
psicosis	75	5	80
otras	5	15	20
Marginal	80	20	100

Siguiendo este proceso de igual manera con las demás categorías, y centrándonos únicamente en el coeficiente *kappa*, la Tabla 3.6 resume los estimadores de *kappa* para evaluar el acuerdo en cada una de las categorías y para el conjunto, junto con su error típico *jackknife*, razones críticas y valores de probabilidad correspondientes

Tabla 3.6. Coeficientes de acuerdo para cada categoría

Categoría	p_o	p_e	*Estimador*	*Error típico jackknife*	*Z*	*P > z*
psicosis	.900	.680	.688	.095	7.240	<.001
neurosis	.930	.860	.500	.181	2.760	.003
orgánico	.950	.780	.773	.102	7.580	<.001
Global	.890	.660	.676	.091	7.430	<.001

En esta tabla puede constatarse que, utilizando la segunda y tercera columnas, el estimador global de *kappa* es igual a la suma de las diferencias entre probabilidades observadas y esperadas $p_o - p_e$ (o sea, los numeradores de los diferentes *kappas* individuales) dividido por la suma de las diferencias $1 - p_e$, es decir,

$$\hat{\kappa} = \frac{(.900-.680)+(.930-.860)+(.950-.780)}{(1-.680)+(1-.860)+(1-.780)} = .676$$

3. 3. Categorías ordinales

3.3.1. El coeficiente *kappa* ponderado

Cohen (1968) generalizó el coeficiente κ de acuerdo –originalmente propuesto para datos nominales– a datos ordinales y, más concretamente, al caso donde era posible valorar la seriedad relativa del desacuerdo (véase asimismo Spitzer, Cohen, Fleiss y Endicott, 1967).

La generalización implica asignar **pesos para el acuerdo** (*agreement weights*), definidos para los jueces A y B respectivamente como $w_{kk'}(k=1,\ldots,K\,;k'=1,\ldots,K)$, que se eligen sobre bases sustantivas y se asignan a cada una de las $K\times K$ casillas de la tabla de acuerdo. Los pesos se restringen a caer dentro del intervalo $0\leq w_{kk'}\leq 1$ y deben cumplir las condiciones siguientes:

1) $w_{kk}=1$ (esto es, el acuerdo exacto recibe el peso máximo);
2) $0\leq w_{kk'}\leq 1$ (en otros términos, todos los desacuerdos reciben menos del peso máximo);
3) $w_{kk'}=w_{k'k}$ (en particular, los jueces A y B se consideran iguales simétricamente).

Dadas estas condiciones, es posible definir con el coeficiente *kappa* la **proporción ponderada de acuerdo observado**, para las casillas que evalúan el acuerdo (para todo $k = k'$), mediante

$$p_o^w = \sum_{k=1}^{K} w_{kk}\, p_{kk} \qquad \text{(Ec. 3.11)}$$

y la **proporción ponderada de acuerdo esperado por azar** mediante

$$p_e^w = \sum_{k=1}^{K} \left(\sum_{k'>k}^{K} w_{kk'}\, p_{kk'} \right) \left(\sum_{k>k'}^{K} w_{kk'}\, p_{kk'} \right) \qquad \text{(Ec.3.12)}$$

El coeficiente *kappa* ponderado se obtiene, aplicando la fórmula RCA,

$$\hat{\kappa}^w = \frac{p_o^w - p_e^w}{1 - p_e^w} \qquad \text{(Ec. 3.13)}$$

Nótese que, cuando $w_{kk'} = 0$ (para todas las casillas que evalúan el desacuerdo, es decir, para $k \neq k'$) y $w_{kk} = 1$ (para todas las casillas que evalúan el acuerdo, o sea, para todo $k = k'$), el coeficiente *kappa* ponderado es idéntico al coeficiente *kappa* no ponderado, y su interpretación, aunque subjetiva, sigue las mismas pautas. Así, $\hat{\kappa}^w \geq .750$ se considera un grado de

acuerdo aceptable, mientras que $\hat{\kappa}^w \leq .400$ representa un grado de acuerdo escaso o insuficiente.

Los pesos pueden definirse de muchas formas posibles. Dos de las formas más conocidas son, los **pesos lineales diferenciales**, que propusieron Cichetti y Allison (1971),

$$w_{kk'} = 1 - \frac{|k - k'|}{(K-1)} \qquad \text{(Ec. 3.14)}$$

y los **pesos cuadráticos diferenciales**, que fueron propuestos en un trabajo de Fleiss y Cohen (1973),

$$w_{kk'} = 1 - \frac{(k - k')^2}{(K-1)^2} \qquad \text{(Ec. 3.15)}$$

La distribución muestral de *kappa* ponderado fue derivada por Fleiss, Cohen y Everitt (1969) y confirmada posteriormente por varios autores (entre ellos, Cichetti y Fleiss, 1977 y Hubert, 1978). La varianza se define mediante

$$V(\hat{\kappa}_w) = \frac{\sum_{k}^{K} \sum_{k'}^{K} p_{k+} p_{+k'} \left[w_{kk'} - (\bar{w}_{k+} + \bar{w}_{+k'}) \right]^2 - (p_e^w)^2}{N(1 - (p_e^w)^2)} \qquad \text{(Ec. 3.16)}$$

donde $\bar{w}_{k+} = \sum_{k'}^{K} w_{kk'} p_{+k'}$ y $\bar{w}_{+k'} = \sum_{k}^{K} w_{kk'} p_{k+}$ .La hipótesis nula puede someterse a prueba, como es habitual, utilizando el valor de la razón crítica

$$z = \frac{\hat{\kappa}_w}{\sqrt{V(\hat{\kappa}_w)}} \qquad \text{(Ec. 3.17)}$$

3.3.2. Ejemplo 3.2: Los datos de von Eye y Schuster

Se ilustra seguidamente el cálculo del coeficiente *kappa* ponderado con los datos empíricos tomados de una investigación sobre la fiabilidad del diagnóstico psiquiátrico (von Eye y Schuster, 2000), que en adelante denominaremos Ejemplo 3.2. Dos psiquiatras reevaluaron los ficheros de N = 129 pacientes internos que habían sido previamente diagnosticados dentro de un amplio intervalo de diagnóstico clínico. Las categorías de clasificación fueron "1: no deprimido" (*nd*), "2: moderadamente deprimido" (*md*) y "3: clínicamente deprimido" (*cd*). Los datos se representan en la Tabla 3.7.

Tabla 3.7. Datos empíricos del Ejemplo 3.2

Diagnóstico psiquiátrico de 129 pacientes				
Categorías	*1 - nd*	*2 - md*	*3 - cd*	Marginales
1 - nd	**11** (.085)	**2** (.016)	**19** (.147)	32 (.248)
2 - md	**1** (.008)	**3** (.023)	**3** (.023)	7 (.054)
3 - cd	**0** (.000)	**8** (.062)	**82** (.636)	90 (.698)
Marginales	12 (.093)	13 (.101)	104 (.806)	129 (1.000)

Utilizando los pesos cuadráticos diferenciales, la matriz de pesos que se obtiene es la siguiente:

Tabla 3.8: Matriz de pesos cuadráticos del Ejemplo 3.2

1.000	.750	.000
.750	1.000	.750
.000	.750	1000

Aplicando la Ecuación 3.11 a los datos empíricos del Ejemplo 3.2 y usando la matriz de pesos cuadráticos de la Tabla 3.8 obtenemos

$$p_o^w = .085 + (.016)(.750) + (.008)(.750) + (.023) + (.023)(.750) + (.062)(.750) + .636 = .826$$

y mediante la Ecuación 3.12 hallamos

$$p_e^w = (.248)(.093) + (.248)(.101)(.750) + (.054)(.093)(.750) + \ldots + (.698)(.806) = .700$$

Finalmente, empleando la Ecuación 3.13, se consigue el valor del coeficiente *kappa* ponderado

$$\hat{\kappa}_w = \frac{.826 - .700}{1 - .700} = .420$$

Este valor difiere sensiblemente del coeficiente *kappa* no ponderado, que alcanza un valor de $\hat{\kappa} = .3745$ para los datos empíricos del Ejemplo 3.2.

La hipótesis nula plantea si el coeficiente es igual a cero y se prueba, una vez calculada la desviación típica del coeficiente mediante la Ecuación 3.16, y siendo $\sqrt{V(\hat{\kappa}_w)} = .079$, utilizando a continuación la Ecuación 3.17, y obteniendo finalmente $z = .420/.079 = 5.332$; $P < .001$, que conduce a la conclusión que se rechaza la hipótesis nula de que el coeficiente sea próximo a cero.

3.3.3. La equivalencia entre *kappa* ponderado y el coeficiente de concordancia

Una forma alternativa de representar la Ecuación 3.13 es (véase Shoukri, 2004:41):

$$\hat{\kappa}_w = 1 - \frac{N \sum_{k=1}^{K} \sum_{k'=1}^{K} (k-k')^2 n_{kk'}}{\sum_{k=1}^{K} \sum_{k'=1}^{K} n_{k+} n_{+k'} \ (k-k')^2} \qquad \text{(Ec. 3.18)}$$

y siendo

$$(k-k')^2 = (k-\bar{x}_A)^2 + (k'-\bar{x}_B)^2 + (\bar{x}_A-\bar{x}_B)^2 + 2(\bar{x}_A-\bar{x}_B)(k-\bar{x}_A) - 2(\bar{x}_A-\bar{x}_B)(k'-\bar{x}_B) - 2(k-\bar{x}_B)(k'-\bar{x}_B) \qquad \text{(Ec. 3.19)}$$

entonces resulta que

$$\sum_{k=1}^{K} \sum_{k'=1}^{K} n_{kk'}(k-k')^2 = N s_A^2 + N s_B^2 + N(\bar{x}_A-\bar{x}_B)^2 - 2N s_{AB}$$
$$\sum_{k=1}^{K} \sum_{k'=1}^{K} n_{k+} n_{+k'}(k-k')^2 = N s_A^2 + N s_B^2 + N(\bar{x}_A-\bar{x}_B)^2 \qquad \text{(Ec. 3.20)}$$

y sustituyendo en la Ecuación 3.13 se obtiene finalmente como resultado

$$\hat{\kappa}_w = \frac{2\, s_{AB}}{s_A^2 + s_B^2 + (\bar{x}_A - \bar{x}_B)^2} \qquad \text{(Ec. 3.21)}$$

que es exactamente la expresión del **coeficiente de correlación de concordancia** propuesta por Lin (1989), empleada con variables de respuesta cuantitativas. La expresión de la Ecuación 3.21 fue no obstante originalmente derivada en un trabajo publicado por Krippendorf (1970) y por esta razón en ocasiones se le conoce como coeficiente de correlación de concordancia de Lin-Krippendorf (Shoukri, 2002), como también demostró la tesis doctoral de Robieson (1999). Sin embargo, conviene precisar que esta equivalencia entre *kappa* ponderada y el coeficiente de concordancia únicamente se cumple cuando se utilizan los pesos cuadráticos diferenciales (Fleiss y Cohen, 1973).

3.3.4. La equivalencia entre *kappa* ponderado y el coeficiente de correlación intraclase para un modelo mixto sin interacción

Asumiendo que las *K* categorías que los evaluadores deben clasificar para cada ítem u objeto son ordenadas y que se decide tomar las categorías observadas para cada juez como si se tratara de una medida cuantitativa continua, otra equivalencia que resulta interesante resaltar en este contexto es la existente

entre el coeficiente *kappa* ponderado y el coeficiente de correlación intraclase, obtenido para un modelo mixto con dos factores sin interacción.

Se ilustra esta equivalencia con los datos del Ejemplo 3.2 utilizado anteriormente, para lo cual definimos un modelo ANOVA mixto no replicado con dos factores: *ítemes*, de efectos aleatorios, con 129 condiciones o niveles, y *juez*, de efectos fijos, con 2 condiciones o niveles. Las observaciones que se toman como variable de respuesta son todas las registradas para cada ítem y para cada juez. El fichero de datos para analizar –por ejemplo– con SPSS o SAS, debe seguir una representación univariante (Ato y Vallejo, 2007) con la forma siguiente:

Tabla 3.9. Representación univariante de los datos del Ejemplo 3.2

Caso	*Variable respuesta*	*Juez* (*J*)	*Ítem* (*I*)
1	1	1	1
...	...	...	...
33	2	1	33
...	...	...	...
40	3	1	40
...	...	...	...
130	1	1	1
...	...	...	...
141	2	2	12
142	2	2	13
143	3	2	14
...	...	...	...
162	1	2	33
163	2	2	34
164	2	2	35
165	2	2	36
166	3	2	37
...	...	...	...
169	2	2	40
...	...	...	...
177	3	2	48
...	...	...	...
258	3	2	129

La salida del correspondiente análisis de varianza se muestra en la Tabla 3.10.

Tabla 3.10: ANOVA para un modelo mixto

Fuentes	*S.C.*	*g.l.*	*M.C.*	*Componentes de la varianza*
Ítemes (*I*)	105.791	128	.8265	$\hat{\sigma}_I^2 = (.8265 - .3166)/2 = .4203$
Jueces (*J*)	4.481	1	4.4806	
Residual	40.519	128	.3166	$\hat{\sigma}_e^2 = .3166$
Total	150.791	257		

A destacar la forma atípica de estimar la varianza del factor juez. En lugar de emplear, como es usual, el estimador estándar del componente de varianza $\hat{\sigma}_J^2 = (4.4806 - .3166)/129 = .0323$, basada en la esperanza de la media cuadrática correspondiente, se utiliza en su lugar como estimador $\hat{\sigma}_J'^2$ que se define como una varianza directa respecto de su media cuadrática, tal y como se representa en la Tabla 3.10. Este hecho diferencial ya se puso de manifiesto en un trabajo de Fleiss y Cohen (1973:617) y se ha visto posteriormente reflejado en varios trabajos (Carrasco y Jover, 2003; Li, 2007; King, Chinchilli y Carrasco, 2007).

El coeficiente de correlación intraclase puede obtenerse así pues mediante

$$\hat{\rho} = \frac{\hat{\sigma}_I^2}{\hat{\sigma}_I^2 + \hat{\sigma}_J'^2 + \hat{\sigma}_e^2} \qquad \text{(Ec. 3.22)}$$

y es igual a $\hat{\rho}_I = .2550/(.2550 + .0350 + .3166) = .420$, que es el mismo valor

que se obtuvo con el coeficiente *kappa* ponderado.

Schuster (2004:247-8) ha mostrado en detalle la conexión entre *kappa* ponderado, el coeficiente de correlación de contingencia y el modelo ANOVA mixto sin interacción. Retomando la relación expuesta en la Ecuación 3.21, para el caso de 2 evaluadores, con un pequeño cambio en el segundo término del denominador para reflejar el carácter finito de los promedios marginales para ambos jueces,

$$\hat{\kappa}_w = \frac{2 s_{AB}}{(s_A^2 + s_B^2) + \left(\frac{N}{N-1}\right)(\bar{x}_A - \bar{x}_B)^2} \qquad \text{(Ec. 3.23)}$$

la expresión equivalente en términos de medias cuadráticas es

$$\hat{\kappa}_w = \frac{MC_I - MC_R}{MC_I + MC_R + \left(\frac{2}{N-1}\right)MC_J} \qquad \text{(Ec. 3.24)}$$

donde MC_I es la media cuadrática debida a ítemes, MC_R es la media cuadrática residual/error y MC_J es la media cuadrática debida a jueces. La equivalencia se fundamenta en el conjunto de ecuaciones siguiente, que representan cada uno de los términos de la Ecuación 3.23 y la Ecuación 3.24:

$$\left(\frac{2}{N-1}\right)MC_R = \frac{N}{N-A}(\bar{x}_A - \bar{x}_B)^2$$
$$MC_I + MC_R = (s_A^2 + s_B^2)$$
$$MC_I + MC_R = 2s_{AB} \qquad \text{(Ec. 3.25)}$$

Utilizando los datos empíricos del Ejemplo 3.2 con la Ecuación 3.24 obtenemos también

$$\hat{\kappa}_w = \frac{.8265 - .3166}{.8265 + .3166 - (2/128)(4.4806)} = .420$$

que es de nuevo el mismo valor obtenido para *kappa* ponderado a través la aplicación estándar.

Esta equivalencia justifica la utilización atípica de la estimación de la varianza para los observadores en la Ecuación 3.22 con el fin de obtener el coeficiente de correlación intraclase.

3.4. Distinguibilidad entre categorías

Darroch y McCloud (1986) destacaron que en su origen el interés hacia los coeficientes de acuerdo no residía tanto en describir cómo 2 observadores acuerdan entre sí como en medir en qué medida difieren en distinguir las categorías de clasificación de los ítemes. En muchas circunstancias, las categorías no tienen definiciones objetivas precisas y, en consecuencia, cabe aceptar, en primer lugar, que diferentes expertos interpreten de forma diferente las definiciones de las categorías y, en segundo lugar, que las categorías no serán completamente distinguibles entre jueces, ni siquiera para un mismo evaluador. Estos dos aspectos – diferencia entre jueces y distinguibilidad entre las categorías– fueron minuciosamente estudiados por estos autores, quienes describieron un modelo para las probabilidades de clasificación conjunta que incorpora las siguientes propiedades:

1) la clasificación de un ítem u objeto por un juez puede ser aleatoria;
2) distintos jueces pueden tener diferentes probabilidades de clasificación para el mismo ítem;
3) no se asume una interacción multiplicativa entre los efectos de ítem y los efectos de juez en las probabilidades de clasificación.

El modelo de Darroch y McCloud (1986) define el grado de **distinguibilidad** entre dos categorías de las probabilidades de clasificación

conjunta para dos jueces, que varía de un par de jueces a otro; sin embargo, el **valor medio de distinguibilidad entre categorías,** que llamaron δ, apenas sufría variación. En consecuencia, concluyeron que el coeficiente *kappa* y sus variantes depende en gran medida de qué par de observadores clasifique un conjunto de ítemes, por lo que recomendaron utilizar en lugar de coeficientes de acuerdo el valor medio de distinguibilidad, que definieron mediante:

$$\delta = 1 - \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{k=k'+1}^{K} \left(\frac{\pi_{k'k} \pi_{kk'}}{\pi_{k'k'} \pi_{kk}} \right) \qquad \text{(Ec.3.26)}$$

Es conveniente señalar las siguientes propiedades deseables de esta definición:

1) $\delta = 1$ sí y sólo si todos los pares de categorías son completamente distinguibles, en cuyo caso $\sum_{k'=1}^{K} \pi_{k'k'} = 1$.
2) $\delta = 0$ sí y sólo si todos los pares de categorías son completamente indistinguibles, lo que implica que $\pi_{k'k} = \pi_{k'+} \pi_{+k}$, para todo k y k'.
3) El valor medio de distinguibilidad δ varía en el rango entre 0 y 1.

El coeficiente *kappa* posee la propiedad 1, pero no las propiedades 2 y 3. Sin embargo, las tres son propiedades que debería cumplir un estimador óptimo

del grado de acuerdo entre dos jueces. Este es el fundamento de una de las críticas que se ha realizado sobre el coeficiente (Cichetti y Feinstein, 1990; Feinsten y Cichetti, 1990).

Capítulo 4

Datos nominales y ordinales: más de dos jueces, dos o más categorías

4.1. Introducción

Los coeficientes descriptivos que se estudiaron en los Capítulos 2 y 3 tienen en común que se derivaron para 2 jueces, de ahí que la notación dominante utilice tablas de contingencia en dos sentidos, que es la representación más usual. No obstante, muchas investigaciones aplicadas utilizan más de 2 observadores y la notación que se aplica raramente se expresa mediante tablas de contingencia, debido a la dificultad de su representación. En este capítulo fundamentalmente

se trata de generalizar los resultados de capítulos anteriores al caso de más de 2 jueces o evaluadores utilizando representaciones de los datos alternativas a las tablas de contingencia o tablas de acuerdo clásicas.

4.2. Formas de representación

En lugar de tablas de contingencia, la representación más conveniente para realizar el cálculo de los coeficientes de acuerdo con más de 2 jueces es una expresión rectangular donde las filas son los ítemes, las columnas son los jueces/observadores y las categorías de clasificación (nominal u ordinal) utilizadas por cada observador para valorar cada ítem se representan en la intersección (véase Tabla 4.1). Hay dos opciones posibles a contemplar en este contexto: cuando el número de ítemes u observaciones es pequeño, la representación más apropiada es utilizar una fila por ítem (opción que llamaremos **representación directa o no abreviada**); por el contrario, cuando el número de ítemes es grande, la representación directa es poco recomendable, y es más común usar una fila para cada una de las combinaciones –o patrones– posibles de respuesta (opción que llamaremos **representación abreviada**).

Ilustramos de forma detallada y minuciosa el cálculo de los coeficientes de acuerdo a continuación que se han definido para más de 2 jueces con las dos representaciones posibles, sirviéndonos de dos ejemplos tratados anteriormente de la literatura.

4.3. Ejemplo 4.1: los datos de Conger (1980)

La Tabla 4.1 representa los datos empíricos que corresponden al Ejemplo 4.1 (Conger, 1980:325), donde se trata de clasificar una muestra de $N = 10$ ítemes por un conjunto de $J = 4$ jueces, en una de $K = 3$ categorías de respuesta, identificadas en la tabla mediante representación directa con las etiquetas *a*, *b* y *c* respectivamente.

Tabla 4.1. Ejemplo 4.1 (representación directa)

Jueces (J)					
Ítemes (I)	*1*	*2*	*3*	*4*	*Frecuencias*
1	*a*	*a*	*a*	*c*	1
2	*a*	*a*	*b*	*c*	1
3	*a*	*a*	*b*	*c*	1
4	*a*	*a*	*c*	*c*	1
5	*a*	*b*	*a*	*a*	1
6	*b*	*a*	*a*	*a*	1
7	*b*	*b*	*b*	*b*	1
8	*b*	*c*	*b*	*b*	1
9	*c*	*c*	*b*	*b*	1
10	*c*	*c*	*c*	*c*	1

Adviértase que existe un total de $K^J = 3^4 = 81$ combinaciones o patrones posibles, 10 de los cuales son los que concretamente se han obtenido y se representan en la Tabla 4.1, de los que únicamente uno de ellos se repite (el patrón *a-a-b-c*). En esta situación, donde hay pocos ítemes, no tendría ningún

interés simplificar los patrones de respuesta producidos. Sólo es pertinente aplicar la representación directa o no abreviada, la que en definitiva se presenta en la Tabla 4.1.

4.3.1. El procedimiento analítico y la generalización de los coeficientes descriptivos

Como apunta Rae (1988:48), la dificultad principal para definir una medida de variación para el caso categórico es la tendencia a pensar en la variabilidad como una medida de desviación de las observaciones respecto de su media, ya que con datos categóricos la media es indefinible. Sin embargo, Gini (1939) observó que la suma de cuadrados de las desviaciones respecto de la media para *N* medidas categóricas podía expresarse igualmente mediante una suma de cuadrados del siguiente modo:

$$SC = \frac{1}{2N} \sum \sum d_{ij}^2 \qquad \text{(Ec. 4.1)}$$

El inconveniente principal que supone partir de la variación y el correspondiente análisis de varianza con datos categóricos, de acuerdo con la formulación de Gini, atañe a la correcta disposición de los datos mediante la utilización de un paquete estadístico estándar. Cuando solamente hay dos categorías, la situación no plantea ningún problema porque los datos empíricos son simplemente ceros y unos, y las diferencias d_{ij} se definen de forma directa e inambigua; pero cuando el número de categorías es mayor de dos, es

necesario definir un vector multinomial para la respuesta que cada observador da a cada ítem, un vector con tantos elementos como categorías requiera el proceso de clasificación. Sin embargo, Light y Margolin (1971) y Margolin y Light (1974) describieron un procedimiento de cálculo simplificado mediante un ANOVA ordinario, que describimos en detalle a continuación. Un procedimiento similar fue también tratado por Landis y Koch (1977b).

Sea *I* la variable que denota los ítemes $(para\ i = 1, \ldots, N)$, *J* la variable que representa los jueces u observadores $(para\ j = 1, \ldots, J)$ y *K* la variable que corresponde a las categorías de respuesta $(para\ k = 1, \ldots, K)$ utilizados en el proceso de clasificación. Se asume así que el número total de ítemes es *N*, el número de jueces es *J* y el número de categorías de respuesta es *K*. Con datos procedentes de un estudio de acuerdo, es posible registrar dos magnitudes básicas: n_{ik} es el número total de jueces que asignaron a una categoría particular *k* cada ítem u objeto *i*, y n_{jk} es el número total de ítemes que cada observador asigna a cada categoría. Desde cualquiera de las dos magnitudes básicas es posible definir $n_{+k} = \sum_{i=1}^{N} n_{ik} = \sum_{j=1}^{J} n_{jk}$, que es el número total de respuestas registradas en la categoría *k*. En consecuencia, el número total de respuestas en todas las categorías de la tabla de contingencia resultante equivale a $\sum_{j=1}^{J} \sum_{i=1}^{N} n_{ij} = NJ.$

Sustituyendo en la Ecuación 4.1, Light y Margolin (1971) y Margolin y Light (1971) demostraron que las cantidades siguientes:

$$[1]: \quad \frac{NJ}{2} \qquad \text{(Ec. 4.2)}$$

$$[2]: \quad \frac{1}{2NJ}\sum_{k=1}^{K} n_{+k}^2 \qquad \text{(Ec. 4.3)}$$

$$[3]: \quad \frac{1}{2J}\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 \qquad \text{(Ec. 4.4)}$$

hacen posible una partición de la varianza, de forma similar a un ANOVA en un sentido con variables numéricas, donde la **SC Total** puede definirse a continuación mediante la diferencia:

$$[1]-[2]=SC_T=\frac{NJ}{2}-\frac{1}{2NJ}\sum_{k=1}^{K} n_{+k}^2 \qquad \text{(Ec. 4.5)}$$

y la partición permite distinguir un componente para la **SC Inter-ítemes** (SC Ítemes o SC_I) y otro componente para la **SC Intra-ítemes** o **SC Error** (SC_E) mediante las diferencias respectivas:

$$[1]-[3]=SC_E=\frac{NJ}{2}-\frac{1}{2J}\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 \qquad \text{(Ec. 4.6)}$$

$$[3]-[2]=SC_I=\frac{1}{2J}\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2-\frac{1}{2NJ}\sum_{k=1}^{K} n_{+k}^2 \qquad \text{(Ec. 4.7)}$$

La correspondiente tabla ANOVA resume el proceso de partición. Para los

datos empíricos del Ejemplo 4.1 (véase la Tabla 4.1 y la Tabla 4.5 donde se resume el proceso de cálculo), aplicando las Ecuaciones 4.2, 4.3 y 4.4, se obtiene

$$[1]: \quad \frac{NJ}{2} = \frac{(10)(4)}{2} = 20$$

$$[2]: \quad \frac{1}{2NJ} \sum_{k=1}^{K} n_{+k}^2 = \frac{538}{(2)(10)(4)} = 6.725$$

$$[3]: \quad \frac{1}{2J} \sum_{i=1}^{N} \sum_{k=1}^{K} n_{ik}^2 = \frac{100}{(2)(4)} = 12.500$$

y aplicando las Ecuaciones 4.5. 4.6 y 4.7, SC_T = [1] – [2] = 13.275, con gl_T = 40 – 1 = 39; SC_I = [3] – [2] = 5.775, con gl_I = 10 – 1 = 9, y SC_E = [1] – [3] = 7.5, con gl_E = 40 – 10 = 30. La tabla ANOVA en un sentido resultante (Tabla 4.2) se muestra a continuación.

Tabla 4.2. ANOVA en un sentido para el Ejemplo 4.1

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	5.775	9	.642	$\hat{\sigma}_I^2 = (.642 - .250)/4 = .098$
Error (SC_E)	7.500	30	.250	$\hat{\sigma}_E^2 = .250$
Total (SC_T)	13.275	39		

Paralelamente, Fleiss (1971) definió un estadístico tipo *kappa* para múltiples observadores, que se ha demostrado que es en esencia equivalente al

coeficiente *pi* de Scott (véase Gwett, 2002), aunque originalmente Fleiss lo concibió como una generalización de *kappa* de Cohen, que en adelante llamaremos **coeficiente *pi* generalizado** (o simplemente π), aplicando la fórmula RCA que tratamos en el Capítulo 2. Este coeficiente puede también obtenerse con facilidad a partir de los resultados que se exponen en la Tabla 4.5. Siendo

$$p_o = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 - NJ}{NJ(J-1)} \qquad \text{(Ec. 4.8)}$$

$$p_e^{\pi} = \frac{\sum_{k} n_{+k}^2}{N^2 J^2} = \sum_{k} p_{+k}^2 \qquad \text{(Ec. 4.9)}$$

y teniendo en cuenta que

$$\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 = N^2 J^2 - 2J\,SC_E \qquad \text{(Ec. 4.10)}$$

$$\sum_{k=1}^{K} n_{+k}^2 = N^2 J^2 - 2NJ\,SC_T \qquad \text{(Ec. 4.11)}$$

sustituyendo en las Ecuaciones 4.8 y 4.9 y simplificando obtenemos

$$p_o = 1 - \frac{2SC_E}{N(J-1)} \qquad \text{(Ec.4.12)}$$

$$p_e^{\pi} = 1 - \frac{2SC_T}{NJ} \qquad \text{(Ec. 4.13)}$$

Con los datos empíricos del Ejemplo 4.1, tomando los valores de la Tabla 4.2,

$$p_o = 1 - [(2)(7.5)]/[(10)(3)] = .500$$

$$p_e^{\pi} = 1 - [(2)(13.275)]/[(10)(4)] = (.375)^2 + (.325)^2 + (.300)^2 = .33625$$

En consecuencia, aplicando la ecuación general RCA, el estadístico *pi* generalizado π de Fleiss se estima mediante

$$\hat{\pi} = \frac{(.500 - .336)}{1 - .336} = .247$$

Por supuesto, si se emplean los resultados del ANOVA en un sentido de la Tabla 4.2, puede alternativamente estimarse el estadístico *pi* generalizado de Fleiss a partir de:

$$\hat{\pi} = \frac{SC_I - \frac{SC_E}{J-1}}{SC_I + SC_E} \qquad \text{(Ec. 4.14)}$$

De nuevo, con los datos empíricos del Ejemplo 4.1, $\hat{\pi} = .247$, que es obviamente el mismo valor que se obtuvo con la ecuación RCA. Además, es posible también deducir (véase Winer,1971:286) que, asumiendo un N suficientemente grande, el coeficiente *pi* generalizado de Fleiss de hecho

estima el coeficiente de correlación intraclase para un modelo ANOVA en un sentido, (i.e.)

$$\pi \approx \rho_1 = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_E^2} \qquad \text{(Ec. 4.15)}$$

Empleando los estimadores que se muestran en la Tabla 4.2, el resultado implica que algo más de un 28% de la varianza de la variable de respuesta puede atribuirse a la variación entre ítemes: $\hat{\rho}_1 = .098/(.098+.250) = .282$. La diferencia con respecto del estimador de π obtenido anteriormente se debe probablemente al escaso número de ítemes utilizado.

Conger (1980) ha mostrado algunos inconvenientes del coeficiente *pi* generalizado propuesto por Fleiss (1971), particularmente cuando los datos empíricos violan el supuesto de independencia y las distribuciones marginales de los jueces u observadores no son estrictamente idénticas, que son los supuestos que asume. Obsérvese que el coeficiente *pi* generalizado asume una probabilidad de acuerdo por azar que no se basa en las probabilidades marginales individuales de los jueces, sino que son una función de las probabilidades globales de que un juez elegido al azar clasifique un ítem elegido también al azar en una categoría específica. Esta idea es similar a la propuesta por Scott (1955), mientras que la probabilidad de acuerdo por azar del coeficiente *kappa* de Cohen (1960) es una función de las probabilidades de clasificación marginales de los jueces. En tal caso se presentan dos consecuencias indeseables relativas al proceso de generalización:

- en primer lugar, π tiende a ser negativo con datos generados por azar,

- en segundo lugar, contrariamente a como fue concebido por Fleiss (1971), el coeficiente pi generalizado no se reduce al coeficiente *kappa* de Cohen cuando el número de jueces es igual a 2.

A pesar de que π se definió en teoría como una generalización del clásico coeficiente *kappa* de Cohen al caso de más de 2 evaluadores (véase Fleiss, 1971), puede de hecho comprenderse como una generalización del coeficiente *pi* de Scott, como se ha propuesto anteriormente. Asimismo, una alternativa en perfecta continuidad con el coeficiente *kappa* de Cohen para cualquier número de jueces fue algún tiempo después propuesta por Conger (1980), que la definió como *exact Fleiss statistics* (p. 325), y que llamaremos aquí **coeficiente *kappa* generalizado** $-(\text{o simplemente } \kappa)-$. Sin embargo, para aplicar esta alternativa es necesario ampliar el modelo ANOVA en un sentido utilizado en la Tabla 4.2 a un modelo en dos sentidos.

Apréciese que el desarrollo del modelo ANOVA en un sentido se ha fundamentado en un formato de representación específico, concretamente el formato de ítemes por categorías, donde las entradas n_{ik} representan el número de jueces que asignan el ítem *i* a una determinada categoría *k*. No obstante, igualmente pueden representarse de forma alternativa siguiendo el formato de jueces por categorías, donde las entradas n_{jk} denotan el número de ítemes que cada observador *j* asigna a una categoría particular *k*. Téngase en cuenta que en ambas representaciones los marginales n_{+k} son iguales y, en consecuencia, la SC Total es coincidente. No obstante, en esta nueva

representación es posible a partir de la variación de error anterior (SC_E) en dos nuevos componentes, uno que reflejaría diferencias entre jueces (SC_J) y un segundo componente residual (SC_R). Como continuación de las Ecuaciones 4.2, 4.3 y 4.4, que definen las magnitudes para los cálculos básicos [1] a [3], se precisa definir entonces una nueva magnitud de cálculo mediante

$$[4]: \quad \frac{1}{2N}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}^2 \qquad \text{(Ec. 4.16)}$$

donde la suma de cuadrados debida a jueces u observadores (SC_J) y la suma de cuadrados residual (SC_R) se obtienen por sustracción,

$$[4]-[2]: \quad SC_J = \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K} n_{jk}^2 - \frac{1}{2JN}\sum_{k=1}^{K} n_{+k}^2 \qquad \text{(Ec. 4.17)}$$

$$[1]-[3]-[4]+[2]: \quad SC_R = SC_E - SC_J \qquad \text{(Ec. 4.18)}$$

Con los datos empíricos del Ejemplo 4.1, la nueva magnitud de cálculo básico resulta igual a (véase Tabla 4.5)

$$[4] = \frac{1}{20}\ (152) = 7.600$$

y para el cálculo de las sumas de cuadrados se procede aplicando las

Ecuaciones 4.17 y 4.18,

$$SC_J = 7.600 - 6.725 = .875$$

$$SC_R = 7.500 - .875 = 6.625$$

La tabla ANOVA resultante corresponde ahora a un modelo ANOVA en dos sentidos sin interacción (Ato y Vallejo, 2007). Asumiendo que los Ítemes representan un efecto aleatorio, es posible considerar dos casos para este modelo ANOVA, a saber, el que contempla los Jueces como efecto fijo ($modelo\ y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$) y el que lo contempla como efecto aleatorio ($modelo\ y_{ij} = \mu + \alpha_i + b_j + e_{ij}$), donde α_i es el efecto fijo debido a los ítemes y β_j o bien b_j son respectivamente el efecto –fijo o aleatorio– debido a los jueces u observadores. Ambos casos se presentan a continuación (Tablas 4.3 y 4.4).

Tabla 4.3. ANOVA en dos sentidos con efecto Jueces aleatorio para el Ejemplo 4.1

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	5.775	9	.642	$\hat{\sigma}_I^2 = (.642 - .245)/4 = .099$
Error (SC_E)	7.500	30	.250	
- Jueces (SC_J)	.875	3	.292	$\hat{\sigma}_J^2 = (.292 - .245)/10 = .005$
- Residual (SC_R)	6.625	27	.245	$\hat{\sigma}_R^2 = .245$
Total (SC_T)	13.275	39		

Tabla 4.4. ANOVA en dos sentido con efecto Jueces fijo para el Ejemplo 4.1

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	5.775	9	.642	$\hat{\sigma}_I^2 = (.642 - .245)/4 = .099$
Error (SC_E)	7.500	30	.250	
- Jueces (SC_J)	.875	3	.292	$\hat{\delta}_J'^2 = .292/10 = .029$
- Residual (SC_R)	6.625	27	.245	$\hat{\sigma}_R^2 = .245$
Total (SC_T)	13.275	39		

Es importante notar de nuevo, como ya se propuso en capítulos anteriores, la atípica estimación del efecto fijo para Jueces. El estadístico *kappa* generalizado fue definido por Conger (1980) como una alternativa exacta al estadístico de Fleiss (aunque, como hemos visto, éste puede considerarse una generalización del coeficiente *pi*), utilizando como aquél la ecuación general RCA. En esta formulación, la probabilidad observada de acuerdo p_o se obtiene de forma similar a la obtenida con la definición de Fleiss (1971), pero se emplea un procedimiento correctivo para obtener p_e^{κ} a partir del valor p_e^{π} conforme a la Ecuación 4.13, como

$$p_e^{\kappa} = p_e^{\pi} - \frac{2\,SC_J}{N\,J(J-1)} \qquad \text{(Ec. 4.19)}$$

Para los datos empíricos del Ejemplo 4.1, $p_e^{\kappa} = .322$ y la aplicación de la ecuación general RCA para estimar el coeficiente *kappa* generalizado de Conger utilizando las Ecuaciones. 4.12 y 4.18 produce como resultado

$$\hat{\kappa} = \frac{p_o - p_e^{\kappa}}{1 - p_e^{\kappa}} = \frac{.5 - .322}{.678} = .263$$

Obviamente, este resultado puede también obtenerse aplicando directamente las sumas de cuadrados de la Tabla ANOVA según la fórmula siguiente

$$\hat{\kappa} = \frac{SC_I - \frac{SC_R}{J-1}}{SC_I + SC_R + \frac{J}{J-1} SC_J} \qquad \text{(Ec. 4.20)}$$

Asumiendo un número de objetos suficientemente grande (Ii.e $N > 100$), el estimador de la Ecuación 4.19 se aproxima a la formulación del coeficiente de correlación intraclase para un modelo ANOVA en dos sentidos sin interacción, considerando que el efecto de los Jueces es fijo (véase Tabla 4.7),

$$\kappa \approx \rho_2 = \frac{\sigma_I^2}{\sigma_I^2 + \delta_J^2 + \sigma_R^2} \qquad \text{(Ec. 4.21)}$$

Esta misma solución se obtiene también mediante la teoría generalizabilidad (Li y Lautenschlager, 1997). Con los datos empíricos del Ejemplo 4.2, aplicando la Ecuación 4.18, $\hat{\kappa} = .263$, y el coeficiente de correlación sería exactamente, aplicando a la Ecuación 4.19 los estimadores de la Tabla 4.7, $\hat{\rho}_2 = .263$.

Aplicando la probabilidad estimada por azar p_e^{π} (Ecuación 4.13) de la

fórmula de Fleiss (1971) y la probabilidad p_e^{κ} (Ecuación 4.19) de la fórmula exacta de Conger, Rae (1988:373) define una **medida de homogeneidad marginal** entre varios jueces que es básicamente el complemento del coeficiente de correlación intraclase del efecto de los jueces en un modelo ANOVA en dos sentidos sin interacción

$$M = 1 - \frac{\sigma_J^2}{\sigma_I^2 + \sigma_J^2 + \sigma_R^2} = \frac{\sigma_I^2 + \sigma_R^2}{\sigma_I^2 + \sigma_J^2 + \sigma_R^2} \qquad \text{(Ec. 4.22)}$$

y puede estimarse de forma simple utilizando las probabilidades esperadas por azar en los dos coeficientes propuestos anteriormente,

$$\hat{M} = 1 - \frac{J(p_e^{\pi} - p_e^{\kappa})}{1 - p_e^{\kappa}} \qquad \text{(Ec. 4.23)}$$

La medida de homogeneidad marginal de Rae, entre otras funciones de interés, permite comprender la naturaleza de los desacuerdos entre observadores. Un grado marcado de asimetría marginal implicaría que los jueces utilizan criterios diferentes en su respuesta. Además, este estimador es una generalización para más de 2 evaluadores de una medida de homogeneidad marginal propuesta por Collis (1985) para el caso de 2 observadores.

En resumen, aunque en un principio se formularon como dos formas alternativas de generalización del coeficiente *kappa* de Cohen (1960) al caso de más de 2 evaluadores, el coeficiente propuesto por Fleiss (1971) y el

propuesto por Conger (1980), y así se siguen empleando en la práctica, pueden considerarse estimaciones diferentes de coeficientes distintos: el primero es una generalización del coeficiente *pi* de Scott y el segundo es una generalización del coeficiente *kappa* de Cohen.

Además de las generalizaciones de *pi* y de *kappa* para múltiples observadores, es posible también generalizar, en este mismo contexto, el coeficiente γ de Gwett (2001, 2008) y el coeficiente σ de Bennet y otros (1950), a más de dos observadores, particularmente en la reformulación propuesta por Maxwell (1977) y por Brennan y Prediger (1981).

Para el coeficiente *gamma* de Gwett, la probabilidad de acuerdo observada se define del mismo modo que con los coeficientes anteriores. Utilizando de nuevo las Ecuaciones 4.8 y 4.12, equivale

$$p_o = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 - NJ}{NJ(J-1)} = 1 - \frac{2SC_E}{N(J-1)} = .500$$

mientras que la probabilidad de acuerdo esperada por azar se obtiene promediando, para cada juez, las proporciones de respuesta dadas por cada uno de los jueces a cada una de las *K* categorías (p_{jk}) y aplicando después la ecuación siguiente sobre las probabilidades marginales p_{+k} y sus complementos $(1-p_{+k})$ resultantes de sumar tales proporciones para el conjunto de los jueces,

$$p_e^{\gamma} = \sum_{k=1}^{K} \frac{(p_{+k})(1-p_{+k})}{K-1} \qquad \text{(Ec. 4.24)}$$

Con los datos empíricos del Ejemplo 4.1, las probabilidades marginales para cada una de las categorías son $p_{+1} = .375$, $p_{+2} = .325$ y $p_{+3} = .300$. La probabilidad esperada por azar para el coeficiente *gamma* por tanto resulta

$$p_e^{\gamma} = \frac{(.375)(.625)+(.325)(.675)+(.300)(.700)}{3-1} = .332$$

y finalmente el estimador de γ es, aplicando la ecuación general RCA,

$$\hat{\gamma} = \frac{.500-.332}{1-.332} = .252$$

El estimador del coeficiente *gamma* es, como puede observarse, bastante similar a los dos estimadores de *kappa* que se vio anteriormente, aunque éste no es siempre el caso. La generalización al caso de 2 jueces y 2 categorías desde esta formulación general es directa y no plantea mayores problemas.

Por último, es posible sugerir igualmente una generalización del coeficiente σ para múltiples observadores aplicando la fórmula general RCA (véase Gwett, 2001; von Eye, 2004:23)

$$\hat{\sigma} = \frac{p_o - \frac{1}{K}}{1 - \frac{1}{K}} \qquad \text{(Ec. 4.25)}$$

donde p_o es la probabilidad de acuerdo observada. Para los datos empíricos del Ejemplo 4.1, con $K = 3$ categorías, resulta igual a

$$\hat{\sigma} = \frac{.500 - 1/3}{1 - 1/3} = .250$$

un valor también similar a los anteriores.

4.3.2. Varianza de los coeficientes de acuerdo para múltiples observadores

Con el objeto de aplicar la inferencia estadística a los coeficientes de acuerdo derivados en la sección anterior, es preciso disponer de una estimación de la varianza de los coeficientes. La dificultad con que nos enfrentamos al estimar una varianza para más de 2 jueces es que cuando hay 2 sólo es posible evaluar aquella porción de la varianza total debida a ítemes, pero no la variación atribuida a los jueces. Por el contrario, con más de 2 observadores pueden evaluarse tanto la variación debida a ítemes como la debida a jueces. En este contexto, se suele distinguir entre la **varianza condicional**, que se obtiene cuando el investigador desea realizar la inferencia a la población de ítemes -pero no al universo de los jueces– y la **varianza no condicional**, cuando la inferencia se desea para todo el universo, tanto de ítemes como de jueces. Nos ceñiremos aquí a la varianza condicional, que suele expresarse como varianza condicional de los ítemes, dada la muestra de jueces utilizada.

Una fuerte controversia ha rodeado el cálculo de la varianza de los coeficientes *kappa* de acuerdo para múltiples observadores. Fleiss (1971)

presentó una fórmula para obtener la varianza del coeficiente *pi* generalizado, pero se ha demostrado (Fleiss, Nee y Landis, 1979; Gwett, 2001) que produce valores excesivamente altos. Aunque también seriamente sometida a crítica, una fórmula más ajustada para el mismo coeficiente fue poco tiempo después propuesta en un trabajo de Fleiss, Nee y Landis (1979) en los términos siguientes:

$$V(\hat{\pi})=\left[\frac{2}{\left(\sum_{k=1}^{K} p_{+k}q_{+k}\right)^2 NJ(J-1)}\right]\left[\left(\sum_{k=1}^{K} p_{+k}q_{+k}\right)^2-\sum_{k=1}^{K} p_{+k}q_{+k}(q_{+k}-p_{+k})\right]$$

(Ec. 4.26)

donde p_{+k} y q_{+k} son, respectivamente, las probabilidades marginales de la categoría *k* y sus correspondientes complementos. Aplicando esta fórmula a los datos empíricos del Ejemplo 4.1 se obtiene como varianza $V(\hat{\kappa}_F)=.013$, cuya desviación típica es $S(\hat{\kappa}_F)=.114$. Pero no existe, que sepamos, ninguna fórmula similar a la de Ecuación 4.26 para calcular la varianza de los otros coeficientes de acuerdo alternativos.

Una expresión matemática para la varianza condicional del coeficiente g*amma* fue además derivada por Gwett (2001:110-4), aunque la fórmula resulta algo más complicada que la de la Ecuación 4.26 y responde a un proceso de computación que desarrollamos con detalle en la Tabla 4.8. Su formulación es general y puede también ser aplicada al caso de otros coeficientes para los que no existe una expresión apropiada, tales como π, κ y σ. Para el caso del coeficiente g*amma*, esta expresión general se

formula de la forma siguiente:

$$V(\gamma)=\left(\frac{1-f}{N}\right)\left(\frac{D^2_{\hat{\gamma}}}{N-1}\right) \qquad \text{(Ec. 4.27)}$$

donde *f* es la fracción de muestreo, que como se ha mostrado en ocasiones anteriores resulta en general ominosa, y

$$D^2_{\hat{\gamma}}=\sum_{i=1}^{N}(\hat{\gamma}_i-\bar{\gamma})^2 \qquad \text{(Ec. 4.28)}$$

expresa las diferencias cuadráticas entre los valores estimados del coeficiente g*amma* $(\hat{\gamma}_i)$ y su media $(\bar{\gamma})$ para cada uno de los objetos. La Tabla 4.8 de la siguiente sección resume con detalle todo el proceso de computación de la varianza del coeficiente *gamma* para los datos empíricos del Ejemplo 4.1. La varianza condicional del coeficiente *gamma* resulta igual a $V(\hat{\gamma})=.032$, siendo su error típico $S(\hat{\gamma})=.178$.

No obstante, la solución estadística más recomendable (Brennan y otros, 1987; Brennan, 2001;Dunn, 2004) para el cálculo de la varianza de un estimador es el procedimiento *jackknife* (Quenouille, 1949; Tukey, 1958). La gran ventaja de este procedimiento es que permite estimar la varianza de cualquier estadístico sin tener que derivar su expresión matemática ni postular distribución alguna de probabilidad. La solución es aplicable a cualquier coeficiente de acuerdo, ya sea alguno de los dos procedimientos basados en *kappa* o los coeficientes g*amma* y *sigma* tratados anteriormente.

Más adelante se detallará el procedimiento para obtener la varianza de los coeficientes de acuerdo mediante un algoritmo *jackknife*.

4.3.3. Un procedimiento de computación simplificado

Los cálculos implicados en los modelos ANOVA en uno y en dos sentidos que hemos considerado anteriormente y en los coeficientes *kappa*, *gamma* y *sigma* para múltiples jueces son complejos y requieren, tanto si se desea proceder al cálculo manual como si se desea programarlos, una estructura computacional apropiada. Con este objeto hemos desarrollado un procedimiento de computación que simplifica notablemente todos los cálculos implicados y permite comprender en mayor medida el significado del proceso con el objeto de facilitar su programación. En esta sección ilustramos dicho proceso mediante el Ejemplo 4.1, propuesto al principio del capítulo.

La estructura del procedimiento computacional se fundamenta una tabla como la Tabla 4.5, en la que se observa como se presenta en las 12 primeras filas los patrones de respuesta para cada ítem, sus frecuencias y a continuación la matriz de respuestas ítemes por categorías (n_{ik}). En las filas siguientes se exponen los patrones de respuesta para cada juez, la matriz de respuestas jueces por categorías (n_{jk}). Dependiendo del número de categorías, varias filas y columnas adicionales se utilizan para presentar un resumen de los patrones para jueces -la última fila– y para ítemes – a última columna– como resultado intermedio. De modo que, la última fila es un resumen de los

resultados para cada juez, sumando todas las categorías. Esta fila presenta uno de los resultados intermedios del proceso con el vector $\sum_{k=1}^{K} n_{jk}^2$, que se destaca con negrita y sombreado suave. La suma de los elementos del vector $\sum_{k=1}^{K} n_{jk}^2$ se expone asimismo en la última fila y es el escalar $\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}^2$, que se destaca con negrita y sombreado fuerte.

Asimismo, la última columna es un resumen de los resultados, para cada uno de los ítemes, sumando todas las categorías. Esta columna ofrece también los resultados intermedios del proceso con el vector $\sum_{k=1}^{K} n_{ik}^2$, destacados de igual forma con negrita y sombreado suave. La suma de los elementos del vector $\sum_{k=1}^{K} n_{ik}^2$ se representa a continuación en la última columna y es el escalar $\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2$, que se destaca también con negrita y sombreado fuerte. Además, en la intersección entre las categorías para las filas y para las columnas se presentan los resultados intermedios del proceso correspondientes al vector n_{+k}^2, que se destaca con negrita y sombreado suave, la suma de cuyos elementos es el escalar $\sum n_{+k}^2$ destacado asimismo con negrita y sombreado fuerte.

Igualmente, con esta misma estructura se pueden obtener las probabilidades marginales de cada una de las categorías para cada juez (p_{jk}), así como su

promedio (p_k), que son apropiadas para el cálculo de las probabilidades observadas, y las probabilidades esperadas por azar para los coeficientes de acuerdo propuestos mediante la ecuación general RCA. A continuación, se presenta detalladamente el procedimiento computacional que se sigue con esta estructura tabular para los datos empíricos del Ejemplo 4.1.

Tabla 4.5. Procedimiento directo para Ejemplo 4.1

Ítemes (I)	*Jueces (J)*				*Número de patrones (f)*	*Categorías (K)*			$\sum_{k=1}^{K} n_{ik}^2$
	1	*2*	*3*	*4*		*a*	*b*	*c*	
1	*a*	*a*	*a*	*c*	1	3	0	1	***10***
2	*a*	*a*	*b*	*c*	1	2	1	1	***6***
3	*a*	*a*	*b*	*c*	1	2	1	1	***6***
4	*a*	*a*	*c*	*c*	1	2	0	2	***8***
5	*a*	*b*	*a*	*a*	1	3	1	0	***10***
6	*b*	*a*	*a*	*a*	1	3	1	0	***10***
7	*b*	*b*	*b*	*b*	1	0	4	0	***16***
8	*b*	*c*	*b*	*b*	1	0	3	1	***10***
9	*c*	*c*	*b*	*b*	1	0	2	2	***8***
10	*c*	*c*	*c*	*c*	1	0	0	4	***16***
Categorías (K)	$n_{jk}(p_{jk})$				$n_{+k}(p_k)$	n_{+k}^2			**100**
a	5 (*.500*)	5 (*.500*)	3 (*.300*)	2 (*.300*)	15 (*.375*)	***225***			
b	3 (*.300*)	2 (*.200*)	5 (*.500*)	3 (*.300*)	13 (*.325*)		***169***		
c	2 (*.200*)	3 (*.300*)	2 (*.200*)	5 (*.500*)	12 (*.300*)			***144***	
$\sum_{k=1}^{K} n_{jk}^2$	***38***	***38***	***38***	***38***	**152**				**538**

Cálculos básicos:

$$[1]: \quad \frac{NJ}{2} = \frac{(10)(4)}{2} = 20$$

$$[2]: \quad \frac{1}{2NJ} \sum n_{+k}^2 = \frac{538}{(2)(10)(4)} = 6.725$$

$$[3]: \quad \frac{1}{2J} \sum \sum n_{ik}^2 = \frac{100}{(2)(4)} = 12.500$$

$$[4]: \quad \frac{1}{2N} \sum \sum n_{jk}^2 = \frac{152}{(2)(10)} = 7.600$$

Obtención de las fuentes de variación:

SC Total: [1] – [2] = 13.275
SC Error: [1] – [3] = 7.500
SC Ítemes: [3] – [2] = 5.775
SC Jueces: [4] – [2] = .875
SC Residual: [1] – [3] – [4] + [2] = 6.625

El ANOVA correspondiente, asumiendo efecto de los Jueces fijo, se resume en la Tabla 4.6.

Tabla 4.6. ANOVA dos sentidos Ejemplo 4.1 con efecto Jueces fijo

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	5.775	9	.642	$\hat{\sigma}_I^2 = (.642 - .245)/4 = .099$
Error (SC_E)	7.5	30	.250	
- Jueces (SC_J)	.875	3	.292	$\hat{\delta}_J'^2 = .292/10 = .029$
- Residual (SC_R)	6.625	27	.254	$\hat{\sigma}_R^2 = .245$
Total (SC_T)	13.275	39		

4.3.4. Estimación de los coeficientes de acuerdo

Aplicando las Ecuaciones 4.14 y 4.20 para el análisis de varianza con los datos del Ejemplo 4.1 se obtiene

$$\hat{\pi} = \frac{SC_I - \frac{SC_E}{J-1}}{SC_T} = .247$$

$$\hat{\kappa} = \frac{SC_I - \frac{SC_R}{J-1}}{SC_I + SC_R + \frac{J}{J-1} SC_J} = .263$$

Si alternativamente se emplean las Ecuaciones 4.12, 4.13 y 4.19 y la fórmula general RCA, siendo para cualquier coeficiente la probabilidad observada de acuerdo $p_o = .500$, la probabilidad esperada por azar para π

es $p_e^{\pi}=.336$ y la ecuación general RCA produce el mismo resultado que con ANOVA: $\hat{\pi}=(.50-.336)/(1-.336)=.247$. De modo análogo, para κ la probabilidad esperada por azar es $p_e^{\kappa}=p_e^{\pi}-2\,SC_J/(N\,J(J-1))=.3214$, y por consiguiente la ecuación general RCA produce también el mismo resultado que el obtenido con el ANOVA: $\hat{\kappa}=(.50-.3214)/(1-.3214)=.263$.

De forma similar, aplicando la fórmula general RCA, la probabilidad esperada por azar para el coeficiente γ resulta igual a $p_e^{\gamma}=[(.375)(.625)+(.325)(.675)+(.300)(.700)]/2=.332$, y en consecuencia el coeficiente se estima mediante

$$\hat{\gamma}=\frac{.50-.332}{1-.332}=.251$$

Por su parte, la fórmula general RCA aplicada al coeficiente σ, siendo su probabilidad esperada por azar $p_e^{\sigma}=1/k=.333$, produce como resultado

$$\hat{\sigma}=\frac{.500-.333}{1-.333}=.250$$

4.3.5. Cálculo de las varianzas

De cara a obtener la varianza del coeficiente *pi* generalizado puede utilizarse la Ecuación 4.26, que para el Ejemplo 4.1 resulta igual a $V(\hat{\pi})=.0084$, siendo su error típico $S(\hat{\pi})=.0914$. La expresión general para la varianza condicional derivada por Gwett (2001:115-121), que se ha formulado en las Ecuaciones. 4.27 y 4.28, se puede utilizar también para obtener la varianza condicional de todos los coeficientes de acuerdo para múltiples evaluadores, cuyo proceso de cálculo se presenta con detalle en la Tabla 4.8.

Para obtener la varianza condicional de cualquiera de los coeficientes se requiere, como se indica con anterioridad, un procedimiento algo más complejo que una simple aplicación de una expresión, que contempla además el cálculo de coeficientes específicos para cada ítem, sujeto u objeto. Los cálculos requeridos se agilizan si se parte de una estructura tabular, similar a la representación usada para la Tabla 4.5, que contenga todas las combinaciones de ítemes por categorías, donde n_{ik} es el número de observadores que responden al ítem *i* con la categoría *k.* La Tabla 4.8 responde a esta demanda e ilustra el cálculo de la varianza condicional para todos los coeficientes tratados en esta sección.

En el proceso de cálculo, las probabilidades observadas para cada ítem se obtienen, utilizando

$$p_{oi}=\frac{\sum_{k=1}^{K} n_{ik}^2 - J}{J(J-1)} \qquad \text{(Ec. 4.29)}$$

siendo su promedio

$$\bar{p}_o = \frac{\sum_{i=1}^{N} p_{oi}}{N} \qquad \text{(Ec. 4. 30)}$$

A continuación se calcula cualquiera de los coeficientes de acuerdo (*pi* de Fleiss, *kappa* de Conger, *gamma* de Gwett o la versión generalizada de *sigma* de Bennet) para cada uno de los ítemes, aplicando la ecuación general RCA. Para ello es preciso conocer la probabilidad esperada por azar p_e para cada uno de los coeficientes.

En primer lugar, las probabilidades marginales p_k de las tres categorías de respuesta del Ejemplo 4.1 se obtienen mediante

$$p_k = \frac{n_{+k}}{NJ} \qquad \text{(Ec. 4.31)}$$

y son, respectivamente, $p_1 = 15/(10 \times 4) = .375$, $p_2 = 13/(10 \times 4) = .325$ y $p_3 = 12/(10 \times 4) = .300$. En consecuencia, la probabilidad esperada por azar para el coeficiente *gamma*, –constante para todos los ítemes–, aplicando la Ecuación 4.24 resulta

$$p_e^{\gamma} = \frac{(.375)(.625)+(.325)(675)+(.300)(.700)}{2} = .332.$$

Del mismo modo, aplicando las Ecuaciones 4.9 ó 4.13 y 4.19 respectivamente se obtiene la probabilidad esperada por azar tanto para el coeficiente *pi* generalizado de Fleiss,

$$p_e^{\pi} = \frac{(15)^2 + (13)^2 + (12)^2}{(10)^2 (4)^2} = 1 - \frac{13.275}{(10)(4)} = .336$$

como la alternativa exacta de Conger,

$$p_e^{\kappa} = .336 - \frac{(2)(.875)}{(10)(4)(3)} = .322$$

Y, la probabilidad esperada por azar para el coeficiente *sigma* es sencillamente $p_e^{\sigma} = 1/K = .333$.

El cálculo de todos los coeficientes de acuerdo, para cada ítem, se muestra en la Tabla 4.8, en las columnas 7, 9, 11 y 13 respectivamente. Una vez aplicada a todos los ítemes, se obtiene posteriormente la media de los coeficientes individuales, que se detalla asimismo en la penúltima fila de las columnas indicadas, y se calculan las diferencias cuadráticas entre cada coeficiente individual y la media global. Tales diferencias se muestran en las columnas 8, 10, 12 y 14. La suma de tales diferencias cuadráticas es igual a

$D^2_{\hat{\kappa}_F} = 1.7654$, $D^2_{\hat{\kappa}_C} = 1.6903$, $D^2_{\hat{\gamma}} = 1.7424$ y $D^2_{\hat{\sigma}} = 1.750$.

Finalmente la varianza condicional de cada uno de los coeficientes, dada la muestra de tasadores utilizada, se obtiene aplicando la ecuación

$$V(\ )=\frac{D^2}{N(N-1)} \qquad \text{(Ec. 4.32)}$$

que produce como resultado $V(\hat{\pi})=.020$, $V(\hat{\kappa})=.019$, $V(\hat{\gamma})=.019$ y $V(\hat{\sigma})=.019$, siendo sus respectivos errores típicos $S(\hat{\pi})=.140$, $S(\hat{\kappa})=.137$, $S(\hat{\gamma})=.139$ y $S(\hat{\sigma})=.139$. Se expone con detalle todo el proceso computacional en la Tabla 4.8. Un conciso resumen de los resultados se muestra en la Tabla 4.7. Adviértase la notable similaridad de las varianzas para todos los coeficientes de acuerdo estimados.

Tabla 4.7. Resumen de los resultados del Ejemplo 4.1

Coeficiente	*Estimador*	*Varianza condicional*	*Error típico condicional*
Pi generalizada de Fleiss	.247	.020	.140
Kappa generalizada de Conger	.263	.019	.137
Gamma de Gwett	.252	.019	.139
Sigma de Bennett	.250	.019	.139

Tabla 4.8. Cálculo de la varianza de los coeficientes descriptivos para el Ejemplo 4.1

Ítemes	*a*	*b*	*c*	$\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$\hat{\pi}$	$(\hat{\pi}-\bar{\pi})^2$	$\hat{\kappa}$	$(\hat{\kappa}-\bar{\kappa})^2$	$\hat{y}_i$	$(\hat{y}_i-\bar{y})^2$	$\hat{\sigma}_i$	$(\hat{\sigma}_i-\bar{\sigma})^2$
1	3	0	1	10	.5000	.2467	.0000	.2629	.0000	.2516	.0000	.2500	.0000
2	2	1	1	6	.1667	-.2555	.2522	-.2285	.2415	-.2473	.2489	-.2500	.2500
3	2	1	1	6	.1667	-.2555	.2522	-.2285	.2415	-.2473	.2489	-.2500	.2500
4	2	0	2	8	.3333	-.0044	.0631	.0172	.0636	.0022	.0622	.0000	.0625
5	3	1	0	10	.5000	.2467	.0000	.2629	.0000	.2516	.0000	.2500	.0000
6	3	1	0	10	.5000	.2467	.0000	.2629	.0000	.2516	.0000	.2500	.0000
7	0	4	0	16	1.000	1.000	.5675	1.000	.5433	1.000	.5600	1.000	.5625
8	0	3	1	10	.5000	.2567	.0000	.2629	.0000	.2516	.0000	.2500	.0000
9	0	2	2	8	.3333	-.0044	.0631	.0172	.0636	.0022	.0622	.0000	.0625
10	0	0	4	16	1.000	1.000	.5675	1.000	.5433	1.000	.5600	1.000	.5625
Suma	*15*	*13*	*12*	*100*	*5.000*	*2.4670*	*1.7654*	*2.6290*	*1.6903*	*2.5164*	*1.7424*	*2.5000*	*1.7500*
Media	*.375*	*.325*	*.300*		p_o=.500	$\hat{\pi}$=.247		$\hat{\kappa}$=.263		$\hat{y}$=.252		$\hat{\sigma}$=.250	
n_{+k}^2	225	169	144		p_e^{π}=.520 p_e^{κ}=.517 p_e^{y}=.480 p_e^{σ}=.500	$V(\hat{\pi})=\frac{1.7654}{(10)(9)}=.020$ $S(\hat{\pi})$=.140		$V(\hat{\kappa})=\frac{1.6893}{(10)(9)}=.019$ $S(\hat{\kappa})$=.137		$V(\hat{y})=\frac{1.7424}{(10)(9)}=.019$ $S(\hat{y})$=.139		$V(\hat{\sigma})=\frac{1.7500}{(10)(9)}=.019$ $S(\hat{\sigma})$=.139	

4.3.6. Cálculo de las varianzas mediante el procedimiento *jackknife*

Un desarrollo similar, aunque considerablemente más complicado, es el que se sigue para obtener la varianza de los coeficientes de acuerdo a través de procedimiento de computación intensiva *jackknife* (Quenoille, 1949; Tukey, 1958). Este desarrollo es también válido para cualquiera de los coeficientes tratados en este capítulo (véase Gwett, 2001:136-146), que ilustraremos nuevamente para todos los coeficientes a continuación.

Se utiliza igualmente una estructura tabular como la de la Tabla 4.8. Todos los cálculos intermedios pueden seguirse en la Tabla 4.9 Para ello, en primer lugar, es preciso distinguir entre probabilidades observadas y estimadas de la muestra total (que son ya conocidas) y las probabilidades observadas y estimadas de la muestra replicada (que se obtienen eliminando uno a uno cada uno de los ítemes que componen la muestra).

Las **probabilidades observadas en la muestra total** se obtienen, como es usual, mediante la Ecuación 4.29:

$$p_{oi} = \frac{\sum_{k=1}^{K} n_{ik}^2 - J}{J(J-1)}$$

y las probabilidades marginales para cada categoría se obtienen en la penúltima fila de la Tabla 4.9 utilizando las sumas marginales (Ecuación 4.31). Así, para la primera categoría, $p_1 = 15/(10 \times 4) = .375$, para la segunda

$p_2 = 13/(10\times 4) = .325$ y para la tercera $p_3 = 12/(10\times 4) = .300$.

Por su parte, la **probabilidad esperada por azar en la muestra total** es una constante para todo ítem, diferente para cada uno de los coeficientes tratados aquí y que se obtiene –como tratamos anteriormente– de forma individualizada utilizando las Ecuaciones 4.9 ó 4.13 (*pi* generalizada de Fleiss: $p_e^{\pi} = .336$), la Ecuación 4.19 (*kappa* generalizada de Conger: $p_e^{\kappa} = .322$) y la Ecuación 4.24 (*gamma* de Gwett: $p_e^{\gamma} = .332$). La probabilidad esperada por azar para la *sigma* de Bennett es la más simple de obtener, ya que es el inverso del número de categorías: (s*igma* de Bennett: $p_e^{\sigma} = .333$).

Para calcular la **probabilidad observada en la muestra replicada** debe excluirse uno a uno cada uno de los ítemes de la muestra y recalcular la probabilidad observada con el caso excluido. La fórmula apropiada para obtener este resultado es la siguiente:

$$P_{oi}^{(-i)} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 - \sum_{k=1}^{K} n_{ik}^2 - [(N-1)(J)]}{(N-1)(J-1)J} \qquad \text{(Ec. 4.33)}$$

y se ha representado en la Tabla 4.9. Nótese que p_{oi} y $p_{oi}^{(-i)}$ son en esencia diferentes, aunque su suma (y por consiguiente su media) es exactamente la misma.

Del mismo modo, de cara a obtener las **probabilidades esperadas por azar para la muestra replicada** se utiliza con cada uno de los coeficientes una ecuación que, para cada ítem, es el resultado de la probabilidad esperada por

azar que se obtendría si se excluyera tal ítem de la muestra. Así, para el coeficiente *pi* generalizado de Fleiss la fórmula es (véase Tabla 4.9):

$$p_{ei}^{\pi(-i)} = \frac{\sum_{k=1}^{K} (n_{+k} - n_{ik})^2}{(N-1)^2 J^2} = \frac{\sum_{k=1}^{K} n_{+k}^{(-i)2}}{(N-1)^2 J^2} \qquad \text{(Ec.4.34)}$$

Así, para el primero y el cuarto elementos,

$$p_{e1}^{(-1)} = \frac{(15-3)^2 + (13-0)^2 + (12-1)^2}{(9^2)(4^2)} = \frac{434}{1296} = .3349$$

$$p_{e4}^{(-4)} = \frac{(15-2)^2 + (13-0)^2 + (12-2)^2}{(9^2)(4^2)} = \frac{438}{1296} = .3380$$

Por su parte, con objeto de obtener la probabilidad esperada por azar para el coeficiente *kappa* generalizado de Conger se utiliza una tabla algo diferente respecto de la Tabla 4.9 (véase Tabla 4.10) debido a la necesidad de emplear los valores n_{jk} que no son necesarios para el cálculo en los restantes coeficientes. La ecuación apropiada para este caso es la siguiente

$$p_{ei}^{\kappa(-i)} = p_{ei}^{\pi(-i)} - \frac{J \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}^{(-i)2} - \sum_{k=1}^{K} n_{+k}^{(-i)2}}{(N-1)^2 J^2 (J-1)} \qquad \text{(Ec.4.35)}$$

donde $\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}^{(-i)2}$ son escalares que representan la suma de los cuadrados de los valores n_{jk} resultantes de eliminar un caso y $\sum_{k=1}^{K} n_{+k}^{(-i)2}$ son los escalares que se utilizan en la Ecuación 4.34. Así, si se elimina el primer ítem los valores resultantes son:

$$\sum_{k=1}^{K}\sum_{j=1}^{J} n_{jk}^{(-i)2} = [(5-1)^2+3^2+2^2]+[(5-1)^2+2^2+3^2] + [(3-1)^2+5^2+2^2] = 29+29+33+29 = 120$$

y los de $\sum_{k=1}^{K} n_{+k}^{(-i)2}$ son, como se mostró anteriormente,

$$\sum_{k=1}^{K} n_{+k}^{(-i)2} = (15-3)^2+(13-0)^2+(12-1)^2 = 434$$

El cálculo de estas dos cantidades se representa también en la Tabla 4.10. En consecuencia, la probabilidad esperada por azar para el primero y el segundo ítem resulta

$$p_{ei}^{\kappa(-1)} = .3349 - \frac{(4)(120)-434}{(81)(16)(3)} = .3349 - .0118 = .3231$$

$$p_{ei}^{\kappa(-2)} = .3349 - \frac{(4)(116)-434}{(81)(16)(3)} = .3349 - .0072 = .3272$$

La probabilidad esperada por azar para la muestra replicada en el caso del coeficiente *gamma* de Gwett se obtiene utilizando la fórmula siguiente:

$$p_{ei}^{\gamma(-i)} = \frac{\sum_{k=1}^{K} [(n_{+k} - n_{ik})(NJ - J - n_{+k} + n_{ik})]}{(K-1)(NJ-J)^2} \qquad \text{(Ec. 4.36)}$$

Así, el primero y el cuarto casos resultan

$$p_{ei}^{\gamma(-1)} = \frac{(12)(24)+(13)(23)+(11)(25)}{(2)(36)^2} = .3326$$

$$p_{ei}^{\gamma(-4)} = \frac{(13)(23)+(13)(23)+(10)(26)}{(2)(36)^2} = .3310$$

Por su parte, la probabilidad esperada por azar en el caso del coeficiente *sigma* se obtiene de forma mucho más simple (véase Tabla 4.12) puesto que es siempre un valor constante, igual al inverso del número de categorías.

A continuación se procede con la estimación de los coeficientes de acuerdo con la muestra replicada, que para cada uno de los ítemes de la misma se obtienen aplicando la fórmula general RCA. Así, para el caso del coeficiente *gamma*,

$$\hat{\gamma}_i^{(i)} = \frac{p_{oi}^{(i)} - p_{ei}^{(i)}}{(1 - p_{ei}^{(i)})} \qquad \text{(Ec 4.37)}$$

Los coeficientes de acuerdo para la muestra replicada se presentan también en las Tablas 4.9, 4.10, 4.11 y 4.12.

Finalmente se obtienen las diferencias cuadráticas $(D_i^{(-i)})$ entre los valores de los coeficientes obtenidos para cada ítem mediante la fórmula RCA y su promedio y se multiplican por el número de ítemes menos uno. El resultado se ha denominado $C_i^{(-i)}$

$$C_i^{(-i)} = (N-1)\,D_i^{(-i)} \qquad \text{(Ec. 4.38)}$$

y se consigna asimismo en la última columna de las Tablas 4.9 a 4.12.

El resumen de los resultados de la aplicación del procedimiento *jackknife* a los coeficientes de acuerdo para el caso de más de 2 jueces aparece en la Tabla 4.13. Si se comparan estos resultados con los de la Tabla 4.7 –proceso de estimación estándar– se observará que, excepto en el caso del coeficiente *sigma*, donde los resultados son exactamente coincidentes, en los restantes coeficientes hay una notable similaridad tanto en la estimación de los coeficientes como en la estimación de las varianzas.

Tabla 4.13. Resumen de los resultados para Ejemplo 4.1

Coeficiente	*Estimador (jackknife)*	*Varianza (jackknife)*	*Error típico (jackknife)*
Pi generalizado de Fleiss	.242	.025	.160
Kappa generalizado de Conger	.260	.022	.148
Gamma de Gwett	.254	.017	.130
Sigma de Bennett	.250	.019	.139

Tabla 4.9: Cálculo de la varianza del coeficiente Pi generalizado de Fleiss mediante jackknife (Ejemplo 4.1)

Ítemes	*a*	*b*	*c*	$\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$p_{ei}^{(-i)}$	$\hat{\kappa}_F^{(-i)}$	$D_{\kappa_F}^{(-i)}$
1	3	0	1	10	.5000	.5000	.3349	.2483	.0003
2	2	1	1	6	.1667	.5370	.3349	.3039	.0343
3	2	1	1	6	.1667	.5370	.3349	.3039	.0343
4	2	0	2	8	.3333	.5185	.3380	.2773	.0838
5	3	1	0	10	.5000	.5000	.3333	.2500	.0005
6	3	1	0	10	.5000	.5000	.3333	.2500	.0005
7	0	4	0	16	1.000	.4444	.3472	.1489	.0783
8	0	3	1	10	.5000	.5000	.3441	.2376	.0001
9	0	2	2	8	.3333	.5185	.3441	.2659	.0050
10	0	0	4	16	1.000	.4444	.3534	.1408	.0925
n_{+k}	*15*	*13*	*12*	$\sum_{i=1}^{N} n_{ik}^2 = 100$	*5.000*	*5.000*	*3.3981*	*2.4222*	*.2545*
n_{+k}^2	225	169	144	$\sum_{k=1}^{K} n_{+k}^2 = 538$	$p_o = .500$	$\bar{p}_{oi}^{(-i)} = .500$	$\bar{p}_{ei}^{-(-i)} = .340$	$\bar{\kappa}_F^{(-i)} = .242$	$V(\hat{\kappa}_F) = .0254$
					$p_e = .336$				$S(\hat{\kappa}_F) = .160$

Tabla 4.11: Cálculo de la varianza del coeficiente Gamma de Gwett mediante jackknife (Ejemplo 4.1)

Ítemes	*Categorías*			$\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$p_{ei}^{(-i)}$	$\hat{\gamma}_F^{(-i)}$	$C_{\hat{\gamma}}^{(-i)}$
	a	*b*	*c*						
1	3	0	1	10	.5000	.5000	.3326	.2508	.0000
2	2	1	1	6	.1667	.5370	.3326	.3064	.0249
3	2	1	1	6	.1667	.5370	.3326	.3064	.0249
4	2	0	2	8	.3333	.5185	.3310	.2803	.0063
5	3	1	0	10	.5000	.5000	.3333	.2500	.0001
6	3	1	0	10	.5000	.5000	.3333	.2500	.0001
7	0	4	0	16	1.000	.4444	.3264	.1753	.0555
8	0	3	1	10	.5000	.5000	.3279	.2560	.0000
9	0	2	2	8	.3333	.5185	.3279	.2836	.0080
10	0	0	4	16	1.000	.4444	.3233	.1790	.0503
Suma	*15*	*13*	*12*	*100*	*5.000*	*5.000*	*3.3009*	*2.5377*	*.1702*
Media	*.375*	*.325*	*.300*		p_o=.500	$\bar{p}_{oi}^{(-i)}$=.500		$\bar{\gamma}^{(-i)}$=.254	$V(\hat{\kappa}_F)$=.0170
n_{+k}^2	225	169	144		p_e=.336		$\bar{p}_{ei}^{-(-i)}$=.330		$S(\hat{\kappa}_F)$=.130

Tabla 4.12: Cálculo de la varianza del coeficiente Sigma de Bennett mediante jackknife (Ejemplo 4.1)

Ítemes	*Categorías*			$\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$p_{ei}^{(-i)}$	$\hat{\sigma}^{(-i)}$	$C_{\hat{\sigma}}^{(-i)}$
	a	*b*	*c*						
1	3	0	1	10	.5000	.5000	.3333	.2500	.0000
2	2	1	1	6	.1667	.5370	.3333	.3056	.0278
3	2	1	1	6	.1667	.5370	.3333	.3056	.0278
4	2	0	2	8	.3333	.5185	.3333	.2778	.0069
5	3	1	0	10	.5000	.5000	.3333	.2500	.0000
6	3	1	0	10	.5000	.5000	.3333	.2500	.0000
7	0	4	0	16	1.000	.4444	.3333	.1667	.0625
8	0	3	1	10	.5000	.5000	.3333	.2500	.0000
9	0	2	2	8	.3333	.5185	.3333	.2778	.0069
10	0	0	4	16	1.000	.4444	.3333	.1667	.0625
Suma	*15*	*13*	*12*	*100*	*5.000*	*5.000*	*3.3333*	*2.500*	*.1944*
Media	*.375*	*.325*	*.300*		p_o=.500	$\bar{p}_{oi}^{(-i)}$=.500		$\bar{\sigma}^{(-i)}$=.250	$V(\hat{\sigma})$=.0194
n_{+k}^2	225	169	144		p_e=.336		$\bar{p}_{ei}^{-(-i)}$=.333		$S(\hat{\sigma})$=.1394

La varianza de cada uno de los coeficientes mediante el algoritmo *jackknife* es el promedio de los valores de $C_i^{(-i)}$. Así, para el coeficiente *gamma*, la varianza *jackknife* es

$$V(\hat{\gamma}) = \frac{C_i^{(i)}}{N} \qquad \text{(Ec .4.39)}$$

4.4. Ejemplo 4.2: los datos de von Eye (2005)

La Tabla 4.14 corresponde al Ejemplo 4.2 (von Eye, 2005:28), donde N = 15 ítemes, son clasificados por J = jueces utilizando una variable de respuesta binaria K, con valores *a* y *b*.

Tabla 4.14. Ejemplo 4.2: representación directa

Jueces (J)				
Ítemes (I)	*1*	*2*	*3*	*Frecuencias*
1	*a*	*a*	*a*	1
2	*a*	*a*	*b*	1
3	*a*	*b*	*b*	1
4	*a*	*a*	*a*	1
5	*b*	*b*	*b*	1
6	*a*	*a*	*a*	1
7	*b*	*a*	*a*	1
8	*b*	*b*	*b*	1
9	*a*	*b*	*b*	1
10	*a*	*a*	*a*	1
11	*a*	*b*	*a*	1
12	*b*	*a*	*a*	1
13	*a*	*a*	*a*	1
14	*b*	*b*	*b*	1
15	*a*	*b*	*a*	1

Nótese que solo hay $K^J = 2^3 = 8$ patrones de respuesta posibles, con frecuencias 0 en dos de ellos, y por tanto sería posible simplificar notablemente el ejemplo presentando solamente los patrones que tienen alguna frecuencia con el objeto de reducir el esfuerzo de computación. La Tabla 4.15 responde a esta simplificación.

Tabla 4.15. Ejemplo 4.2: representación abreviada

Ítemes (I)	*Jueces (J)*			*Número de patrones (f)*
	1	*2*	*3*	
1-5	*a*	*a*	*a*	5
6	*a*	*a*	*b*	1
7-8	*a*	*b*	*a*	2
9-10	*a*	*b*	*b*	2
11-12	*b*	*a*	*a*	2
13-15	*b*	*b*	*b*	3

La Tabla 4.16 resume el procedimiento abreviado para obtener todos los escalares básicos con los datos empíricos del Ejemplo 4.2. Se trata de una tabla de acuerdo para $N = 15$ ítemes, $J = 3$ jueces y $K = 2$ categorías de respuesta, pero el procedimiento abreviado solamente contempla seis patrones ya que son los únicos registrados. Véase que el único cambio requerido concierne al papel que desempeñan ahora las frecuencias (*f*), que representan el número de veces que se repiten los patrones de respuesta registrados. El cambio afecta tanto al listado de los ítemes –cuya numeración se corresponde con las frecuencias acumuladas– como a reflejar el cálculo de las dos últimas columnas, donde $45 = (5)\,[(3)^2 + (0)^2]$, $5 = (1)\,[(2)^2 + 1^2]$,.., y $27 = (3)\,[(3)^2 + (0)^2]$. Sin embargo, como puede observarse, no quedan en absoluto afectados los resultados básicos.

Tabla 4.16. Procedimiento abreviado para Ejemplo 4.2

Ítemes (I)	*Jueces (J)*			*Número de patrones (f)*	*Categorías (K)*		$f\sum_{k=1}^{K} n_{ik}^2$
	1	*2*	*3*		*a*	*b*	
1-5	*a*	*a*	*a*	5	3	0	***45***
6	*a*	*a*	*b*	1	2	1	***5***
7-8	*a*	*b*	*a*	2	2	1	***10***
9-10	*a*	*b*	*b*	2	1	2	***10***
11-12	*b*	*a*	*a*	2	2	1	***10***
13-15	*b*	*b*	*b*	3	0	3	***27***
Categorías (K)	$f\,n_{jk}(p_{jk})$			$n_{+k}(p_{+k})$	n_{+k}^2		**107**
a	10 (.670)	8 (.530)	9 (.600)	27 (.600)	***729***		
b	5 (.330)	7 (.470)	6 (.400)	18 (.400)		***324***	
$\sum_{k=1}^{K}(f\,n_{jk})^2$	***125***	***113***	***117***	**355**			**1053**

Cálculos básicos:

$$[1]:\quad \frac{NJ}{2}=\frac{(15)(3)}{2}=22.500$$

$$[2]:\quad \frac{1}{2\,NJ}\sum_{k=1}^{K} n_{+k}^2=\frac{1053}{(2)(15)(3)}=11.700$$

$$[3]: \quad \frac{1}{2J}\sum_{i=1}^{N}\left(f\sum_{k=1}^{K}n_{ik}^{2}\right)=\frac{107}{(2)(3)}=17.833$$

$$[4]: \quad \frac{1}{2N}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(f\,n_{jk}\right)^{2}=\frac{355}{(2)(15)}=11.833$$

Obtención de las fuentes de variación:

SC Total: [1] – [2] = 10.800

SC Error: [1] – [3] = 4.667

SC Ítemes: [3] – [2] = 6.133

SC Jueces: [4] – [2] = .133

SC Residual: [1] – [3] – [4] + [2] = 4.534

El ANOVA correspondiente, asumiendo el efecto de los Jueces fijo, se resume en la Tabla 4.17

Tabla 4.17. ANOVA dos sentidos para Ejemplo 4.1 con efecto Jueces fijo

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	6.133	14	.438	$\hat{\sigma}_I^2=(.438-.162)/4=.092$
Error (SC_E)	4.667	30	.156	
- Jueces (SC_J)	.133	2	.067	$\hat{\delta}_J'^2=.067/15=.004$
- Residual (SC_R)	4.534	28	.162	$\hat{\sigma}_R^2=.162$
Total (SC_T)	10.800	44		

4.4.1 Estimación de los coeficientes de acuerdo

En primer lugar, el coeficiente *pi* generalizado (Fleiss, 1971) se estima utilizando la Ecuación 4.14 y el coeficiente k*appa* generalizado (Conger, 1984) empleando la Ecuación 4.20. Aplicando tales ecuaciones a los datos del Ejemplo 4.2 se obtiene, respectivamente,

$$\hat{\pi} = \frac{SC_I - \frac{SC_E}{J-1}}{SC_T} = .352$$

$$\hat{\kappa} = \frac{SC_I - \frac{SC_R}{J-1}}{SC_I + SC_R + \frac{J}{J-1} SC_J} = .356$$

Recuérdese que también se pueden obtener ambos coeficientes mediante la fórmula general RCA de corrección del azar utilizando las Ecuaciones 4.12, 4.13 y 4.19. Siendo la probabilidad observada de acuerdo $p_o = 1 - [2\,SC_E]/[N(J-1)] = .689$ y la probabilidad esperada por azar $\bar{p}_e = 1 - [(2)(SC_T)]/NJ = .520$, para el coeficiente *pi* generalizado la fórmula RCA arroja como resultado $\hat{\pi} = (.689 - .520)/(1 - .520) = .352$. Del mismo modo, para el coeficiente *kappa* generalizado la probabilidad esperada por azar es $p_e = \bar{p}_e - 2\,SC_J/(N\,J\,(J-1)) = .517$ y la fórmula general RCA resulta igual a $\hat{\kappa} = (.689 - .517)/(1 - .517) = .356$.

Igualmente puede estimarse el coeficiente γ de Gwett (2001, 2008)

aplicando la ecuación RCA. Siendo $p_o = .689$, y consultando la Tabla 4.17, que contiene tanto las probabilidades marginales para las combinaciones de jueces por categorías (p_{jk}) como las probabilidades marginales por categorías (p_k) en las últimas filas de la tabla, puede utilizarse la Ecuación 4.22 para obtener –puesto que solo hay dos categorías– la probabilidad esperada por azar $p_e = (.600)(.400) + (.400)(.600) = .480$ .Por último la aplicación de la fórmula RCA produce como resultado

$$\hat{y} = \frac{.689 - .480}{1 - .480} = .402$$

un valor muy similar a los resultantes con las versiones generalizadas de los coeficientes *pi* y *kappa*.

Y finalmente para estimar la versión generalizada para múltiples jueces y múltiples categorías del coeficiente σ se aplica la Ecuación 4.25. Puesto que existen 2 categorías, la probabilidad esperada por azar es $1/k = .500$. Así, para los datos del Ejemplo 4.2,

$$\hat{\sigma} = \frac{.689 - .500}{1 - .500} = .378$$

4.4.2. Cálculo de las varianzas de los coeficientes de acuerdo

Para obtener la varianza del coeficiente *pi* generalizado puede aplicarse, aunque no es de uso común, la Ecuación 4.26, que resulta igual a $V(\hat{\pi})=.013$, siendo su error típico $S(\hat{\pi})=.114$.

Más conveniente es que todos los coeficientes se obtengan aplicando la expresión derivada por Gwett (2001), cuyo proceso de cálculo es más complejo y se detalla en la Tabla 4.19 Los cálculos requeridos se simplifican si se parte de una estructura tabular que contenga todas las combinaciones de ítemes por categorías, donde n_{ik} es el número de jueces que responden al ítem *i* con la categoría *k*. La Tabla 4.19 refleja esta demanda e ilustra el cálculo de la varianza para todos los coeficientes de acuerdo en el caso de múltiples jueces u observadores.

En el proceso de computación, las probabilidades observadas de cada ítem se obtienen, , utilizando la Ecuación 4.29 para el cálculo de las probabilidades directas, y la Ecuación 4.30 para calcular su promedio.

A continuación se calculan los coeficientes correspondientes para cada uno de los ítemes, aplicando la ecuación general RCA. Para ello se precisa conocer la probabilidad esperada por azar p_e, que en todos los casos es una constante que se obtiene mediante las Ecuaciones 4.9 ó 4.13 (*pi* generalizado de Fleiss), la Ecuación 4.19 (*kappa* generalizada de Conger) y la Ecuación 4.24 (*gamma* de Gwett). Siendo las probabilidades marginales p_k de ambas categorías $p_1=27/(15)(3)=.600$ y $p_2=18/(15)(3)=.400$ respectivamente las probabilidades esperadas por azar para los diferentes coeficientes de acuerdo

son $p_e^{\pi}=.520$, $p_e^{\kappa}=.517$, $p_e^{\gamma}=.480$ y $p_e^{\sigma}=.500$.

Los coeficientes de acuerdo correspondientes a cada ítem se obtienen aplicando la fórmula RCA y se representan en la Tabla 4.19 Una vez aplicada a todos los ítemes, se obtiene la media de cada uno de los coeficientes *gamma* individuales, que se detallan también en la fila de sumas de la Tabla 4.19 y se calculan las diferencias cuadráticas entre cada coeficiente individual y su media. La suma de tales diferencias cuadráticas se representa también a continuación de cada uno de los coeficientes. Repárese que las sumas de las probabilidades no se corresponden directamente con las sumas de las probabilidades de cada uno de los patrones; es preciso en este caso utilizar las frecuencias de cada patrón para obtener la suma total.

Finalmente, la varianza de cada uno de los coeficientes se obtiene aplicando la Ecuación 4.32 que produce como resultado $V(\hat{\pi})=.034$ para el coeficiente *pi* generalizado de Fleiss, $V(\hat{\kappa})=.034$ para el coeficiente *kappa* de Conger, $V(\hat{\gamma})=.029$ para el coeficiente g*amma* y $V(\hat{\sigma})=.032$ para el coeficiente *sigma,* siendo sus respectivos errores típicos $S(\hat{\pi})=.185$, $S(\hat{\kappa}_F)=.184$, $S(\hat{\gamma})=.171$ y $S(\hat{\sigma})=.178$ (véase Tabla 4.20).

A continuación se muestra un resumen de todos los coeficientes descriptivos de acuerdo a los que pueden aplicarse los procedimientos inferenciales convencionales, (Tabla 4.18).

Tabla 4.18. Resumen del Ejemplo 4.2

Coeficiente	*Estimador*	*Varianza*	*Error típico*
Pi generalizada	.352	.034	.185
Kappa generalizada	.356	.034	.184
Gamma	.402	.029	.171
Sigma	.378	.032	.178

4.4. 3. Cálculo de las varianzas mediante el procedimiento *jackknife*

De forma similar se procede para obtener la varianza de los coeficientes de acuerdo mediante el procedimiento de computación intensiva *jackknife*, que tratamos más arriba. Como se vio anteriormente, el desarrollo es también válido para cualquiera de los coeficientes tratados en este capítulo, y lo ilustramos aquí de nuevo para los datos empíricos del Ejemplo 4.2, pero en este caso utilizando la representación abreviada. Se emplea igualmente una estructura tabular como la de la Tabla 4.19.

Tabla 4.19. Cálculo de la varianza de los coeficientes descriptivos, procedimiento abreviado (Ejemplo 4.2)

Patrones	*a*	*b*	*f*	$f\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$\hat{\pi}$	$(\hat{\pi}-\bar{\pi})^2$	$\hat{\kappa}$	$(\hat{\kappa}-\bar{\kappa})^2$	$\hat{y}_i$	$(\hat{y}_i-\bar{y})^2$	$\hat{\sigma}_i$	$(\hat{\sigma}_i-\bar{\sigma})^2$
1	3	0	5	45	1.000	1.000	.4202	1.000	.4149	1.000	.3576	1.000	.3872
2	2	1	1	5	.3333	-.3890	.5288	-.3803	.5419	-.2821	.4676	-.3333	.5056
3	2	1	2	10	.3333	-.3890	.5288	-.3803	.5419	-.2821	.4676	-.3333	.5056
4	1	2	2	10	.3333	-.3890	.5288	-.3803	.5419	-.2821	.4676	-.3333	.5056
5	2	1	2	10	.3333	-.3890	.5288	-.3803	.5419	-.2821	.4676	-.3333	.5056
6	0	3	3	27	1.000	1.000	.4202	1.000	.4149	1.000	.3576	1.000	.3872
Suma	*27*	*18*		*107*	*10.333*	*5.2773*	*7.2031*	*5.3379*	*7.1125*	*6.0253*	*6.1339*	*5.6667*	*6.6370*
Media	*.600*	*.400*			p_o=.689	$\bar{\pi}$=.352		$\bar{\kappa}$=.356		$\bar{y}$=.402		$\bar{\sigma}$=.378	
n_{+k}^2	729	324			p_e^{π}=.520 p_e^{κ}=.517 p_e^{y}=.480 p_e^{σ}=.500	$V(\hat{\pi})=\frac{7.2031}{(15)(14)}=.034$		$V(\hat{\kappa})=\frac{7.1125}{(15)(14)}=.034$		$V(\hat{y})=\frac{6.1339}{(15)(14)}=.029$		$V(\hat{\sigma})=\frac{6.6470}{(15)(14)}=.032$	
						$S(\hat{\pi})$=.185		$S(\hat{\kappa})$=.184		$S(\hat{y})$=.171		$S(\hat{\sigma})$=.178	

Adviértase que el número de patrones necesario para la representación abreviada no es el mismo en todos los coeficientes. Es igual para los coeficientes *pi* generalizado, *gamma* y *sigma,* pero es diferente para *kappa* generalizado. Esta diferencia se debe al hecho de que para obtener el coeficiente *kappa* generalizado se precisa disponer tanto de la representación por ítemes como de la representación por jueces .

Las **probabilidades observadas en la muestra total** se obtienen, como es usual, mediante la Ecuación 4.29 y las probabilidades marginales para cada categoría se representan en la penúltima fila de la Tabla 4.20, utilizando las sumas marginales (Ecuación 4.31). Puesto que solamente hay dos categorías, la probabilidad marginal para la primera es $p_1 = 27/(15 \times 3) = .600$ y para la segunda es $p_2 = 18/(15 \times 3) = .400$.

Por su parte, la **probabilidad esperada por azar en la muestra total** es una constante para todo ítem, pero es diferente para cada uno de los coeficientes tratados aquí, que se obtiene – como hemos visto en la sección anterior– de forma individualizada utilizando las Ecuaciones 4.9 ó 4.13 (*pi* generalizada de Fleiss: $p_e^{\kappa} = .336$), la Ecuación 4.19 (*kappa* generalizada de Conger: $p_e^{\kappa} = .322$) y la Ecuación 4.24 (*gamma* de Gwett: $p_e^{\gamma} = .332$). La probabilidad esperada por azar para la *sigma* de Bennett es la más simple de obtener, por ser el inverso del número de categorías: (*sigma* de Bennett: $p_e^{\sigma} = .333$).

En cuanto a la **probabilidad observada en la muestra replicada,** se obtiene aplicando la Ecuación 4.33. Para ello debe excluirse uno a uno cada uno de los sujetos de la muestra y recalcular la probabilidad observada con el caso excluido, tal y como se ha representado en las Tablas 4.20 a 4.23.

Tabla 4.20. Cálculo de la varianza del coeficiente Pi generalizado de Fleiss mediante jackknife (Ejemplo 4.2)

Patrones	*Categorías*		f	$\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$p_{ei}^{(-i)}$	$\hat{\pi}^{(-i)}$	$C_{\pi}^{(-i)}$
	a	*b*							
1	3	0	5	9	1.000	.6667	.5102	.3194	.0127
2	2	1	5	5	.3333	.7143	.5181	.4071	.0463
3	1	2	2	5	.3333	.7143	.5283	.3942	.0278
4	0	3	3	9	1.000	.6667	.5408	.2741	.0797
n_{+k}	*27*	*18*		$\sum_{i=1}^{N}\sum_{k=1}^{K} n_{ik}^2 = 107$	*10.333*	*10.333*	*7.8205*	*5.2442*	*.5901*
n_{+k}^2	729	324		$\sum_{k=1}^{K} n_{+k}^2 = 1053$	$p_o = .689$	$\bar{p}_o^{(-i)} = .689$	$\bar{p}_{ei}^{-(-i)} = .521$	$\bar{\pi}^{(-i)} = .350$	$V(\hat{\pi}) = .039$
p_k	.600	.400			$p_e^{\pi} = .520$				$S(\hat{\pi}) = .198$

Tabla 4.21: Cálculo de la varianza del coeficiente Kappa generalizado de Conger mediante jackknife (Ejemplo 4.1)

Patrones	*f*	*Jueces*			*Categorías*		$f\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$\sum\sum n_{jk}^{(-i)2}$	$\sum n_{+k}^{(-i)2}$	$p_{ei}^{(-i)}$	$\hat{\kappa}^{(-i)}$	$C_{\kappa}^{(-i)}$
		1	*2*	*3*	*a*	*b*								
1	5	*a*	*a*	*a*	3	0	***9***	1.000	.6667	304	900	.5068	.3242	.0127
2	1	*a*	*a*	*b*	2	1	***5***	.3333	.7143	310	914	.5136	.4127	.0477
3	2	*a*	*b*	*a*	2	1	***5***	.3333	.7143	306	914	.5170	.4085	.0412
4	2	*a*	*b*	*b*	1	2	***5***	.3333	.7143	312	932	.5272	.3958	.0241
5	2	*b*	*a*	*a*	2	1	***5***	.3333	.7143	314	914	.5102	.4167	.0546
6	3	*b*	*b*	*b*	0	3	***9***	1.000	.6667	322	954	.5374	.2795	.0783
Total	15	*Jueces*					**107**	*10.3333*	*10.3333*			*7.7683*	*6.119*	*.5857*
Categorías	*a*	10	8	9	27		***729***	p_o=.689	$p_o^{(-i)}$=.689			$p_e^{(-i)}$=.518	*.3543*	
	b	5	7	6		18	***324***	p_e^{π}=.520					$V(\hat{\kappa}^{(-i)})$=.039	
$\sum_{j=1}^{J} n_{jk}^2$		***125***	***113***	***117***	**355**		**1053**						$S(\hat{\kappa}^{(-i)})$=.198	

Tabla 4.22: Cálculo de la varianza del coeficiente Gamma de Gwett mediante jackknife (Ejemplo 4.1)

Patrones	*Categorías*		f	$f\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$p_{ei}^{(-i)}$	$\hat{\gamma}^{(-i)}$	$C_{\gamma}^{(-i)}$
	a	*b*							
1	3	0	5	45	1.000	.6667	.4898	.3497	.0446
2	2	1	5	25	.3333	.7143	.4819	.4486	.0290
3	1	2	2	10	.3333	.7143	.4717	.4592	.0442
4	0	3	3	27	1.000	.6667	.4592	.3836	.0053
n_{+k}	*27*	*18*		$\sum_{i=1}^{N} n_{ik}^2 = 107$	*10.333*	*10.333*	*7.1791*	*6.0460*	*.4718*
n_{+k}^2	729	324		$\sum_{k=1}^{K} n_{+k}^2 = 1053$	$p_o = .689$	$\bar{p}_o^{(-i)} = .689$	$\bar{p}_{ei}^{-(-i)} = .479$	$\bar{\gamma}^{(-i)} = .403$	$V(\hat{\gamma}) = .0315$
					$p_e^{\pi} = .480$				$S(\hat{\gamma}) = .177$

Tabla 4.23: Cálculo de la varianza del coeficiente Sigma de Bennett mediante jackknife (Ejemplo 4.1)

Patrones	*Categorías*		f	$f\sum_{k=1}^{K} n_{ik}^2$	p_{oi}	$p_{oi}^{(-i)}$	$p_{ei}^{(-i)}$	$\hat{\sigma}^{(-i)}$	$C_{\sigma}^{(-i)}$
	a	*b*							
1	3	0	5	45	1.000	.6667	.5000	.3334	.0276
2	2	1	5	25	.3333	.7143	.5000	.4286	.0361
3	1	2	2	10	.3333	.7143	.5000	.4286	.0361
4	0	3	3	27	1.000	.6667	.5000	.3334	.0276
n_{+k}	*27*	*18*		$\sum_{i=1}^{N} n_{ik}^2 = 107$	*10.333*	*10.333*	*7.500*	*5.667*	*.4735*
n_{+k}^2	729	324		$\sum_{k=1}^{K} n_{+k}^2 = 1053$	$p_o = .689$	$\bar{p}_o^{(-i)} = .689$	$\bar{p}_{ei}^{-(-i)} = .500$	$\bar{\sigma}^{(-i)} = .378$	$V(\hat{\sigma}) = .0316$
					$p_e^{\pi} = .500$				$S(\hat{\sigma}) = .178$

Sin embargo, los valores resultan iguales en todos los coeficientes. Así, para obtener las probabilidades observadas del primero y el segundo casos se utiliza

$$p_{o1}^{(-1)} = \frac{107-9-42}{(14)(2)(3)} = .6667$$
$$p_{o2}^{(-2)} = \frac{107-5-42}{(14)(2)(3)} = .7143$$

Nótese que p_{oi} y $p_{oi}^{(-i)}$ son diferentes, aunque su suma –y por consiguiente su media– es exactamente la misma. Pero para calcular la suma debe ponderarse cada coeficiente por el número de patrones (*f*) que representa; así,

$$\sum_{i=1}^{N} p_{oi}^{(-i)} = (5)(.6667)+(5)(.7143)+(2)(.7143)+(3)(.6667) = 10.3333.$$

Igualmente, a fin de calcular las **probabilidades esperadas por azar para la muestra replicada** se utiliza para cada uno de los coeficientes una ecuación que, para cada ítem, es el resultado de la probabilidad esperada por azar que se obtendría si se excluyera tal ítem de la muestra. Así, para el coeficiente *kappa* generalizado de Fleiss se utiliza la Ecuación 4.34, siendo el primero y segundo elementos

$$p_{e1}^{\pi(-1)} = \frac{(27-3)^2+(18-0)^2}{(14)^2(3)^2} = .5102$$

$$p_{e2}^{\pi(-2)} = \frac{(27-2)^2+(18-1)^2}{(14)^2(3)^2} = .5181$$

Por su parte, con objeto de obtener la probabilidad esperada por azar para el coeficiente *kappa* generalizado de Conger se requiere ampliar la tabla anterior para incluir la matriz de jueces por categorías (Tabla 4.21). Además, el número de patrones requerido no es el mismo que el de la Tabla 4.20, debido a la necesidad de utilizar los valores n_{jk}. Una vez obtenidos los escalares $\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}^{(-i)2}$ y $\sum_{k=1}^{K} n_{+k}^{(-i)2}$, que aparecen en la Tabla 4.21 para cada uno de los patrones representados, se aplica la Ecuación 4.35 sobre los dos primeros patrones y resulta

$$p_{e1}^{\kappa(-1)} = .5102 - \frac{(3)(304)-900}{2\times(14)^2\times 3^2} = .5068$$

$$p_{e2}^{\kappa(-2)} = .5181 - \frac{(3)(310)-914}{2\times(14)^2\times 3^2} = .5136$$

La probabilidad esperada por azar para la muestra replicada en el caso del coeficiente *gamma* de Gwett se obtiene, de forma similar a los anteriores, utilizando la Ecuación 4.36 (véase Tabla 4.22). Para los dos primeros patrones de la Tabla 4.22,

$$p_{e1}^{\gamma(-1)} = \frac{(24)(18)+(18)(24)}{(1)(42)^2} = .4898$$

$$p_{e2}^{\gamma(-2)} = \frac{(25)(17)+(17)(25)}{(2)(36)^2} = .4819$$

Por su parte, la probabilidad esperada por azar en el caso del coeficiente *sigma* se obtiene de forma mucho más simple puesto que es siempre un valor constante, e igual al inverso del número de categorías(véase Tabla 4.23).

Después se procede a estimar los coeficientes de acuerdo para la muestra replicada, que para cada uno de los ítemes se consiguen aplicando la fórmula general RCA. De manera que, para el caso del coeficiente *gamma*,

$$\hat{\gamma}_i^{(i)} = \frac{p_{oi}^{(i)} - p_{ei}^{(i)}}{(1 - p_{ei}^{(i)})} \qquad \text{(Ec. 4.37)}$$

Los coeficientes de acuerdo para la muestra replicada se representan también en las Tablas 4.20 a 4.23.

En último lugar se obtienen las diferencias cuadráticas $(D_i^{(-i)})$ entre los valores de los coeficientes obtenidos para cada ítem mediante la fórmula RCA y su promedio, y se multiplican por el número de ítemes menos uno. El resultado es la Ecuación 4.38, cuyo valor se consigna en la última columna de las Tablas 4.20 a 4.23. Nuevamente, para obtener la suma de los valores de $C_i^{(-i)}$ para cada uno de los patrones es preciso ponderar cada valor por la

frecuencia (*f*) de cada patrón.

La varianza de cada uno de los coeficientes a través de el procedimiento *jackknife* es, finalmente, el promedio de los valores de $C_i^{(-i)}$ una vez que han sido ponderados. Así, para el coeficiente *gamma*, la varianza *jackknife* es

$$V(\hat{\gamma}) = \frac{.4718}{15} = .0315$$

El resumen de los resultados de la aplicación del algoritmo *jackknife* a los coeficientes de acuerdo para el caso de más de 2 jueces se resume en la Tabla 4.24.

Tabla 4.24. Resumen del procedimiento jackknife con el Ejemplo 4.2

Coeficiente	*Estimador (jackknife)*	*Varianza (jackknife)*	*Error típico (jackknife)*
Pi generalizado de Fleiss	.350	.039	.198
Kappa generalizado de Conger	.354	.039	.198
Gamma de Gwett	.403	.032	.177
Sigma de Bennett	.378	.032	.178

Si se comparan de nuevo estos resultados con los de la Tabla 4.19 –obtenidos mediante el proceso de estimación estándar– se observará que, excepto excepto en el caso del coeficiente *sigma*, donde los resultados son exactamente coincidentes, en los restantes coeficientes hay una notable similaridad tanto en la estimación de los coeficientes como en la estimación de las varianzas.

Capítulo 5

El coeficiente de acuerdo *iota*

5.1. Introducción

Los coeficientes descriptivos tratados en los Capítulos 2, 3 y 4 tienen en común utilizar una fórmula convencional de corrección del azar –la ecuación RCA– para obtener sus estimaciones, basándose en dos "deseables" propiedades que debe poseer una medida de acuerdo:

1) ser capaz de detectar la magnitud verdadera del acuerdo, más allá de lo que se espera por azar, y

2) ser directamente aplicable a la evaluación de la fiabilidad.

Sin embargo, ésta no es la única vía existente para desarrollar medidas de acuerdo entre jueces. Aparte de otros enfoques válidos para evaluar el acuerdo, entre los cuales cabe destacar una propuesta del grupo de Koch (Landis y Koch, 1977 a, b, c), basada en la metodología *GSK* (e.g. véase Ato y López, 1996), una línea alternativa que ha sido empleada en ocasiones en la literatura psicológica se fundamenta en un enfoque propuesto por Berry y Mielke (1988: 922), quienes señalaron una tercera propiedad también "deseable" de una medida de acuerdo, a saber,

3) resultar aplicable a cualquier nivel de medición.

Además, hay otras dos propiedades deseables que, en opinión de Berry y Mielke (1988:922), reúnen interés estadístico para constituir una medida de acuerdo, a saber:

4) tener una base estadística, i.e., permitir desarrollar un test de significación estadística apropiado, ya que de no ser así estaría severamente limitada en situaciones prácticas, y
5) ser capaz de analizar datos multivariantes, esto es, no limitarse a valorar un único criterio de respuesta.

Es obvio que ninguno de los coeficientes descriptivos tratado hasta ahora cumplen con la tercera propiedad, ni con las restantes propiedades deseables anteriormente citadas. Con el propósito de solucionar esta problemática, Berry y Mielke (1988) redefinieron el clásico coeficiente Kappa de Cohen en un nuevo coeficiente, que posteriormente Janson y Olsson (2001, 2002) han llamado **iota**, y que formularon como el complemento de la razón entre el desacuerdo observado y el desacuerdo estimado, mediante

$$\iota = 1 - \frac{\delta_o}{\delta_e} \qquad \text{(Ec. 5.1)}$$

donde δ_o representa la distancia euclídea media entre la valoración de 2 observadores de un mismo ítem y es la proporción de desacuerdo observado, mientras que δ_e representa la distancia entre la valoración de un juez de un ítem específico y la valoración de cualquier otro juez de cualquier otro ítem, siendo la proporción de desacuerdo esperado. En la Ecuación 5.1, $\delta_o = 1 - P_o$ representa en parte el numerador de la ecuación RCA y $\delta_e = 1 - P_e$ es la expresión exacta del denominador de la ecuación RCA utilizado en la formulación de *kappa;* por tanto puede demostrarse fácilmente la equivalencia

$$\iota = 1 - \frac{1 - P_o}{1 - P_e} = \frac{1 - P_e - 1 + P_o}{1 - P_e} = \kappa \qquad \text{(Ec. 5.2)}$$

5.2. Generalización a *kappa* para dos observadores

Llamando y_{iA} e y_{iB} a las categorías seleccionadas por 2 observadores –identificados como *A* y *B* para simplificar la exposición– Berry y Mielke (1988) y Janson y Olsson (2001) definieron δ_o en los términos siguientes:

$$\delta_o = \frac{1}{N}\sum_{i=1}^{N} \Delta(y_{iA}, y_{iB}) \qquad \text{(Ec. 5.3)}$$

donde $\Delta(y_{iA}, y_{iB}) = \sqrt{\sum_{k=1}^{K} (y_{iA} - y_{iB})^2}$ es el desacuerdo producido en la selección de las categorías para el ítem *i* entre los 2 jueces. En consecuencia, dado un par de jueces, δ_o es la media –para el conjunto de los ítemes– de las distancias euclídeas cuadráticas entre la selección de las categorías de los mismos ítemes realizada por ambos jueces, que puede tomar dos valores: igual a 0 $(si\ y_{iA} = y_{iB})$; o bien igual a 1 $(si\ y_{iA} \neq y_{iB})$.

La Tabla 5.1 presenta un ejemplo simple con $N = 5$ ítemes, $J = 3$ jueces y $K = 3$ categorías de respuesta por ítem (que etiquetamos como *a*, *b* y *c*), pero nuestro interés se centra en un primer momento en los dos primeros jueces. Siendo las categorías seleccionadas por el juez *A* para los cinco ítemes [*a, b, b, a, c*] y las del juez *B* [*a, c, b, b, c*], la suma de las distancias euclídeas para cada ítem serán [0, 1, 0, 1, 0] , donde los unos corresponden a los desacuerdos entre las categorías seleccionadas por ambos jueces para cada uno de los ítemes, y la estimación de δ_o –que llamamos d_o por tratarse de un estimador– será

igual a

$$d_o = (0+1+0+1+0)/(5) = 2/5 = .400$$

Tabla 5.1. Datos del Ejemplo 5.1

Ítemes (I)	*Jueces (J)*		
	1	2	3
1	*a*	*a*	*a*
2	*b*	*c*	*b*
3	*b*	*b*	*b*
4	*a*	*b*	*b*
5	*c*	*c*	*c*

De forma similar, Berry y Mielke (1988) y Janson y Olsson (2001) definieron también δ_e, como

$$\delta_e = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \Delta(y_{iA}, y_{i'B}) \qquad \text{(Ec. 5.4)}$$

donde $\Delta(y_{iA}, y_{i'B}) = \sum_{i,i'} (y_{iA} - y_{i'B})^2$ es el desacuerdo producido en la selección de las categorías de respuesta entre la respuesta al ítem *i* (para el juez *A*) y la respuesta a todos y cada uno de los ítemes (para el juez *B*), que se ha representado por *i'*. En este sentido, δ_e es la media –para el conjunto de los ítemes– de las distancias euclídeas cuadráticas entre la selección de las categorías realizada por ambos jueces para ítemes iguales y diferentes. Así,

para el primer ítem la suma de las distancias entre la respuesta del juez *A* y las respuestas al conjunto de los ítemes del juez *B* es igual a ($0 + 1^2 + 1^2 + 1^2 + 1^2 = 4$), y para los ítemes 2 a 5 tales sumas son respectivamente 3, 3, 4 y 3. Nótese que el número de comparaciones entre ítemes es en este caso el cuadrado del número de ítemes (5×5) en comparación con la estimación anterior de δ_o.

Igualmente, la estimación de δ_e para el conjunto de los ítemes y ambos observadores resulta igual a $d_e=(4+3+3+4+3)/(5)^2=17/25=.680$. En consecuencia, aplicando la Ecuación 5.1, la estimación del coeficiente *iota* será igual a

$$\hat{\iota}=1-\frac{.400}{.680}=.412$$

La ventaja fundamental de esta nueva formulación es su interpretación como una razón de medidas de distancia (o desacuerdo), en la que la distancia se mide simplemente contando valores que consisten en ceros y unos. De esta forma, el clásico coeficiente *kappa* se convierte en una medida de acuerdo basada en la proximidad de las clasificaciones, y se mide a través de la distancia euclídea cuadrática entre las clasificaciones de ambos jueces. Desde este punto de vista podría ser perfectamente generalizable –como se tratará posteriormente –desde la medida nominal a la escala ordinal y de intervalo.

5.3. Generalización a *kappa* para más de dos observadores

Sean ahora y_{ij} e $y_{ij'}$ las categorías seleccionadas por 2 jueces cualesquiera –identificados como *j* y *j'*– de un conjunto de *J* jueces. Berry y Mielke (1988) y Janson y Olsson (2001) definieron δ_o para más 2 evaluadores en los términos siguientes:

$$\delta_o = \frac{1}{N\binom{J}{2}} \sum_{i=1}^{N} \sum_{i<i'} \Delta(y_{ij}, y_{ij'}) \qquad \text{(Ec. 5.5)}$$

donde $\Delta(y_{ij}, y_{ij'}) = \sum_{k=1}^{K} (y_{ij} - y_{ij'})^2$ representa el desacuerdo producido en la selección de las categorías para el ítem *i* entre los 2 evaluadores. Por consiguiente, se trata de una generalización a más de 2 evaluadores del caso tratado anteriormente, donde δ_o es además la media –para el conjunto de los ítemes– de las distancias euclídeas cuadráticas entre la selección de las categorías realizadas por 2 observadores de cualesquiera de los mismos sujetos/ítemes, selección que puede ser igual a 0 $(si\ y_{ij} = y_{ij'})$ o igual a 1 $(si\ y_{ij} \neq y_{ij'})$.

En el Ejemplo 5.1 presentado en la Tabla 5.1 se calculan las distancias euclídeas entre los jueces *A* y *B*, cuya suma es 2. Procediendo con los jueces *A* y *C* las distancias euclídeas cuadráticas son [0, 0, 0, 1^2,0], cuya suma es 1; del mismo modo, las distancias euclídeas cuadráticas entre los jueces *B* y *C* son

[0,1^2 ,0 ,0 ,0], cuya suma da –igualmente– como resultado 1. Por lo tanto, la estimación de δ_o será entonces igual al promedio

$$d_o = (2+1+1)/(5)(3) = 4/15 = .267.$$

De forma similar, Berry y Mielke (1988) y Janson y Olsson (2001) definieron δ_e como

$$\delta_e = \frac{1}{N^2\binom{J}{2}} \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{j<j'} \Delta(y_{ij}, y_{i'j'}) \qquad \text{(Ec. 5.6)}$$

donde $\Delta(y_{ij}, y_{i'j'}) = \sum_{k=1}^{K} (y_{ij} - y_{i'j'})^2$ denota el desacuerdo producido en la selección de las categorías de respuesta entre la respuesta al ítem *i* (para el juez *j*) y la respuesta a cada uno de los *i′* ítemes (para el juez *j′*). En este sentido, δ_e es la media –para el conjunto de los ítemes y combinaciones de pares de jueces– de las distancias euclídeas cuadráticas entre la selección de las categorías realizada por los jueces para ítemes iguales y diferentes. En consecuencia, la suma de las distancias cuadráticas entre el primer ítem del juez *A* y el conjunto de los ítemes del juez *B* es, como se mostró antes, (4 + 3 + 3 + 4 + 3 = 20); para las restantes combinaciones entre pares de observadores son (4 + 2 + 2 + 4 + 4 = 16), para los observadores *A* y *C*, y (4 + 2 + 2 +2 + 4 = 14) para los observadores *B* y *C*. Obsérvese que el número de comparaciones

entre ítemes es el cuadrado del número de ítemes –aquí 5×5– en comparación con la estimación anterior. Así, el promedio –o estimación de δ_e – del conjunto de los ítemes resulta igual a

$$d_e=(17+16+16)/(5)^2(3)=49/75=.653$$

Por lo tanto, aplicando la Ecuación 5.1, la estimación del coeficiente *iota* será

$$\hat{\iota}=1-\frac{.267}{.653}=.591$$

Sin embargo, el procedimiento de cálculo utilizado hasta este capítulo solamente es practicable cuando hay pocos evaluadores y un reducido número de ítemes, pero resulta considerablemente más complejo cuando se incrementa el número de jueces y/o el número de ítemes.

Hay dos procedimientos de estimación del coeficiente *iota* para cualquier número de observadores y de objetos que resultan de interés intrínseco, puesto que representan una generalización de los procedimientos de estimación tratados en el capítulo anterior. El primero es un procedimiento **directo** basado en técnicas de permutación (Mielke e Iver, 1982; Berry y Mielke,1988) equivalente al método utilizado en el ejemplo ficticio empleado hasta ahora en el capítulo; el segundo es un procedimiento **indirecto**, mediante un análisis de varianza en dos sentidos, aprovechando la equivalencia entre la medida

euclídea y el ANOVA tratado en capítulos anteriores.

Para ilustrar el cálculo del coeficiente *iota* con los dos procedimientos –directo e indirecto– señalados anteriormente utilizaremos uno de los ejemplos desarrollados en el capítulo anterior, en concreto el Ejemplo 4.1, donde un conjunto de $N = 10$ ítemes, es clasificado por un conjunto de $J = 4$ jueces en una de $K = 3$ categorías de respuesta, en concreto, *a*, *b,* ó *c*.

La Tabla 5.2 resume nuevamente los datos empíricos y los resultados esenciales utilizados durante el proceso de cálculo.

Tabla 5.2. Datos empíricos del Ejemplo 4.1

Ítemes (I)	*Observadores (J)*				*Número patrones*	*Categorías (K)*			$\sum_{k=1}^{K} n_{ik}^2$
	A	*B*	*C*	*D*		*a*	*b*	*c*	
1	*a*	*a*	*a*	*c*	1	3	0	1	***10***
2	*a*	*a*	*b*	*c*	1	2	1	1	***6***
3	*a*	*a*	*b*	*c*	1	2	1	1	***6***
4	*a*	*a*	*c*	*c*	1	2	0	2	***8***
5	*a*	*b*	*a*	*a*	1	3	1	0	***10***
6	*b*	*a*	*a*	*a*	1	3	1	0	***10***
7	*b*	*b*	*b*	*b*	1	0	4	0	***16***
8	*b*	*c*	*b*	*b*	1	0	3	1	***10***
9	*c*	*c*	*b*	*b*	1	0	2	2	***8***
10	*c*	*c*	*c*	*c*	1	0	0	4	***16***
Categorías (K)	$n_{jk}(p_{jk})$				$n_{+k}(p_k)$				**100**
a	5	5	3	2	15	***225***			
b	3	2	5	3	13		***169***		
c	2	3	2	5	12			***144***	
$\sum_{k=1}^{K} n_{jk}^2$	***38***	***38***	***38***	***38***	**152**				**538**

5.3.1. Procedimiento directo

Para obtener una estimación de δ_o, Berry y Mielke (1988:924) aplican la Ecuación 5.3 en la primera parte de la Tabla 5.2 sumando –para cada par de jueces– los desacuerdos o no coincidencias –que se puntúan con 1, mientras

que los acuerdos se puntúan con 0– en el conjunto de los ítemes. Por tanto, las no coincidencias entre los jueces *A* y *B* son, para el conjunto de los 10 ítemes del Ejemplo 4.1,

$$0 + 0 + 0 + 0 + 1 + 1 + 0 + 1 + 0 + 0$$

cuya suma es 3. Generalizando para el conjunto de las 6 combinaciones posibles entre pares de evaluadores, el resultado final es, aplicando la Ecuación 5.5,

$$d_0 = \frac{3+5+6+6+7+3}{(10)(6)} = .500$$

De forma similar, para obtener una estimación de δ_e se utiliza una simplificación de la Ecuación 5.4 con la segunda parte de la Tabla 5.2 consistente en sumar –para cada par de jueces– los productos de todos los desacuerdos posibles entre categorías, en el conjunto de las categorías (es decir, las columnas de n_{jk}). De modo que la suma de los productos para los jueces *A* y *B* –columnas 1 y 2– resulta igual a

$$5\times 2+5\times 3+3\times 3+5\times 3+5\times 2+2\times 2=63$$

y la de los jueces *A* y *C* –columnas 1 y 3– es

$$5\times5+5\times2+3\times2+3\times3+3\times2+5\times2=66$$

Generalizando para el conjunto de las 6 combinaciones posibles entre pares de observadores, se obtiene como resultado

$$d_e=\frac{63+66+71+69+69+69}{(10)^2(6)}=.678$$

Finalmente, el coeficiente *iota* se estima aplicando la Ecuación 5.1:

$$\hat{\iota}=1-\frac{d_o}{d_e}=1-\frac{.500}{.678}=.263$$

Este resultado coincide exactamente con el valor del coeficiente *kappa* generalizado (Conger, 1980) examinado en el capítulo previo.

5.3.2. Procedimiento indirecto vía ANOVA

Una alternativa computacional indirecta, aunque generalmente más efectiva, se basa en la equivalencia entre la distancia euclídea cuadrática y el análisis de varianza en dos sentidos que se desarrolló en capítulos anteriores. Este procedimiento implica realizar un ANOVA en dos sentidos convencional, y estimar posteriormente las magnitudes δ_o y δ_e mediante las ecuaciones

$$d_o = \frac{J\,SC_E}{N\binom{J}{2}} \qquad \text{(Ec. 5.7)}$$

$$d_e = \frac{(J-1)\,SC_T + SC_J}{N\binom{J}{2}} \qquad \text{(Ec. 5.8)}$$

Utilizamos para ello la tabla ANOVA correspondiente, asumiendo el efecto de los observadores fijo, como se resume en la Tabla 5.3.

Tabla 5.3. ANOVA dos sentidos del Ejemplo 4.1 con efecto Jueces fijo

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	5.775	9	.642	$\hat{\sigma}_I^2 = (.642 - .245)/4 = .099$
Error (SC_E)	7.500	30	.250	
- Jueces (SC_J)	.875	3	.292	$\hat{\sigma}_J'^2 = .292/9 = .032$
- Residual (SC_R)	6.625	27	.245	$\hat{\sigma}_R^2 = .245$
Total (SC_T)	13.275	39		

Conviene observar la atípica estimación del componente de varianza para jueces, cuyos detalles se explican más adelante.

Aplicando las Ecuaciones 5.5 y 5.6 a los resultados de la Tabla 5.3 obtenemos

$$d_o = \frac{(4)(7.5)}{(10)(6)} = .500$$

$$d_e = \frac{(3)(13.275)+0.875}{(10)(6)} = .678$$

o, de forma equivalente,

$$\tilde{\rho}_I = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_J'^2 + \sigma_R^2} = \frac{.099}{.099 + .032 + .245} = .263$$

obteniéndose el mismo resultado para el coeficiente *iota:*

$$\hat{\iota} = 1 - \frac{d_o}{d_e} = 1 - \frac{.500}{.678} = .263$$

Este último coeficiente varía en magnitud del mismo modo que varían *kappa* y –como se verá posteriormente– el coeficiente de correlación intraclase. El límite superior es 1 (obtenido cuando $d_o = 0$), indicando que el

acuerdo es perfecto; el límite inferior es $-1/(J-1)$. Los valores de *iota* pueden ser interpretados asimismo de forma categórica, por ejemplo mediante el esquema de valoración propuesto por Landis y Koch (1977).

Berry y Mielke (1988:929) recomiendan utilizar para la inferencia estadística una prueba *T* definida como

$$T = \frac{\delta_o - \delta_e}{\sigma_\delta} \qquad \text{(Ec. 5.9)}$$

donde la media y varianza de *T* son 0 y 1 respectivamente, y su desviación típica σ_δ puede obtenerse mediante *jackknife* o *bootstrap*. La distribución de *T* se aproxima a la distribución normal conforme el número de ítemes, objetos o sujetos (*N*) tiende a infinito.

5.4. Generalización a medidas cuantitativas

En el caso de medidas cuantitativas, la aplicación de las Ecuaciones 5.4 y 5.5 es más simple que en el caso de medidas categóricas, dado que no existe necesidad de transformar en una puntuación 0-1 el acuerdo o desacuerdo en la selección de las categorías por parte de los jueces, sino simplemente obtener la medida directa de la distancia.

Para ilustrar el cálculo del coeficiente *iota* con datos cuantitativos

utilizaremos un ejemplo ficticio –que llamaremos Ejemplo 5.2– con N = 5 ítemes y J = 3 observadores que evalúan actitudes sobre una escala cuantitativa.

La Tabla 5.4 presenta los datos empíricos y los resultados esenciales del proceso de cálculo con el procedimiento directo.

Tabla 5.4. Datos empíricos del Ejemplo 5.2 con procedimiento directo

Ítemes (*I*)	*Jueces (J)*			d_o			Suma
	Juez A	*Juez B*	*Juez C*	*A-B*	*A-C*	*B-C*	
1	*1*	*1*	*1*	0	0	0	0
2	*3*	*2*	*2*	1	1	0	2
3	*5*	*4*	*5*	1	0	1	2
4	*4*	*4*	*4*	0	0	0	0
5	*2*	*2*	*3*	0	1	1	2
Suma	15	13	15	2	2	2	**6**
d_e				90	100	90	**280**

Con J = 3 jueces, para obtener un estimador de δ_o se calculan las diferencias cuadráticas ítem a ítem entre los tres pares de jueces posibles. Así, las diferencias entre los jueces *A* y *B* para los 5 ítemes son (0, 1^2, 1^2, 0, 0), cuya suma es 2. La misma estructura se observa también con las diferencias entre los jueces *A* y *C,* que para el conjunto de los ítemes es (0, 1^2, 0, 0, -1^2), cuya suma es 2, y entre los jueces *B* y *C*, que es (0, 0, -1^2, 0, -1^2), cuya suma es igualmente 2. En consecuencia, aplicando la Ecuación 5.5, obtenemos

$$d_0 = (2+2+2)/(5)(3) = 6/15 = .400$$

De forma similar, para obtener un estimador de δ_e, se calculan para cada par de jueces las diferencias cuadráticas entre cada ítem de un juez y el conjunto de los ítemes del otro juez. De modo que, las diferencias entre los jueces *A* y *C*, para el ítem 1 del juez *A* son (0, 1^2, 4^2, 3^2, 2^2) , cuya suma es 30; para el ítem 2 del juez *A* son (2^2, 1^2, 2^2, 1^2, 0), cuya suma es 10; para el ítem 3 del juez *A* son (4^2, 3^2, 0, 1^2, 2^2), cuya suma es 30; para el ítem 4 del juez *A* son (3^2, 2^2, 1^2, 0, 1^2), cuya suma es 15 y para el ítem 5 del juez *A* son (1^2, 0, 3^2, 2^2, 1^2), cuya suma es 15. Por consiguiente, las suma de las diferencias cuadráticas euclídeas para los jueces *A* y *C* es 100. El mismo proceso se aplica de manera análoga a las diferencias entre los jueces *A* y *B*, cuya suma es 90, y entre los jueces *B* y *C*, cuya suma es también 90. Por lo tanto, aplicando la Ecuación 5.5, obtenemos

$$d_e = (90+100+90)/(5^2)(3) = 280/75 = 3.733$$

Finalmente, aplicando la Ecuación 5.1 para estimar el coeficiente *iota*, obtenemos

$$\hat{\iota} = 1 - \frac{.400}{3.733} = .893$$

cuya interpretación es ya conocida.

A continuación, la Tabla 5.5 presenta el análisis de varianza de los datos del Ejemplo 5.2 como resumen del proceso de cálculo con el procedimiento indirecto.

Tabla 5.5. ANOVA dos sentidos del Ejemplo 5.2 con efecto Jueces fijo

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	25.733	4	6.433	
Error (SC_E)	2.000	10	.200	
- Jueces (SC_J)	.533	2	.267	$\hat{\sigma}'^2_J = .267/4 = .067$
- Residual (SC_R)	1.467	8	.183	$\hat{\sigma}^2_R = .183$
Total (SC_T)	27.733	14		

Nota: La estimación de es atípica (vid. Cap.2)

Aplicando las Ecuaciones 5.5 y 5.6 para calcular los estimadores de δ_o y de δ_e obtenemos

$$d_o = \frac{J\,SC_E}{N\binom{J}{2}} = \frac{(3)(2.000)}{(5)(3)} = .400$$

$$d_e = \frac{(J-1)SC_T + SC_R}{N\binom{J}{2}} = \frac{(2)(25.733)+.533}{(5)(3)} = 3.733$$

y finalmente el coeficiente *iota* resulta

$$\hat{\iota} = 1 - \frac{.400}{3.733} = .893$$

Janson y Olsson (2001) han demostrado que el coeficiente *iota* es equivalente al coeficiente de correlación intraclase, en concreto es equivalente al Modelo IIIA, formulado por McGraw y Wong (1996), que asume la interacción ausente –como se demostró en el Capítulo 1– y es incluso equivalente al coeficiente de correlación de contingencia, ya que como se expuso asimismo en el Capítulo 1, el coeficiente de correlación de contingencia puede estimarse a partir del Modelo IIIA del coeficiente de correlación intraclase (Carrasco y Jover, 2003) . Utilizando los componentes de la varianza de la Tabla 5.5 –que asume fijo el factor jueces– y aplicando la Ecuación 1.12 obtenemos

$$ICC = CCC = \frac{\hat{\sigma}_I^2}{\hat{\sigma}_I^2 + \hat{\sigma}_J'^2 + \hat{\sigma}_R^2} = \frac{2.083}{2.083 + .067 + .183} = .893$$

donde $\hat{\sigma}_I^2$ es la estimación del componente de varianza para ítemes y $\hat{\sigma}_J'^2$ es la estimación del mismo componente para evaluadores, que como se trató en el Capítulo 1 se asume fijo pero se calcula de forma atípica, como si se tratara de un efecto aleatorio, prescindiendo de la *MC* del denominador en el cálculo del componente; si el tamaño muestral es pequeño –i.e. $N < 100$– suelen tomarse los grados de libertad –en lugar de los niveles del factor ítemes– como divisor

de la razón, pero si es grande se utiliza –como es usual– el número de niveles del factor ítemes (véase Capítulo 1). Finalmente, $\hat{\sigma}_e^2$ es la varianza de error del modelo.

5.5. Generalización al caso multivariante

Una de las más interesantes propiedades del coeficiente *iota*, considerada como propiedad deseable en el trabajo de Berry y Mielke (1988) en contraste con otros coeficientes de acuerdo de uso común, reside en la posibilidad de ser utilizado para valorar el acuerdo en el caso multivariante, ya que permite operar con dos –o más– características o comportamientos a un tiempo bajo la forma de dos –o más– conjuntos de ítemes, que supuestamente representan un constructo multivariante. Siendo M el número de características o elementos del constructo cuyo grado de acuerdo se pretende valorar, para $m = 1,\ldots, M$, una sencilla modificación de las Ecuaciones 5.5 y 5.6 conduce a las ecuaciones

$$\delta_o = \frac{1}{N\binom{J}{2}} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j<j'} \Delta(y_{iA}, y_{iB}) \qquad \text{(Ec. 5.10)}$$

$$\delta_e = \frac{1}{N^2\binom{J}{2}} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{j<j'} \Delta(y_{ij}, y_{i'j'}) \qquad \text{(Ec. 5.11)}$$

donde la diferencia afecta únicamente al primer sumatorio y queda incorporada la propiedad multivariante del acuerdo. De forma similar, empleando el procedimiento ANOVA, una pequeña modificación de las Ecuaciones 5.7 y 5.8

$$d_0 = \frac{J \sum_{m=1}^{M} w_m SC_{E_m}}{N \binom{J}{2}} \qquad \text{(Ec. 5.12)}$$

$$d_e = \frac{(J-1) \sum_{m=1}^{M} w_m SC_{T_m} + \sum_{m=1}^{M} w_m SC_{J_m}}{N \binom{J}{2}} \qquad \text{(Ec. 5.13)}$$

donde w_m es el peso utilizado con las sumas cuadráticas, permite igualmente valorar el grado de acuerdo de un constructo multivariante. Adviértase que se trata de una suma ponderada de términos, cada uno de los cuales se toma del correspondiente ANOVA realizado con los datos de cada elemento del constructo multivariante.

Para ilustrar esta propiedad del coeficiente *iota*, utilizaremos una ampliación del Ejemplo 5.2 –que llamaremos Ejemplo 5.3– con dos conjuntos de ítemes, que muestra la Tabla 5.6.

Tabla 5.6. Ejemplo 5.3 (ficticio)

Jueces	*1*		*2*		*3*	
Variables	*A*	*B*	*A*	*B*	*A*	*B*
Ítem 1	1	2	1	3	1	2
Ítem 2	3	5	2	5	2	5
Ítem 3	5	3	4	3	5	3
Ítem 4	4	2	4	3	4	3
Ítem 5	2	2	2	2	3	2

Utilizando el procedimiento ANOVA, la valoración del acuerdo multivariante pasa por realizar un primer ANOVA con los datos del elemento *A* del constructo que ya se realizó anteriormente (véase Tabla 5.5), obteniendo un coeficiente de $\hat{\iota}_A = .893$, y un segundo ANOVA con los datos del elemento *B*, cuyo análisis de varianza se resume en la Tabla 5.7.

Tabla 5.7. ANOVA en dos sentidos del elemento B del Ejemplo 5.3

Fuentes	*SC*	*gl*	*MC*	*Componentes de la varianza*
Ítemes (SC_I)	16.667	4	4.167	
Error (SC_E)	1.333	10	.133	
- Jueces (SC_J)	.400	2	.200	$\hat{\sigma}'^2_J = .200/4 = .050$
- Residual (SC_R)	.933	8	.117	$\hat{\sigma}^2_R = .117$
Total (SC_T)	18.000	14		

La estimación del grado de acuerdo para el elemento *B* del constructo multivariante, siendo $d_o = (3)(1.333)/15 = .267$ y $d_e = [(2)(18) + .400]/15 = 2.427$, es $\hat{\iota}_B = 1 - (.267/2.427) = .890$. Aplicando las Ecuaciones 5.12 y 5.13 para

estimar el coeficiente multivariante, se obtiene

$$d_o = \frac{(3)(2+1.333)}{(5)(3)} = .667$$

$$d_e = \frac{(2)(27.333+18)+(.533+.400)}{(5)(3)} = 6.107$$

y el coeficiente *iota* resulta igual a

$$\hat{\iota}_{AB} = 1 - \frac{.667}{6.107} = .891$$

un valor muy similar al promedio de los dos elementos.

Obviamente, no se ha utilizado ningún tipo de ponderación con los datos de los dos elementos del constructo multivariante, puesto que ambos utilizan una escala de medida similar. En ocasiones en que las escalas de medida son diferentes puede emplearse alguna forma de estandarización y/o ponderación con el objeto de homogeneizar las varianzas de los diferentes elementos del constructo, para así interpretar apropiadamente el índice de acuerdo multivariante hallado (véase Janson y Olsson, 2001).

Capítulo 6

Modelos loglineales

6.1. Introducción

Los coeficientes de acuerdo tratados en los capítulos anteriores, y en particular el coeficiente *kappa* de Cohen, gozan hoy por hoy de una gran popularidad. Varias razones podrían explicar este hecho (von Eye y Mun, 2005:31):

- en primer lugar, todos los coeficientes de acuerdo son aplicables en la práctica, con excepción de aquellos casos en los que algunas de las categorías raramente son seleccionadas por los jueces,

- en segundo lugar, todos los coeficientes de acuerdo condensan información que es relativamente fácil de interpretar;

- en tercer lugar, todos los coeficientes de acuerdo son en la práctica sencillos de calcular y algunos de ellos –en particular, el coeficiente *kappa*, tanto nominal como ordinal– se han incorporado a los grandes paquetes estadísticos (e.g. SAS, SPSS, SYSTAT, etc.), permitiendo la aplicación de un amplio repertorio de técnicas de inferencia estadística (como pruebas de significación e intervalos de confianza).

Pero, pese a su popularidad, los índices descriptivos para evaluar el acuerdo entre jueces han generado multitud de críticas en su aplicación práctica. La mayoría de ellas se han centrado en el coeficiente *kappa* de Cohen –y en sus correspondientes generalizaciones– y se han tratado con detalle en el Capítulo 2; presentan problemas de sesgo –de los jueces– y de prevalencia –de ciertas categorías de respuesta– . Además de estos reparos, es preciso apuntar algunas críticas generales que motivan la elaboración de este capítulo, a saber:

- en primer lugar, la impresión general de pérdida de información que supone resumir todas las características de una tabla de contingencia –o tabla de acuerdo– en un único índice descriptivo;

- en segundo lugar, la dificultad de comparar coeficientes de acuerdo procedentes de diferentes tablas de contingencia, aunque este tema ha sido en parte tratado por Barnhart y Williamson, (2002);

- y finalmente, excepto en casos muy excepcionales, los coeficientes descriptivos no permiten describir la estructura de la distribución conjunta de las decisiones de los jueces y, lo que es más problemático, no es posible conocer la bondad del ajuste de la estimación de un coeficiente de acuerdo respecto a la tabla de acuerdo que sirvió de base para su cálculo.

En el presente capítulo se presentan los detalles del **enfoque del modelado estadístico** para evaluar el acuerdo entre observadores, que surgió en la década de los 80 en respuesta a la interpretación simplista de los coeficientes de acuerdo hasta entonces dominante en la investigación aplicada (véase Tanner y Young, 1985a,b; Agresti, 1988, 1992). Varios son los modelos estadísticos que pueden aplicarse en este contexto, pero nos centraremos en los dos cuya investigación ha despertado mayor interés; en concreto, los **modelos loglineales** (en este capítulo) y los **modelos mixtura** (en el capítulo siguiente).

6.2. Modelos loglineales

El concepto de análisis loglineal en tablas de acuerdo es análogo al concepto de ANOVA para variables de respuesta de naturaleza cuantitativa que asumen una distribución normal. Cuando la variable de respuesta no tiene naturaleza numérica sino categórica, y más concretamente en el caso de los recuentos de una tabla de contingencia particular –la tabla de acuerdo– se asume una distribución de Poisson (Clogg y Shihadeh, 1994; Vermunt, 1996).

Los modelos loglineales modelan el acuerdo observado en términos de componentes y constituyen una jerarquía que varía desde modelos muy simples, que dejan muchos grados de libertad residuales y por tanto utilizan pocos parámetros, a modelos muy complejos, que consumen muchos grados de libertad a costa de introducir un considerable número de parámetros.

Dos de las ventajas cruciales de los modelos loglineales, respecto de los coeficientes descriptivos, residen en que los parámetros que determinan tales modelos:

- pueden ser definidos mediante el **enfoque de la matriz de diseño** derivado del análisis con modelos lineales y no lineales (véase Evers y Namboodiri, 1978; Kirk, 1995; Ato y López, 1996; Christensen, 1997; von Eye y Mun, 2005) ,y además

- pueden ser sometidos a prueba empírica mediante el **enfoque de la comparación de modelos** (véase Ato y Vallejo, 2007), verificando si el

ajuste producido en un modelo aumentado –o lo que es lo mismo, con la inclusión de uno o más grados de libertad adicionales– mejora respecto del ajuste de un modelo restringido de referencia –es decir, sin la inclusión de grados de libertad adicionales– .

En lo que sigue utilizaremos ambos enfoques para exponer las líneas maestras de la aplicación de los modelos loglineales a la evaluación del acuerdo entre jueces.

Todos los modelos -sean lineales o loglineales- se agrupán en familias cuyos componentes pueden identificarse por la estructura que la matriz de diseño presenta. En la práctica, hay dos familias de modelos que permiten mayor aplicabilidad en el área del acuerdo entre jueces: los modelos de **cuasi-independencia** y los de modelos de **cuasi-simetría**. Este trabajo se centra en la primera de las familias.

6.3. Ejemplo 6.1: Los datos de Dillon y Mullani (1984)

Para ilustrar la aplicación de los modelos loglineales en la evaluación del acuerdo entre jueces nos valdremos del que denominaremos Ejemplo 6.1 tomado de un trabajo de Dillon y Mullani (1984), en el que $J = 2$ jueces, A y B registraron un conjunto de $N = 164$ respuestas cognitivas elicitadas en un estudio de comunicación persuasiva sobre una escala con $K = 3$ categorías de respuesta (“positiva”, “neutral” y “negativa”). Los datos empíricos se

representan en la Tabla 6.1.

En una tabla de acuerdo como la Tabla 6.1, las frecuencias conjuntas observadas para la categoría elegida por el juez *A* –para $k = 1, ..., K$– y la categoría elegida por el juez *B* –para $k' = 1, ..., K$– se denotan por $n_{kk'}$, las frecuencias marginales del juez *A* por $n_{k+} = n_k^A$ y las frecuencias marginales del juez *B* por $n_{+k'} = n_{k'}^B$. El número total de ítemes clasificados por los observadores es, por tanto, $\sum_{k=1}^{K} \sum_{k'=1}^{K} n_{kk'} = N$. Para los datos del Ejemplo 6.1, los coeficientes descriptivos –y sus respectivos errores típicos– obtenidos por *jackknife* son: $\hat{\kappa} = .565(.053)$, $\hat{\pi} = .557(.056)$, $\hat{\gamma} = .555(.061)(.053)$ y $\hat{\sigma} = .579(.053)$.

Tabla 6.1. Frecuencias y sus probabilidades del Ejemplo 6.1

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61** (.372)	**26**(.159)	**5**(.030)	92 (.561)
neutral	**4** (.025)	**26**(.159)	**3**(.018)	33 (.201)
negativa	**1** (.006)	**7**(.043)	**31** (.189)	39(.238)
Marginales	66(.402)	59(.360)	39(.238)	164(1.000)

6.3.1. El modelo de independencia (Modelo I)

En toda familia de modelos hay uno básico que da nombre a la familia. En el caso de la familia de cuasi-independencia, el modelo básico es obviamente el **modelo de cuasi-independencia**. No obstante en todas las familias suele existir un modelo de referencia, que es el último recurso al que se acoge el investigador cuando todos los modelos de la familia se ajustan perfectamente. Para las dos familias citadas anteriormente (cuasi-independencia y cuasi-simetría), el modelo de referencia es el **modelo de_independencia** (Ato y López, 1996).

Asumiendo que se utiliza **codificación ficticia** para convertir las categorías del factor en vectores numéricos (véase Ato y López, 1996), el modelo de independencia se representa por I **(Modelo I)** y se define para cualquier categoría de *k* de *A* y *k'* de *B*, mediante

$$\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B \qquad \text{(Ec. 6.1)}$$

donde $m_{kk'}$ son las frecuencias esperadas o ajustadas por el modelo, λ representa el valor –en escala logarítmica– de la categoría de referencia y λ_k^A y $\lambda_{k'}^B$ son los efectos –en escala logarítmica– correspondientes al juez A y al juez *B*, respectivamente. El modelo de la Ecuación 6.1 puede ser utilizado para obtener estimaciones por máxima verosimilitud de tales efectos.

La medida de la discrepancia entre valores empíricos observados y valores

esperados por el modelo se realiza utilizando varios indicadores, de los cuales el más importante es la desvianza -*deviance*-, (L^2) que representa la desviación respecto del modelo perfecto o saturado; es decir, del modelo donde valores observados y valores esperados son coincidentes y no hay grados de libertad residuales. La desvianza sólo puede valorarse en función del número de grados de libertad residuales. En general, el modelo se considera aceptable cuando la desvianza es inferior al número de grados de libertad residuales. Puesto que la desvianza se distribuye aproximadamente según χ^2_{gl}, la probabilidad asociada a la magnitud de la desvianza en la distribución χ^2_{gl} puede servir para valorar el ajuste. Se considera por lo general aceptable el ajuste cuando el valor de probabilidad es igual o mayor de 0.10.

La Tabla 6.2 resume los resultados del ajuste del Modelo I en tres secciones consecutivas, la primera para las frecuencias esperadas (y sus correspondientes probabilidades entre paréntesis), la segunda para los estimadores de los parámetros (con sus probabilidades entre paréntesis) y la tercera para el ajuste del modelo.

Tabla 6.2. Resumen del ajuste del Modelo I

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**37.024** (.226)	**33.098** (.202)	**21.878** (.133)	92.000 (.561)
neutral	**13.280** (.081)	**11.872** (.072)	**7.848** (.048)	33.000 (.201)
negativa	**15.695** (.096)	**14.030** (.086)	**9.274** (.057)	39.000 (.238)
Marginales	66.000 (.402)	59.000 (.360)	39.000 (.238)	164.000 (1.000)

Parámetros	*positiva*	*neutral*	*negativa*	
λ				*3.612 (37.024)*
λ_k^A	*.000 (1.000)*	*-1.025 (.359)*	*-.858 (.424)*	
$\lambda_{k'}^B$	*.000 (1.000)*	*-.112 (.894)*	*-.526 (.591)*	

Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	118.600	4	.000	98.200

Algunos comentarios para la tabla 6.2 permitirán aclarar algunos de los términos utilizados.

- En la primera sección se presentan los valores esperados –o ajustados– y sus correspondientes probabilidades condicionales para cada una de las casillas de la tabla de acuerdo. Nótese que los valores esperados o ajustados son considerablemente diferentes de los valores observados (Tabla 6.1), pero las sumas marginales son exactamente coincidentes en ambas tablas.

- En la segunda sección se presentan los estimadores de los parámetros

en escala loglineal, y entre paréntesis en escala exponencial, para cuya interpretación se asume que se ha empleado codificación ficticia. Así, $\lambda = 3.612$ representa en unidades logarítmicas el valor esperado de la casilla de referencia, a saber, la casilla <11>, y por tanto su exponencial resulta igual a $e^{3.162} \approx 37.024$. Del mismo modo, $\lambda_1^A = -1.025$ representa la disminución, en unidades logarítmicas, en la respuesta del evaluador *A* a la segunda categoría respecto de su respuesta a la primera y por tanto su exponencial

- es $e^{-1.025} \approx .359$ (puede comprobarse que, en efecto, la reducción en la frecuencia marginal del juez *A* a la categoría 2 respecto de la categoría 1 es $(92)(.359) \approx 33)$. Y, análogamente, $\lambda_2^A = -.858$ representa la disminución, en unidades logarítmicas, en la respuesta del juez *A* a la tercera categoría respecto de su respuesta a la primera y de ahí su exponencial $e^{-.858} \approx .424$ (igualmente puede comprobarse que la reducción en la frecuencia marginal del juez *A* a la categoría 3 respecto de la categoría 1 es $(92)(.424) \approx 39)$. De modo similar se interpretan los parámetros para el observador *B*.

- Finalmente, en la tercera sección de la Tabla 6.2 se muestran los resultados del ajuste del modelo, cuya desvianza –con 4 grados de libertad residuales–, produce un valor de probabilidad prácticamente nulo, lo que conduce a la conclusión de que el modelo de independencia no es un modelo apropiado para explicar los datos empíricos de la Tabla 6.1.

La razón que justifica el inapropiado ajuste del Modelo QI hay que buscarla en el hecho de que los evaluadores *A* y *B* no clasifican de forma independiente las categorías de la tabla de contingencia. Los datos apuntan a que puede existir alguna destreza común –asociación– entre ambos jueces que explique el alto grado de coincidencia o acuerdo en la clasificación de las categorías.

6.3.2. El modelo de cuasi-independencia (Modelo QI)

A diferencia de la forma estándar de proceder con los modelos loglineales, el modelo de cuasi-independencia (**Modelo QI**) es el más complejo, cuya naturaleza da nombre al conjunto de modelos que constituye la familia. Se basa en el modelo de independencia, pero además de los parámetros del Modelo I este modelo incluye un conjunto de parámetros ideados para explicar la coincidencia o acuerdo en la clasificación de las categorías por ambos jueces, es decir, las casillas diagonales de la tabla de acuerdo. Para el resto de las casillas, el Modelo QI se comporta como un modelo de independencia.

El Modelo QI se formula para cualquier categoría *k* de *A* y *k'* de *B*, como

$$\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \delta_k \qquad \text{(Ec. 6.2)}$$

donde δ_k son **parámetros diagonales**, puesto que afectan a las casillas

diagonales (en particular, a las casillas donde se concentra el acuerdo entre los jueces) de la tabla de acuerdo. La característica esencial de este modelo está en considerar que cada categoría tiene una ponderación diferente para la evaluación del acuerdo entre jueces, y por consiguiente se excluye de la tabla de acuerdo "fijando" los valores empíricos diagonales (Bergan, 1980). Eso conlleva que el Modelo QI puede considerarse como una medida del desacuerdo existente en la tabla (Bishop, Fienberg y Holland, 1975). En contraste, el Modelo I, que incluye tanto las casillas diagonales como las no diagonales, puede considerarse como una medida conjunta de acuerdo y desacuerdo.

La Tabla 6.3 resume los resultados del ajuste del Modelo QI.

Tabla 6.3. Resumen del ajuste del Modelo QI

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61.000** (.372)	**26.319** (.161)	**4.682** (.027)	92.000 (.561)
neutral	**3.681** (.022)	**26.000** (.159)	**3.318** (.020)	33.000 (.201)
negativa	**1.319** (.008)	**6.681** (.041)	**31.000** (.189)	39.000 (.238)
Marginales	66.000 (.402)	59.000 (.360)	39.000 (.238)	164.000 (1.000)
Parámetros	*positiva*	*neutral*	*negativa*	
λ				*1.647 (5.194)*
λ_k^A	*.000 (1.000)*	*- .344 (.709)*	*-1.371 (.709)*	
$\lambda_{k'}^B$	*.000 (1.000)*	*1.623 (5.067)*	*- .104 (.901)*	
δ_k	*2.463 (11.750)*	*0.332 (1.394)*	*3.261 (26.083)*	
Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	.182	1	.669	- 4.920

- En la primera sección de la Tabla 6.3 se exhiben los valores esperados. Adviértase que no hay ninguna discrepancia entre valores observados y valores esperados en las casillas diagonales (casillas <11>, <22> y <33>), ya que el modelo "fija" tales casillas con sus valores muestrales. El resto de las casillas presenta un ajuste casi perfecto entre valores observados y valores esperados; en concreto, los valores no diagonales pueden atribuirse al azar.

- En la segunda sección, se muestran los estimadores por máxima verosimilitud de este modelo, cuya principal novedad son los parámetros diagonales. La magnitud de los parámetros diagonales se asocia con la consistencia en la respuesta conjunta por parte de sendos jueces: a mayor magnitud, mayor consistencia.

- En la tercera sección se expone el ajuste de este modelo, que con un grado de libertad residual produce una desvianza de un valor de $L_1^2 = .182; P = .669$, lo que evidencia que se trata de un modelo óptimamente ajustado.

En un trabajo muy influyente, Guggenmoos-Holtzman (1989) sugirió estimar medidas de acuerdo basadas en modelos loglineales, que son similares en forma a los coeficientes de acuerdo de tipo descriptivo, utilizando los parámetros diagonales δ a través de la Ecuación 6.3, que es de naturaleza similar a la ecuación RCA tratada en el Capítulo 2 de este trabajo.

$$\psi = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\delta_k)} \right] \qquad \text{(Ec. 6.3)}$$

donde p_{kk}, son las probabilidades empíricas de las casillas diagonales de la tabla de acuerdo, y $\exp(\delta_k)$ son los exponenciales de los parámetros diagonales. Apréciese que esta formulación distingue dos partes bien delimitadas: el **acuerdo observado** (representado por las probabilidades p_{kk}) y el **acuerdo esperado por azar** (representado por las razones $p_{kk}/\exp(\delta_k)$, cuya magnitud produce de hecho un efecto corrector de mayor o menor grado en función del valor del parámetro diagonal δ_k).

De este modo, para el Modelo QI la medida de acuerdo derivada de la formulación propuesta por Guggenmoos-Holtzman (1989) resulta igual a

$$\hat{\psi}_{QI} = \left(.372 - \frac{.372}{11.750} \right) + \left(.159 - \frac{.159}{1.394} \right) + \left(.189 - \frac{.189}{26.080} \right) = .567,$$

un valor muy similar a los coeficientes descriptivos. Adviértase no obstante que esta medida de acuerdo utiliza una escala diferente, que varía de 0 –cuando la corrección debida al azar es igual a la probabilidad de acuerdo observada– a 1 –cuando no hay frecuencias en las casillas no diagonales, y por tanto no existe corrección debida al azar– en lugar de la escala $-1/1$ que emplean los coeficientes descriptivos. Por esta razón, ambas medidas de acuerdo no son estrictamente comparables.

Por otra parte, nótese que este modelo se acerca considerablemente al modelo saturado, aunque dejando al menos un grado de libertad residual. En general el Modelo QI se ajusta aceptablemente la mayoría de las ocasiones en que se aplica, particularmente con tablas de acuerdo pequeñas.

Una característica deseable de los modelos estadísticos es su parsimonia, la economía de sus parámetros. En consecuencia, los restantes modelos de la familia QI que se verán a continuación se proponen con el objetivo de simplificar el modelo de cuasi-independencia introduciendo algún tipo de restricción. La naturaleza de la restricción utilizada justifica en gran medida el nombre que adopta el modelo.

6.3.3. El modelo de cuasi-independencia constante (Modelo QIC)

Una forma de simplificar el modelo de cuasi-independencia radica en reducir el número de parámetros diagonales, que utilizan tantos grados de libertad como categorías tiene la variable a clasificar, a un único parámetro que sea constante para todas las casillas diagonales. El modelo resultante se designa de cuasi-independencia constante (**Modelo QIC**). Fue inicialmente propuesto por Tanner y Young (1985a), que lo denominaron *equal weight agreement model* y posteriormente ha sido tratado por Wickens (1989), Agresti (2002) y von Eye y Mun (2005). La característica esencial del Modelo QIC reside en que e**l acuerdo se produce con la misma consistencia –o constancia– en todas las categorías**, por lo que, cada casilla diagonal se considera igualmente

importante en la evaluación del acuerdo.

La ecuación general del modelo es para cualquier categoría k de A y k' de B

$$\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \delta \qquad \text{(Ec. 6.4)}$$

donde δ es un parámetro que refleja el mismo peso para todas las categorías diagonales. Adviértase que, a diferencia del Modelo QI, este parámetro es único y, por consiguiente supone una mayor reducción en el número de grados de libertad consumidos (por lo que aumentan los grados de libertad residuales).

La Tabla 6.4 presenta los resultados del ajuste del Modelo QIC.

- En primer lugar, los casillas diagonales ya no se "fijan", y por tanto los valores esperados no coinciden con los observados. Por lo demás, se observa cierta discrepancia entre valores observados y esperados, más patente en las casillas externas a la diagonal.

- En segundo lugar, el estimador del parámetro diagonal constante es $\delta = 1.978$ en unidades logarítmicas.

- En tercer lugar, el ajuste no resulta aceptable, puesto que el modelo con 3 grados de libertad residuales, $L_3^2 = 10.129, P = .017,$ lo que implica que los datos observados no son compatibles con el

supuesto del modelo de que ambos jueces no otorgan el mismo peso a todas las categorías.

Tabla 6.4. Resumen del ajuste del Modelo QIC

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61.072** (.372)	**21.212** (.129)	**9.715** (.059)	92.000 (.561)
neutral	**1.627** (.010)	**29.504** (.180)	**1.870** (.011)	33.000 (.201)
negativa	**3.301** (.020)	**8.284** (.051)	**27.415** (.167)	39.000 (.238)
Marginales	66.000 (.402)	59,000 (.360)	39.000 (.237)	164.000 (1.000)
Parámetros	*positiva*	*neutral*	*negativa*	
λ				*2.134 (8.452)*
λ_k^A	*.000 (1.000)*	*- 1.648 (.192)*	*-.940 (.391)*	
$\lambda_{k'}^B$	*.000 (1.000)*	*.920 (2.510)*	*.139 (1.149)*	
δ				*1.978 (7.226)*
Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	10.129	3	.017	- 5.170

Para ilustrar el significado del parámetro diagonal constante, puede comprobarse que, para dos categorías diferentes –*k* y *k'*– de *A* y *B* respectivamente, la razón de *odds* diagonal resulta igual a

$$\theta_{kk'} = \frac{m_{kk} m_{k'k'}}{m_{kk'} m_{k'k}} = \exp(2\delta) \qquad \text{(Ec. 6.5)}$$

De este modo, $\theta_{12}=\theta_{13}=\theta_{23}=\exp(2\times 1.978)=52.200.$ Si no hubiera ninguna coincidencia entre ambos jueces en la probabilidad de elegir una casilla diagonal –o lo que es lo mismo, una casilla donde se produce acuerdo– respecto a la probabilidad de elegir una casilla no diagonal –donde no se produce acuerdo– la magnitud de θ –y por ende de δ – sería igual a 1. Cuanto mayor resulte su tamaño, mayor será la probabilidad de que se produzca acuerdo entre ambos jueces.

Aplicando la Ecuación 6.3 se puede obtener el equivalente, para este modelo, de un coeficiente descriptivo que resulta ser igual a

$$\hat{\psi}_{QIC}=.372-\frac{.372}{7.226}+.180-\frac{.180}{7.226}+.167-\frac{.167}{7.226}=.620$$

un valor bastante mayor que cualquiera de los cuatro coeficientes descriptivos anteriormente citados, aunque maneja una escala de medida diferente, como se apuntó más arriba.

Esta medida de acuerdo, como se verá en el capítulo siguiente, es equivalente al coeficiente de acuerdo propuesto por Aickin (1990).

6.3.4. El modelo de cuasi-independencia homogéneo (Modelo QIH)

Otro procedimiento para simplificar el modelo de cuasi-independencia descansa en asumir que ambos jueces son homogéneos en su respuesta al clasificar una proporción similar de ítemes; dicho en otras palabras, postular que sus distribuciones marginales son idénticas y por tanto la igualdad de sus parámetros, $\lambda_k^A = \lambda_{k'}^B$. El número de grados de libertad que se liberan como consecuencia de esta restricción, en el caso de 2 jueces, es igual a $K-1$. Por el tipo de restricción utilizada, lo llamaremos como modelo de cuasi-independencia homogéneo (**Modelo QIH**). Fue considerado por Schuster y Smith (2002) quienes lo denominaron modelo de jueces homogéneos (*homogeneous raters model*), en contraposición al modelo de cuasi-independencia, al que llamaron modelo de observadores heterogéneos (*heterogeneous raters model*). Su ecuación es, para cualquier categoría de *k* de *A* o *k'* de *B,*

$$\log(m_{kk'}) = \lambda + \lambda_k^{A=B} + \delta_k \qquad \text{(Ec. 6.6)}$$

donde $\lambda_k^{A=B}$ es un conjunto de *K* parámetros que se asumen iguales para ambos evaluadores.

La Tabla 6.5 presenta el resumen del ajuste del Modelo QIH. Destacamos algunas de sus características.

Tabla 6.5. Resumen del ajuste del Modelo QIH

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61.000** (.372)	**15.000** (.091)	**3.000** (.018)	79.000 (.482)
neutral	**15.000** (.092)	**26.000** (.159)	**5.000** (.030)	46.000 (.280)
negativa	**3.000** (.018)	**5.000** (.030)	**31.000** (.189)	39.000 (.238)
Marginales	79.000 (.482)	46.000 (.280)	39.000 (.237)	164.000 (1.000)
Parámetros	*positiva*	*neutral*	*negativa*	
λ				*2.197 (9.000)*
$\lambda_k^{A=B}$	*.000 (1.000)*	*-0.511 (1.667)*	*-1.099 (.333)*	
δ_k	*1.914 (6.778)*	*0.0392 (1.040)*	*3.434 (31.000)*	
Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	22.585	3	.000	7286

- En primer lugar, las frecuencias marginales esperadas son exactamente iguales y las casillas no diagonales son simétricas, una consecuencia directa del supuesto de jueces homogéneos; además, como en el caso del Modelo QI, los valores diagonales vuelven a ser "fijados" y se estiman tantos parámetros diagonales como niveles tiene la variable categórica.

- En segundo lugar, los parámetros diagonales del modelo apuntan que el grado de acuerdo entre jueces es mayor –en efecto, hay más acuerdo puro ya que existe menor participación del azar– en la categoría 3, después le sigue la categoría 1 y finalmente la categoría 2.

- Y en tercer lugar, este modelo tampoco se ajusta, ya que la desvianza del modelo, con 3 grados de libertad residuales, es $L_3^2 = 22.585$; $P = .000$.

Asimismo, puede obtenerse un coeficiente de acuerdo para este particular modelo loglineal aplicando la Ecuación 6.3. La **restricción de homogeneidad marginal** es precisamente la característica esencial que distingue el coeficiente π de los restantes coeficientes descriptivos; el producido por este modelo es equivalente al coeficiente π, siendo las diferencias atribuibles al cambio de escala utilizado. En este caso el valor del coeficiente es

$$\psi_{QIH} = .372 - \frac{.372}{6.778} + .158 - \frac{.158}{1.040} + .189 - \frac{.189}{31.000} = .506$$

un valor considerablemente menor que cualquiera de los coeficientes descriptivos clásicos.

La característica más destacada del Modelo QIH radica en la restricción de homogeneidad marginal, como queda patente en la primera parte de la Tabla 6.5, lo que implica que ambos jueces no clasifican las categorías de la variable de respuesta de la misma forma. Pero en muchas ocasiones prácticas donde quepa suponer un grado similar de experiencia y la aplicación de unos criterios estrictos para la clasificación de los ítemes, este modelo probablemente ajustará bastante bien.

6.3.5. El modelo de cuasi-independencia constante y homogéneo (Modelo QICH)

Cuando el Modelo QIH se ajusta bien, una forma de simplificarlo es postulando un parámetro diagonal constante, en lugar del conjunto de parámetros diagonales. El modelo resultante es una fusión de los Modelos QIC y QIH, y su característica esencial es el uso de dos restricciones: la **restricción de homogeneidad marginal** y la **restricción de parámetro diagonal constante**. Precisamente por esta razón se denomina modelo de cuasi-independencia homogéneo con parámetro diagonal constante (**Modelo QICH**). La ecuación del modelo, para cualquier combinación k de A y k' de B, es la siguiente

$$\log(m_{kk'}) = \lambda + \lambda_k^{A=B} + \delta \qquad \text{(Ec. 6.7)}$$

donde al igual que en el Modelo QIH, se asume que $\lambda_k^A = \lambda_{k'}^B$. La Tabla 6.6 resume el ajuste del Modelo QICH, del que destacamos lo siguiente:

- En la primera sección, puede observarse la homogeneidad de los marginales y la simetría de los elementos no diagonales de la tabla de acuerdo.

- En la segunda sección, el parámetro diagonal único revela que el grado de acuerdo entre jueces es moderado.

- Y en la tercera sección se demuestra que el ajuste dista mucho de ser satisfactorio: $L_5^2 = 40.059$; $P = .000$ y $gl_R = 5$.

Tabla 6.6. Resumen del ajuste del Modelo QICH

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61.782** (.377)	**9.075** (.055)	**8.143** (.050)	79.000 (.482)
neutral	**9.075** (.055)	**31.143** (.190)	**5.782** (.035)	46.000 (.280)
negativa	**8.143** (.050)	**5.782** (.035)	**25.075** (.153)	39.000 (.238)
Marginales	79.000 (.482)	46.000 (.280)	39.000 (.238)	164.000 (1.000)

Parámetros	*positiva*	*neutral*	*negativa*	
λ				*2.548 (12.782)*
$\lambda_k^A = \lambda_{k'}^B$	*.000 (1.000)*	*-.343 (.710)*	*-.451 (.637)*	
δ				*1.576 (4.833)*

Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	40.059	5	.000	14.560

La medida de acuerdo derivada de este coeficiente, aplicando la Ecuación 6.3, resulta

$$\psi_{QICH} = .377 - \frac{.377}{4.833} + .190 - \frac{.190}{4.833} + .153 - \frac{.153}{4.833} = .571$$

un valor prácticamente similar de los coeficientes descriptivos clásicos.

6.3.6. El modelo de cuasi-independencia uniforme (Modelo QIU)

Un paso más en la simplificación del modelo de cuasi-independencia es la **restricción de uniformidad de las categorías**, que se basa en considerar iguales todas las combinaciones de respuestas que no corresponden a casillas de acuerdo (dicho de otro modo, las casillas no diagonales). En este caso, no es preciso incluir parámetros que tomen en cuenta las diferencias entre categorías, es decir, el conjunto de parámetros λ_k^A y $\lambda_{k'}^B$. No obstante, el modelo requiere contemplar los parámetros diagonales δ_k.

Este modelo es equivalente al que se asume para obtener el coeficiente σ de Bennet y otros (1954), así como al de las versiones de Maxwell (1977) y de Brennan y Prediger (1981).

La ecuación general del modelo, para cualquier combinación de categorías *k* de *A* y *k'* de *B*, es la siguiente

$$\log(m_{kk'}) = \lambda + \delta_k \qquad \text{(Ec. 6.8)}$$

donde se ha prescindido de los parámetros λ_k^A y $\lambda_{k'}^B$ al asumir que se produce una respuesta uniforme en todas las categorías de respuesta.

La Tabla 6.7 es un resumen del **Modelo QIU**. Entre las características de este modelo destacamos:

- En primer lugar se presentan los valores ajustados, donde se observa que todas las casillas externas a la diagonal principal tienen el mismo valor esperado y como resultado los marginales son homogéneos entre ambos jueces; igualmente, las casillas diagonales son "fijadas" en su valor muestral, como resultado de incluir el conjunto de parámetros diagonales en el proceso de ajuste.

- En segundo lugar, la estimación de los parámetros prescinde de los estimadores λ_k^A y $\lambda_{k'}^B$, y contiene únicamente los parámetros diagonales δ_k, donde se puede observar que la estimación es sensiblemente mejor en la primera categoría que en las restantes.

- Por último, el ajuste del modelo no es tampoco aceptable, con 5 grados de libertad residuales $L_5^2 = 43.047$; $P = .000$.

Tabla 6.7. Resumen del ajuste del Modelo QIU

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61.000** (.371)	**7.667** (.047)	**7.667** (.047)	76.333 (.465)
neutral	**7.667** (.047)	**26.000** (.158)	**7.667** (.047)	41.333 (.252)
negativa	**7.667** (.047)	**7.667** (.047)	**31.000** (.189)	46.333 (.283)
Marginales	76.333 (.465)	41.333 (.252)	46.334 (.283)	164.000 (1.000)
Parámetros	*positiva*	*neutral*	*negativa*	
λ				*2.037 (7.667)*
δ_k	*2.074 (7.957)*	*1.221 (3.391)*	*1.397 (4.044)*	
Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	43.047	5	.000	17.548

La restricción de uniformidad marginal se asume cuando se utiliza el coeficiente σ. Aplicando la Ecuación 6.3 se obtiene un coeficiente equivalente con el Modelo QIU que, aunque con una métrica diferente, resulta igual a

$$\psi_{QIC} = .372 - \frac{.372}{7.957} + .159 - \frac{.159}{3.931} + .189 - \frac{.189}{4.0435} = .579$$

y es exactamente el mismo valor que se obtuvo con el coeficiente σ.

6.3.7. El modelo de cuasi-independencia constante con asociación uniforme (Modelo QICAU)

Se han considerado hasta ahora en este capítulo modelos que asumen que los evaluadores clasifican los ítemes en categorías nominales. Pero en muchas ocasiones las categorías presentan niveles de medida ordinal o incluso intervalar. Para estos casos, Agresti (1988, 1992) propuso utilizar modelos loglineales que contengan uno o más parámetros con el objeto de evaluar la **asociación lineal –o asociación uniforme– entre las respuestas de los jueces**. La ecuación de este modelo, para cualquier combinación de categorías k de A y k' de B, es la siguiente

$$\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \beta u_k v_{k'} + \delta \qquad \text{(Ec. 6.9)}$$

donde las puntuaciones u_k –para el juez A– y $v_{k'}$ –para el juez B– son puntuaciones fijas conocidas de las categorías ordinales y β es el peso del producto escalar de las puntuaciones. Lo más común es emplear como puntuaciones los índices de orden de las variables (1, 2, 3, ... , K), aunque hay otras alternativas para definir tales puntuaciones (véase Clogg y Shihadeh, 1994). Con el producto de las puntuaciones u_k y v'_k tratadas como rangos, el resultado es un término de interacción que implica un único parámetro, al que Goodman (1979) denominó **asociación uniforme**.

La Tabla 6.8 resume el ajuste del **Modelo QICAU.** Entre sus propiedades más interesantes destacamos:

Tabla 6.8. Resumen del ajuste del Modelo QICAU

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**62.121** (.379)	**24.948** (.152)	**4.932** (.030)	92.000 (.561)
neutral	**2.811** (.017)	**26.000** (.158)	**4.189** (.025)	33.000 (.252)
negativa	**1.068** (.007)	**8.052** (.049)	**29.879** (.182)	39.000 (.283)
Marginales	66.000 (.402)	59.000 (.360)	39.000 (.237)	164.000 (1.000)
Parámetros	*positiva*	*neutral*	*negativa*	
λ				*2.106 (8.216)*
λ_k^A	*.000 (1.000)*	*- 2.891 (.056)*	*-4.768 (.009)*	
$\lambda_{k'}^B$	*.000 (1.000)*	*- .708 (.493)*	*- 3.238 (.039)*	
δ				*1.114 (3.046)*
β				*.909 (7.640)*
Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	1.074	2	.585	- 9.130

- En la primera sección se presentan los valores ajustados, donde no se detecta homogeneidad ni uniformidad marginal.
- En segundo lugar, los estimadores de los parámetros que, respecto del Modelo QIC, utilizan como novedad la introducción del componente de asociación uniforme.
- Y al final, el ajuste del modelo, igualmente aceptable: $L_2^2 = 1.074$; $P = .585$ y $gl_R = 2$.

Adviértase que este modelo es un compuesto de un modelo básico –el Modelo I– al que se ha añadido un parámetro diagonal constante y un parámetro de asociación uniforme. A partir del modelo de independencia pueden por tanto definirse tres modelos posibles, a saber: el Modelo QIC, el modelo de asociación uniforme AU –que incluye efectos para los jueces *A* y *B* y un parámetro de asociación uniforme– y el Modelo QICAU que tratamos aquí, que es una fusión de los modelos QIC y AU.

Una característica relevante del Modelo AU que justifica su denominación es el hecho de que el grado de asociación, medido en términos de una razón de *odds,* es uniforme para todas las razones de *odds* locales, es decir, para todas las razones que involucran categorías contiguas, tanto para las razones de *odds* diagonales involucradas en el acuerdo (casillas <12-12> y <23-23>) como para las razones *odds* no diagonales involucradas en el desacuerdo (casillas <12-23> y <23-12>). Si bien el modelo de asociación uniforme no pertenece a la familia de modelos de cuasi-independencia, el valor de su razón de *odds* es $\theta = 6.575$. Cuando se introduce el parámetro diagonal, el resultado es por lo tanto un modelo de la familia de cuasi-independencia, y la razón de *odds* derivada es uniforme para las casillas diagonales y uniforme asimismo, aunque de magnitud diferente, para las casillas no diagonales. En concreto,

$$\theta_{12-12} = \frac{(62.121)(26)}{(24.948)(2.811)} = 23.030 ; \theta_{23-23} = \frac{(26)(29.879)}{(4.189)(8.052)} = 23.030$$
$$\theta_{12-23} = \frac{(24.948)(4.189)}{(4.932)(26)} = .815 ; \theta_{23-12} = \frac{(2.811)(8.052)}{(26)(1.068)} = .815$$

lo que refleja que la asociación es considerablemente mayor –más de 28 veces mayor– en las casillas involucradas con el acuerdo que en las involucradas con el desacuerdo, como es de esperar en estudios que evalúan el acuerdo entre observadores.

6.3.8. Comparaciones entre modelos

Si hay más de un modelo óptimamente ajustado, subsiste finalmente la cuestión acerca de qué modelo es más apropiado y debe por tanto ser objeto de interpretación. En el ejemplo que nos ocupa, hay dos que se ajustan aceptablemente, a saber, los Modelos QI y QICAU (véase Tabla 6.9).

El procedimiento más común es la comparación de modelos que se aplica a dos o más modelos anidados(e.g. Ato y López, 1996). Sin embargo, aunque los dos modelos pertenecen a la misma familia, no son estrictamente comparables, porque el primero contiene el conjunto de parámetros diagonales δ_k y el segundo un parámetro diagonal único, δ. Una forma complementaria para decidirse por uno de los dos modelos que se ajustan aceptablemente (i.e. Ato y Vallejo, 2007), consiste en hacer uso del criterio de información bayesiano de Schwartz Raftery (*Bayesian Information Criterium*, criterio BIC), por el que de dos modelos que se comparan, es mejor el modelo con puntuación más baja. El más aceptable aquí es –según este criterio– el **Modelo QICAU**.

Tabla 6.9. Comparación de modelos mediante criterio BIC

QI

Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	.182	1	.669	- 4.920

QICAU

Ajuste	L^2	*gl residual*	*probabilidad*	*BIC*
	1.074	2	.585	- 9.130

No obstante, puede observarse que comparando los Modelos QIC y QICAU el primero no se ajusta aceptablemente ($p < .10$) y el segundo sí se ajusta; y por otra parte comparando los Modelos AU y QICAU, el primero tampoco se ajusta $(L_3^2 = 12.823\,;\, P = .005)$ mientras que el segundo sí.

En consecuencia, parece sensato concluir que el componente de asociación uniforme del Modelo QICAU es el determinante esencial para el ajuste del modelo.

6.4. Algunas generalizaciones del modelo loglineal

La familia de modelos loglineales que se ha considerado en este capítulo, permite algunas generalizaciones de gran interés aplicado:

- Todos los modelos admiten la inclusión de covariantes, tanto

numéricas como categóricas, que no son posibles con coeficientes de tipo descriptivo.

- Los modelos pueden ampliarse con facilidad para evaluar el acuerdo entre más de 2 jueces.

En ambos casos, se plantean ciertas cuestiones de los modelos resultantes que es preciso especificar, como tratamos más adelante.

6.4.1. La inclusión de covariantes

La inclusión de covariantes se puede realizar de una de dos formas posibles (von Eye y Mun, 2005:41-53), dependiendo de si las covariantes tienen naturaleza categórica o cuantitativa:

- Como ejemplo del primer caso la/s covariante/s serán de naturaleza categórica, un estudio en el que dos jueces valoran las severidad de un determinado síntoma en una muestra de estudiantes, la mitad de los cuales son varones; el análisis con los coeficientes descriptivos no permitiría discutir el efecto del sexo sobre el acuerdo, porque los coeficientes no son comparables de muestra a muestra, pero la inclusión de covariantes con modelos loglineales es totalmente válida.

- Como ejemplo del segundo caso, i.e. covariante de naturaleza numérica, usaremos el mismo estudio anterior en el que se valora el síntoma específico y otro síntoma afín que potencialmente podría afectarle.

Graham (1995) propuso extender el Modelo QIC de Tanner y Young (1985) con la incorporación de variables categóricas. Von Eye y Mun (2005) han hecho lo propio con variables cuantitativas. Con dos observadores, el modelo propuesto por Graham (1995) es, para cualquier combinación de categorías k de A y k' de B,

$$\log(m_{kk'l}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \lambda_l^C + \delta^{AB} + \delta_{kl}^{AC} + \delta_{k'l}^{BC} + \delta_{kk'l}^{ABC} \qquad \text{(Ec. 6.10)}$$

donde $\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \delta^{AB}$ corresponde al Modelo QIC, al que se incorporan como nuevos parámetros δ_{kl}^{AC}, que representa la asociación parcial entre el juez A y la covariante categórica C, controlando al juez B, $\delta_{k'l}^{BC}$ que representa la asociación entre el juez B y la covariante C, controlando al juez A. Y puesto que δ^{AB} es un parámetro diagonal constante que no asume variación entre los niveles de la covariante categórica, el modelo incluye el conjunto de parámetros $\delta_{kk'l}^{ABC}$ que asume la variación entre los niveles de la covariante, con tantos grados de libertad como niveles tenga ésta.

6.4.1.1. Ejemplo 6.2: Los datos de Jackson y otros (2001)

Un estudio de Jackson y otros (2001; véase asímismo von Eye y Mun, 2005) presentó una cuestión de una encuesta en dos ocasiones en el tiempo –variables *A* y *B*– a dos muestras –variable *C*– de un total de 442 adolescentes, la primera con historia familiar de alcoholismo y la segunda sin historia familiar. La cuestión se refería a la consideración acerca del nivel de la ingesta de alcohol en dos ocasiones separadas por 3 años, con categorías "abstemio", "ocasional", "moderado" e "intenso". Los datos originales, que constituyen el Ejemplo 6.2, se muestran en la Tabla 6.10.

Tabla 6.10. Datos empíricos del Ejemplo 6.2

Primera ocasión (*A*)	Segunda ocasión (*B*) - 3 años después-				
	abstemio	*ocasional*	*moderado*	*intenso*	Marginal
Con historia familiar de alcoholismo					
abstemio	20	8	3	2	33
ocasional	6	8	3	2	19
moderado	2	13	8	10	33
intenso	9	8	27	96	140
Marginal	37	37	41	110	225
Sin historia familiar de alcoholismo					
abstemio	13	8	6	0	27
ocasional	0	6	13	1	20
moderado	3	13	26	15	57
intenso	3	0	26	84	113
Marginal	19	27	71	100	217

La Tabla 6.11 resume los resultados del proceso de ajuste con varios de los modelos probados en el trabajo de Jackson y otros, ninguno de los cuales se ajusta adecuadamente.

- En primer lugar, el modelo básico QIC propuesto por Graham (1995) y formulado en la Ecuación 6.10 obtiene una desvianza de $L^2_{17} = 124.7$ *;* $P = .000$, con BIC = 21.145.

- En segundo lugar, puesto que el modelo básico no se ajusta, se incorporó un parámetro de asociación uniforme y el ajuste descendió significativamente con una desvianza de $L^2_{16} = 50.383$ *;* $P = .000$, y criterio BIC = 18.383.

- Puesto que el modelo anterior tampoco ajustaba, en tercer lugar, se optó por utilizar una covariante numérica, que para este caso se definió tomando las proporciones de la transición de la primera a la segunda ocasión de medida. El ajuste resultante descendió hasta $L^2_{15} = 27.702$ *;* $P = .024$ de desvianza, con un criterio BIC = -2.298.

- El modelo final utilizado en el estudio, el único que podría considerarse para su interpretación, es el siguiente:

$$\log(m_{kk'l}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \lambda_l^C + \lambda^X + \delta^{AB} + \delta_{kl}^{AC} + \delta_{k'l}^{BC} + \delta_{kk'l}^{ABC} + \beta u_k v_{k'} \quad \text{(Ec. 6.11)}$$

De los parámetros del Modelo QICAU + covariante numérica que se muestran en la Tabla 6.11, se deduce que los efectos del parámetro diagonal constante $(\delta_{kk'l}^{ABC})$, así como los del parámetro de asociación uniforme (β) y los del parámetro para la covariante cuantitativa (λ^X), son todos estadísticamente significativos.

Tabla 6.11. Resultados del ajuste de modelos para los datos del Ejemplo 6.2

Ajuste	L^2	*gl residual*	ΔL^2	$P > \chi^2_{gl}$	*BIC*
QIC	124.697	17	-	.000	21.145
QICAU	50.383	16	74.314(1)	.000	18.383
QICAU + covariante	27.702	15	22.681(1)	.024	-2.298
	abstemio	*ocasional*	*moderado*	*intenso*	
λ					1.357
λ_k^A	*.000*	*-.886*	*-1.045*	*-.090*	*
$\lambda_{k'}^B$	*.000*	*-.706*	*-1.946*	*-2.918*	**
λ_l^C	*.000 (l1)*	*-.552 (l2)*			(NS)
λ^X					3.174***
λ_{kl}^{AC}	*.000*	*-.328*	*.157*	*-.778*	**
$\lambda_{k'l}^{BC}$	*.000*	*.399*	*1.472*	*1.117*	**
δ^{ABC}					-.366**
β_u					.324***

Nota: * p<.05; ** p<.01; *** p<.001

6.4.2. Más de dos jueces. Ejemplo 6.3: Los datos de von Eye y Mun (2005)

Muchos estudios de acuerdo utilizan más de 2 observadores. La segunda generalización que se discute en este capítulo es el tratamiento de la familia de modelos de cuasi-independencia con 3 o más jueces. Para ello, nos serviremos del Ejemplo 6.3, en el que $J = 3$ jueces clasifican $N = 113$ objetos en una escala de $K = 3$ categorías de respuesta. Los datos empíricos (von Eye y Mun, 2005:65) se reproducen en una tabla de acuerdo multidimensional como la presentada en la Tabla 6.12.

Tabla 6.12. Datos empíricos del Ejemplo 6.3

Juez A = 1	*Juez C*			
Juez B	*grado 1*	*grado 2*	*Grado 3*	Marginal
grado 1	4	3	6	13
grado 2	2	1	3	6
grado 3	2	2	17	21
Marginal	8	6	26	40
Juez A = 2	*grado 1*	*grado 2*	*grado 3*	Marginal
grado 1	0	1	2	3
grado 2	1	1	1	3
grado 3	0	0	4	4
Marginal	1	2	7	10
Juez A = 3	*grado 1*	*grado 2*	*grado 3*	Marginal
grado 1	0	1	3	4
grado 2	0	1	8	9
grado 3	0	4	96	100
Marginal	0	6	107	113

El modelo más básico posible es el de independencia (**Modelo I**), que se formula para cualquier combinación de las categorías k de A, k' de B y k'' de C mediante

$$\log(m^{ABC}_{kk'k''}) = \lambda + \lambda^{A}_{k} \lambda^{B}_{k'} + \lambda^{C}_{k''} \qquad \text{(Ec. 6.12)}$$

que obtiene un ajuste de $L^2_{20} = 75.100; P = .000$, de modo que el modelo se rechaza; existe algún tipo de asociación en la tabla de acuerdo multidimensional que es necesario desvelar.

El segundo modelo a considerar es el de cuasi-independencia (**Modelo QI**), que se formula para cualquier combinación de las categorías k, k' y k'' de A, B y C

$$\log(m_{kk'k''}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \lambda_{k''}^C + \delta_k \qquad \text{(Ec. 6.13)}$$

al que, respecto al modelo anterior, se han incorporado tres parámetros diagonales, que afectan a las casillas <111>, <222> y <333> de la tabla de acuerdo. El ajuste del modelo es $L^2_{17} = 18.626\,;\, P = .350\,(BIC = -67.800)$, que puede aceptarse para su interpretación.

El siguiente modelo a considerar es el de cuasi-independencia constante (**Modelo QIC**), para el que existen varias versiones posibles.

- La primera es el que contempla un parámetro diagonal por cada uno de los pares de jueces (AB, AC y BC) cuya formulación es:

$$\log(m_{kk'k''}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \lambda_{k''}^C + \delta^{AB} + \delta^{AC} + \delta^{BC} \qquad \text{(Ec. 6.14)}$$

  que incluye un parámetro para cada una de las posibles combinaciones de pares de jueces; el modelo obtiene un ajuste aceptable:

  $L^2_{17} = 17.026\,;\, P = .453$ (BIC = -69.460)

- Un segundo modelo, que prescinde de los parámetros de las combinaciones de pares de jueces y propone en su lugar un parámetro diagonal único para todos los jueces, cuya formulación es:

$$\log(m_{kk'k''}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \lambda_{k''}^C + \delta^{ABC} \qquad \text{(Ec. 6.15)}$$

obtiene un ajuste aceptable $L^2_{19} = 19.428$; $P = .430$ (BIC = -77.236).

y, atendiendo al BIC, resulta incluso algo mejor que los anteriores.

- Una tercera alternativa es un modelo conjunto que incluya un parámetro diagonal para cada par de jueces y otro parámetro para el conjunto, cuyo ajuste es $L^2_{16} = 16.502$; $P = .419$ (BIC = -64.900),

pero como se puede observar no mejora en absoluto el ajuste de los dos modelos.

Todos los modelos que implican homogeneidad entre jueces también se descartan, por elevar considerablemente la desvianza pero con un incremento no equivalente de los grados de libertad residuales. Eso implica, que la mejor opción es la segunda, i.e. el Modelo QIC con parámetro diagonal único.

Capítulo 7

Modelos mixtura

7.1. Introducción

Una generalización del enfoque loglineal consiste en incluir en los modelos de una determinada familia una o más variables latentes no observables y asumir que los ítemes que los jueces deben clasificar se extraen de una población que representa una mezcla –o mixtura– de dos subpoblaciones finitas. Los modelos resultantes se basan en los modelos loglineales, y se estiman asimismo por máxima verosimilitud y se denominan **modelos de clase latente** o, más precisamente, **modelos mixtura** (*mixture model*s), o también **modelos con mezcla de distribuciones.**

En un modelo mixtura cada subpoblación identifica un conglomerado de ítemes, sujetos u objetos homogéneos, por ejemplo, la subpoblación que representa acuerdo sistemático entre tales ítemes (o subpoblación *X1*), que afecta únicamente a las casillas de la diagonal principal de la tabla de acuerdo, y la subpoblación que representa al acuerdo aleatorio y desacuerdo (o subpoblación *X2*), la cual afecta por igual a todas las casillas de la tabla de acuerdo (Agresti, 1989, 1992; Guggenmoos-Holtzman, 1993; Guggenmoos-Holtzman y Vonk, 1998; Schuster, 2002; Schuster y Smith, 2002). De ahí se deduce la definición de dos subtablas de acuerdo que son de hecho clases latentes y que no necesariamente asumen la misma distribución, sino que representan una mezcla de distribuciones (Ato, Benavente, Rabadán y López, 2004).

La distribución conjunta resultante es por tanto una mezcla de una distribución que asume acuerdo perfecto entre jueces u observadores (la subpoblación *X1*), con probabilidad μ, y una distribución que asume independencia entre esos jueces (la subpoblación *X2*), con probabilidad $1-\mu$ (Agresti, 1989). Esta composición se basa en el **supuesto de independencia local** que deben cumplir los modelos de clase latente, por el cual los jueces *A* y *B* son independientes dentro de una clase latente determinada. Esta propuesta abre nuevas e interesantes perspectivas para la interpretación del acuerdo entre jueces y posibilitan la definición de medidas de acuerdo similares a las derivadas del modelado loglineal pero con una definición más sutil y detallada del acuerdo.

Existen otros enfoques alternativos para evaluar el acuerdo con datos categóricos con modelos de clase latente y modelos de rasgo latente (Clogg,

1979; Coull y Agresti, 2003; Dawid y Skine, 1979; Kraemer, 1979; Uebersax, 1992; Uebersax y Grove, 1990 y Williamson y Manatunga, 1997), pero no seguiremos ninguna de estas líneas en este trabajo. Nos ceñimos fundamentalmente al enfoque originalmente propuesto en los trabajos citados de Agresti (1989, 1992), Guggenmoos-Holtzma-Volk (1993, 1998) y Schuster-Smith (2002), porque en nuestra opinión se trata de la línea más coherente desarrollada hasta ahora.

7.2. La familia QI de modelos mixtura

De forma similar a los modelos loglineales, los modelos mixtura se definen a partir de una familia con un modelo base, que da nombre a la familia, y un conjunto de modelos más restrictivos, cada uno de los cuales supone la aplicación de alguna/s forma/s de restricción.

En este capítulo se introducen los modelos mixtura que se desarrollan a partir de la familia de Modelos QI utilizando el mismo ejemplo (Ejemplo 6.1, Tabla 6.1 del Capítulo 6) que sirvió para ilustrar los modelos loglineales tratados en el capítulo anterior. Recuérdese que los coeficientes descriptivos – con sus respectivos errores típicos entre paréntesis obtenidos por *jackknife*– fueron los que se muestran a continuación: $\hat{\kappa}=.5653(.053)$, $\hat{\pi}=.5567(.056)$, $\hat{\gamma}=.5556(.061)$ y $\hat{\sigma}=.5793(.053)$.

7.2.1. El modelo mixtura básico: Modelo mixtura QI

El modelo mixtura de cuasi-independencia puede definirse de dos formas posibles, mediante la formulación de un modelo loglineal o bien a través de la estimación de probabilidades. Utilizando la primera de las dos opciones, la formulación del modelo, para cada combinación de *k* de *A* y *k′* de *B*, resulta

$$\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \xi_k \qquad \text{(Ec. 7.1)}$$

donde, respecto del modelo loglineal de la Ecuación 6.2, lo único que cambia es la sustitución de los parámetros ξ_k por los parámetros δ_k.

Los parámetros diagonales ξ_k de los modelos mixtura se relacionan estrechamente con los parámetros δ_k de los modelos loglineales ya que (Guggenmoos-Holtzman, 1989),

$$\exp(\xi_k) = \exp(\delta_k) - 1 \qquad \text{(Ec. 7.2)}$$

y la estimación de una medida de acuerdo basada en el Modelo mixtura QI tiene lugar (Goggenmoos-Holtzman y Vonk 1998; Schuster y Smith, 2002), aplicando una modificación de la Ecuación 6.3, a saber

$$\mu = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi_k) + 1} \right] \qquad \text{(Ec. 7.3)}$$

y su complemento es $1-\mu$. En la Ecuación 7.3, μ representa la proporción que corresponde a la subpoblación de acuerdo sistemático; y $(1-\mu)$ simboliza la proporción que representa la subpoblación de acuerdo aleatorio y desacuerdo. Como consecuencia, el ajuste de los modelos mixtura es el mismo que el de los modelos loglineales, pero la distinción entre las dos clases latentes permite una interpretación más detallada del acuerdo existente entre jueces.

Tabla 7.1. Resultados del Modelo mixtura QI

Frecuencias observadas					*Probabilidades observadas*			
	k1	*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**61.000**	**26.319**	**4.682**	92	**.3720**	**.1605**	**.0285**	.5610
k2	**3.681**	**26.000**	**3.318**	33	**.0225**	**.1585**	**.0202**	.1564
k3	**1.319**	**6.681**	**31.000**	39	**.0080**	**.0407**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.0000
Prob. condicionales X1					*Prob. condicionales X2*			
k1	**.3403**	**.0000**	**.0000**	.3403	**.0317**	**.1605**	**.0285**	.2207
k2	**.0000**	**.0448**	**.0000**	.0448	**.0224**	**.1138**	**.0202**	.1564
k3	**.0000**	**.0000**	**.1818**	.1818	**.0080**	**.0407**	**.0072**	.0559
Total	.3403	.0448	.1818	***.5668***	.0621	.3150	.0559	***.4332***
Prob. clase latente X1 (ϕ_k) $(\mu = .5668)$					*Prob. clase latente X2* $(\psi_k^A, \psi_{k'}^B)$ $(1-\mu = .4332)$			
A	**.6003**	**.0790**	**.3207**	1.0000	**.5095**	**.3612**	**.1293**	1.0000
B	**.6003**	**.0790**	**.3207**	1.0000	**.1435**	**.7272**	**.1293**	1.0000

La Tabla 7.1 presenta todos los resultados intermedios del proceso de cálculo para el Modelo mixtura QI.

- En la primera sección se muestran las frecuencias –a la izquierda– y probabilidades –a la derecha– observadas de la tabla de acuerdo.

- En la segunda sección se exponen las frecuencias y probabilidades esperadas.
- En la tercera sección se presenta las probabilidades condicionales de la tabla, dado un nivel de la clase latente;esta descomposición es el objetivo esencial que persigue el análisis de clase latente.
- Y, finalmente se indican las probabilidades de ambas clases latentes ; para cada uno de los dos evaluadores .

Nótese que para la clase latente *X1* (subtabla de acuerdo sistemático), las probabilidades de clase latente –que llamaremos ϕ_k – son iguales en *A* y *B*, ya que la subtabla sólo contiene valores en la diagonal principal. Por el contrario, para la segunda clase latente *X2* (subtabla de acuerdo aleatorio y desacuerdo), las probabilidades de clase latente no coinciden en *A* y *B;* por ello las distinguimos utilizando ψ_k^A y $\psi_{k'}^B$.

Las probabilidades esperadas del modelo se obtienen a través de una suma ponderada de las probabilidades condicionales de ambas clases latentes (Schuster y Smith, 2002),

$$\hat{p}_{kk'} = \mu \phi_k + (1-\mu)\psi_k^A \psi_{k'}^B \qquad \text{(Ec. 7.4)}$$

donde ϕ_k son las probabilidades de clase latente de la subtabla de acuerdo sistemático y ψ_k^A y $\psi_k^{B\prime}$ son, respectivamente, las probabilidades de clase latente –para los jueces *A* y *B*– de la subtabla de acuerdo aleatorio y

desacuerdo. De modo que, para obtener la probabilidad esperada de la casilla <11> , se utiliza la siguiente composición

$$p_{11}=(.5668)(.6003)+(.4332)(.5095)(.1435)=.3403+.0317=.3720$$

En la Tabla 7.2. se expone el ajuste del Modelo mixtura QI y sus parámetros básicos:

- En primer lugar los parámetros diagonales ξ_k junto con sus equivalentes loglineales.
- En segundo lugar los parámetros ϕ_k y $\psi_k^A, \psi_{k'}^B$ y el parámetro μ para probabilidad latente de la subtabla de acuerdo sistemático y su complementario para probabilidad latente de la subtabla de acuerdo aleatorio y desacuerdo.

Tabla 7.2. Ajuste del Modelo mixtura QI

Ajuste Modelo QI	L^2	*gl residual*	*Probabilidad*	*BIC*
	.182	1	.669	- 4920
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta_k)$	*10.745*	*.394*	*25.083*	
$\exp(\xi_k)$	*11.745*	*1.394*	*26.083*	$\mu=.5668$
ϕ_k	*.6003*	*.0790*	*.3207*	$1-\mu=.4332$
ψ_k^A	*.5095*	*.3612*	*.1293*	
$\psi_{k'}^B$	*.1435*	*.7272*	*.1293*	

Nótese que se ha prescindido de toda información acerca de los parámetros λ_k^A y $\lambda_{k'}^B$ debido a que en este contexto no ofrecen información relevante. Adviértase que los parámetros ϕ_k ocupan una línea, ya que la subtabla con la que se asocia sólo contiene valores en la diagonal principal, y en consecuencia son iguales para ambos jueces.

El parámetro μ fue propuesto primero por Agresti (1989) y posteriormente tratado por Schuster (2002) y Schuster y Smith (2002), quienes lo llamaron **modelo de observadores heterogéneos** (*heterogeneous raters model*). Su interpretación es muy simple en este contexto, entendiéndose como la proporción de la subtabla de acuerdo sistemático; por tanto, es una medida depurada del grado de acuerdo existente entre los dos observadores. El valor del coeficiente descriptivo κ=.5653 es análogo al de μ para estos datos empíricos.

Una medida de acuerdo muy similar, también basada en el Modelo QI pero justificada desde los modelos de respuesta múltiple, fue propuesta por Martín y Femia (2005), que designaron como Δ. Ambas medidas son en la práctica equivalentes, pero se definen con un rango de valores diferente: mientras que μ queda dentro del rango 0/1, Δ oscila dentro de -1/+1.

Algunas de las propiedades del Modelo mixtura QI respecto de otros modelos que se consideraran más adelante son los siguientes:

1) El Modelo mixtura QI es el modelo básico, pero complejo, que no contiene restricciones; todos los modelos de la familia que se definen en relación al mismo asumen la introducción de algún tipo de

restricción.

2) Como consecuencia de la ausencia de restricciones, en el Modelo mixtura QI no hay relación entre los parámetros de clase latente, es decir, $\phi_k \neq \psi_k^A \neq \psi_{k'}^B$.

3) El índice de acuerdo para el Modelo mixtura QI se denomina μ (Agresti, 1989; Schuster, 2002; Schuster y Smith, 2002) y si se desea se puede obtener aplicando la ecuación general RCA, calculando la probabilidad esperada por azar a partir de las probabilidades latentes de la subtabla de acuerdo aleatorio

$$p_e^{\mu} = \sum_{k=1}^{K} \psi_k^A \psi_{k'}^B \qquad \text{(Ec. 7.5)}$$

que para los datos empíricos de la Tabla 7.1 es $p_e^{\mu} = .3525$, y obteniendo después el coeficiente

$$\mu = \frac{p_o - p_e^{\mu}}{1 - p_e^{\mu}} \qquad \text{(Ec. 7.6)}$$

que resulta $\mu = .5668$.

7.2.2. El Modelo mixtura QIC

Como modelo loglineal, la formulación del Modelo QIC, respecto del Modelo QI, asume que los parámetros diagonales son iguales entre sí, y por tanto se restringe su número a uno independientemente del número de categorías de la variable. Esta restricción es tanto más importante cuanto mayor es la dimensión de la tabla de contingencia. El Modelo mixtura QIC (modelo de cuasi-independencia constante) es similar al loglineal y se formula como

$$\log(m_{kk'}) = \lambda + \lambda_k^A + \lambda_{k'}^B + \xi \qquad \text{(Ec.7.7)}$$

donde ξ, como δ en los modelos loglineales, es un parámetro diagonal único. Respecto del Modelo QIC, $\exp(\xi) = \exp(\delta) - 1 = 7.2295 - 1 = 6.2295$.

La estimación de una medida de acuerdo asociada con el Modelo mixtura QIC se basa en un modelo de población que fue originalmente propuesto por Kraemer y Bloch (1988) y después reformulado por Aickin (1990), quien lo denominó α. En su origen la propuesta se formuló contra el supuesto de que el azar actuaba del mismo modo con todos los objetos. El coeficiente α es otra forma de aplicar la corrección del azar, amparada en el supuesto de que sólo una proporción de objetos son clasificados aleatoriamente. Aplicando la Ecuación 7.3, el coeficiente α de Aickin es

$$\alpha = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi) + 1} \right] = .6200$$

y su complemento $1-\alpha=.3800$. Como en todo modelo mixtura, α representa la proporción que corresponde a la subpoblación de acuerdo sistemático y $(1-\alpha)$ la proporción para la subpoblación de acuerdo aleatorio y desacuerdo o, lo que es lo mismo, la proporción de casos clasificados por azar. El ajuste del modelo mixtura es el mismo que el que se obtuvo para el modelo loglineal, pero la distinción entre las dos clases latentes permite en este caso una interpretación más cabal del acuerdo existente entre evaluadores, como se mostrará después. La Tabla 7.3 contiene un resumen de todos los resultados del proceso de cálculo del Modelo mixtura QIC.

Tabla 7.3: Resultados del Modelo mixtura QIC

Frecuencias observadas					*Probabilidades observadas*			
	k1	*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**61.076**	**21.212**	**9.713**	92	**.3724**	**.1293**	**.0592**	.5609
k2	**1.625**	**29.506**	**1.869**	33	**.0099**	**.1799**	**.0114**	.2012
k3	**3.299**	**8.283**	**27.419**	39	**.0201**	**.0505**	**.1672**	.2378
Total	66	59	39	164	.4024	.3598	.2378	1.0000
Prob. condicionales X1					*Prob. condicionales X2*			
k1	**.3209**	**.0000**	**.0000**	.3209	**.0515**	**.1293**	**.0592**	.2400
k2	**.0000**	**.1550**	**.0000**	.1550	**.0099**	**.0249**	**.0114**	.0462
k3	**.0000**	**.0000**	**.1441**	.1441	**.0201**	**.0505**	**.0231**	.0937
Total	.3209	.1550	.1441	***.6200***	.0815	.2047	.0937	***.3799***
Prob. clase latente X1 (Φ_k) $(\alpha = .6200)$					*Prob. clase latente X2* $(\psi_k^A, \psi_{k'}^B)$ $(1-\alpha = .4332)$			
A	**.5176**	**.2500**	**.2324**	1.0000	**.6318**	**.1216**	**.2467**	1.0000
B	**.5176**	**.2500**	**.2324**	1.0000	**.2148**	**.5387**	**.2467**	1.0000

La Tabla 7.4. presenta el ajuste del Modelo mixtura QIC y sus parámetros básicos:

- En primer lugar el parámetro diagonal único ξ junto con su equivalente loglineal

- En segundo lugar los parámetros ϕ_k y $\psi_k^A, \psi_{k'}^B$ y la medida de acuerdo α, para la probabilidad latente de la subtabla de acuerdo sistemático y su complementario $1-\alpha$, para la probabilidad latente de la subtabla de acuerdo aleatorio y desacuerdo.

Tabla 7.4. Ajuste del Modelo mixtura QIC

Ajuste Modelo QIC	L^2	*gl residual*	*Probabilidad*	*BIC*
	10.129	3	.018	4.13
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta)$				7.2295
$\exp(\xi)$				6.2295
ϕ_k	*.5176*	*.2500*	*.2324*	$\alpha=.6200$
ψ_k^A	*.6318*	*.1216*	*.2467*	$1-\alpha=.3800$
$\psi_{k'}^B$	*.2146*	*.5387*	*.2467*	

Resaltamos algunas de las propiedades más interesantes de este modelo.

1) Respecto del Modelo mixtura QI, el Modelo QIC incluye la **restricción de constancia diagonal**, que implica que ajusta mejor cuanto menores son las discrepancias entre los elementos de la diagonal principal. Una consecuencia de esta restricción es que las razones *odds* diagonales $\theta_{kk'}$ para todo $k \neq k'$ de la tabla de probabilidades esperadas son iguales: $\theta_{12}=\theta_{13}=\theta_{23}=\exp(2\delta)=52.300$, mientras que las razones *odds* diagonales de la tabla de probabilidades condicionales son homogéneas e iguales a la unidad: $\exp(\theta_{12})=\exp(\theta_{13})=\exp(\theta_{23})=1$.

Otra característica de la restricción es que se asume una **constante de proporcionalidad** (*Constant of Proportionality*, CP) para todas las categorías entre las subtablas de acuerdo sistemático (los parámetros ϕ_k) y acuerdo aleatorio y desacuerdo (los parámetros ψ_k^A y $\psi_{k'}^B$), es decir,

$$CP = \frac{\phi_k}{\psi_k^A \psi_{k'}^B} \qquad \text{(Ec. 7.8)}$$

Esta es la razón por la que Aickin (1990) denominó también a este modelo como **modelo de probabilidad predictiva constante**(*constant predictive probability model*). Para los datos empíricos del Ejemplo 6.1, la constante de proporcionalidad es $CP = 1/\sum \psi_k^A \psi_{k'}^B = 3.820$. Esto es, el producto de las probabilidades latentes marginales de la clase *X2* es proporcional a la probabilidad latente marginal de la clase *X1*, de tal modo que para la categoría k = 1, .5176 = (.6318)(.2146)(3.82), para k = 2, .2500 = (.1216)(.5487)(3.82), y para k = 3, 2324 = (.2467)(.2467)(3.820).

2) En el Modelo mixtura QIC, la probabilidad de acuerdo aleatorio se define –de igual modo que en el Modelo mixtura QI– como la suma de los productos de las probabilidades de acuerdo aleatorio: $p_e^{\alpha} = \sum_k \psi_k^A \psi_{k'}^B = .2620$. Por consiguiente, es posible calcular la

probabilidad de acuerdo sistemático –y la medida de acuerdo para el Modelo mixtura QIC– utilizando la ecuación general RCA,

$$\alpha = \frac{p_o - p_e^{\alpha}}{1 - p_e^{\alpha}} = \frac{.7195 - .2620}{1 - .2620} = .6200.$$

7.2.3. El Modelo mixtura QIH

Como modelo loglineal la formulación del Modelo QIH, respecto del Modelo QI, asume que los observadores *A* y *B* actúan de forma homogénea; eso significa que sus valores marginales son iguales, y por tanto se restringe el número de parámetros para contemplar solamente uno para los jueces. El Modelo mixtura QIH (modelo de cuasi-independencia homogéneo) resultante es similar al loglineal y se formula como

$$\log(m_{kk'}) = \lambda + \lambda_k^{A=B} + \xi_k \qquad \text{(Ec.7.9)}$$

donde ξ_k son los parámetros diagonales. Fue propuesto por Schuster y Smith (2002), quienes lo llamaron **modelo de observadores homogéneos** (*homogeneous raters model*).

La estimación de una medida de acuerdo asociada con el Modelo mixtura QIH fue propuesta por Schuster y Smith (2002) y se ha denominado λ.

Aplicando la Ecuación 7.3, esta medida de acuerdo resulta igual a

$$\lambda = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi_k) + 1} \right] = .5061$$

y su complemento es $1 - \lambda = .4939$. En esta ecuación, λ representa la proporción que corresponde a la subpoblación de acuerdo sistemático y $(1 - \lambda)$ la proporción para la subpoblación de acuerdo aleatorio y desacuerdo. La Tabla 7.5 presenta los resultados esenciales del Modelo QIH.

Tabla 7.5.Resultados del Modelo mixtura QIH

Frecuencias observadas					*Probabilidades observada*			
	k1	*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**61**	**15**	**3**	79	**.3720**	**.0915**	**.0183**	.4818
k2	**15**	**26**	**5**	46	**.0915**	**.1585**	**.0305**	.2805
k3	**3**	**5**	**31**	39	**.0183**	**.0305**	**.1890**	.2378
Total	79	46	39	164	.4818	.2805	.2378	1.0000
Prob. condicionales X1					*Prob. condicionales X2*			
k1	**.3171**	**.0000**	**.0000**	.3171	**.0549**	**.0915**	**.0183**	.1647
k2	**.0000**	**.0061**	**.0000**	.0061	**.0915**	**.1524**	**.0305**	.2744
k3	**.0000**	**.0000**	**.1829**	.1829	**.0183**	**.0305**	**.0061**	.0549
Total	.3171	.0661	.1829	***.5061***	.1647	.2744	.0549	***.4939***
Prob. clase latente X1 (ϕ_k) $(\lambda=.5061)$					*Prob. clase latente X2* $(\psi_k^A, \psi_k^B{}')$ $(1-\lambda=.4940)$			
A	**.6265**	**.0120**	**.3614**	1.0000	**.3333**	**.5556**	**.1111**	1.0000
B	**.6265**	**.0120**	**.3614**	1.0000	**.3333**	**.5556**	**.1111**	1.0000

De igual manera, la Tabla 7.6 muestra los parámetros y el ajuste del Modelo mixtura QIC. Obsérvese que el ajuste es similar al del modelo loglineal, pero los parámetros y la interpretación del modelo difieren.

Tabla 7.6. Ajuste del Modelo mixtura QIH

Ajuste Modelo QIH	L^2	*gl residual*	*Probabilidad*	*BIC*
	22.585	3	.000	7.290
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta_k)$	*6.778*	*1.040*	*31.000*	
$\exp(\xi_k)$	*5.778*	*0.040*	*30.000*	
ϕ_k	*.6225*	*.0120*	*.3164*	$\lambda = .5061$
ψ_k^A	*.3333*	*.5556*	*.1111*	$1-\lambda = .4939$
$\psi_{k'}^B$	*.3333*	*.5556*	*.1111*	

Algunas características de interés de este modelo –en comparación con los anteriores– son las siguientes:

1) Respecto del Modelo QI, éste incorpora la **restricción de homogeneidad marginal**, que asume igualdad de las probabilidades marginales en la subtabla de acuerdo aleatorio, lo que implica que $\psi_k^A = \psi_{k'}^B$. En consecuencia de esta restricción, las probabilidades condicionales de la clase *X2* presentan la propiedad de simetría.
2) La probabilidad de acuerdo corregido por azar es la suma de los productos de las probabilidades latentes de la clase *X2* (acuerdo aleatorio y desacuerdo) para ambos evaluadores:

Por consiguiente, es posible estimar también la proporción de acuerdo sistemático aplicando la fórmula general RCA mediante

$$\hat{\lambda} = \frac{p_o - p_e^{\lambda}}{1 - p_e^{\lambda}} = .5061$$

7.2.4 El Modelo mixtura QICH

La formulación del Modelo mixtura QICH, respecto del Modelo mixtura QI, asume en primer lugar que las casillas diagonales son iguales y en segundo lugar que los evaluadores *A* y *B* actúan de forma homogénea . Por lo tanto se restringe el número de parámetros para incluir solamente un parámetro diagonal y un conjunto de parámetros para uno de los jueces. El Modelo QICH (modelo de cuasi-independencia constante y homogéneo) se formula como

$$\log(m_{kk'}) = \lambda + \lambda_k^{A=B} + \xi \qquad \text{(Ec.7.10)}$$

donde ξ es un parámetro diagonal único y $\lambda_k^{A=B}$ es conjunto de *K* parámetros que asumen la homogeneidad entre los observadores *A* y *B*. Estas restricciones son equivalentes a las del coeficiente π de Scott (1955).

La estimación de una medida de acuerdo asociada con el Modelo mixtura QICH se realiza de forma similar a las anteriores y se ha denominado η . Aplicando la Ecuación 7.3, esta medida de acuerdo resulta igual a

$$\eta = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi)+1} \right] = .5707$$

y su complemento es $1-\eta = .5707$. La Tabla 7.7 muestra los resultados esenciales del Modelo mixtura QICH.

Tabla 7.7. Resultados del Modelo mixtura QIH

Frecuencias observadas					*Probabilidades observadas*			
	k1	*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**61.782**	**9.075**	**8.143**	79	**.3767**	**.0553**	**.0497**	.4817
k2	**9.075**	**31.143**	**5.782**	46	**.0553**	**.1899**	**.0353**	.2805
k3	**8.143**	**5.782**	**25.075**	39	**.0497**	**.0353**	**.1529**	.2378
Total	79	46	39	164	.4818	.2805	.2378	1.0000
Prob. condicionales X1					*Prob. condicionales X2*			
k1	**.2988**	**.0000**	**.0000**	.2908	**.0779**	**.0553**	**.0497**	.1829
k2	**.0000**	**.1506**	**.0000**	.1506	**.0553**	**.0393**	**.0353**	.1299
k3	**.0000**	**.0000**	**.1213**	.1213	**.0497**	**.0353**	**.0316**	.1166
Total	.2908	.1506	.1213	***.5707***	.1829	.1299	.1166	***.4293***
Prob. clase latente X1 (ϕ_k) $(\eta = .5707)$					*Prob. clase latente X2* $(\psi_k^A = \psi_{k'}^B)$ $(1-\eta = .4293)$			
A	**.5236**	**.2639**	**.2125**	1.0000	**.4261**	**.3025**	**.2714**	1.0000
B	**.5236**	**.2639**	**.2125**	1.0000	**.4261**	**.3025**	**.2714**	1.0000

Por su parte, la Tabla 7.8 muestra el ajuste del modelo y sus correspondientes parámetros.

Tabla 7.8. Ajuste del Modelo mixtura QICH

Ajuste Modelo QICH	L^2	*gl residual*	*Probabilidad*	*BIC*
	40.059	5	.000	14.560
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta)$				4.8334
$\exp(\xi)$				3.8334
ϕ_k	*.5236*	*.2639*	*.2125*	$\eta = .5707$
ψ_k^A	*.4261*	*.3025*	*.2714*	$1-\eta = .4293$
$\psi_{k'}^B$	*.4261*	*.3025*	*.2714*	

Algunas de las propiedades más destacables de esta propuesta se resumen a continuación.

1) El Modelo mixtura QICH incluye la **restricción de homogeneidad marginal**, por la que los marginales de los evaluadores *A* y *B* se asumen homogéneos o iguales $(\psi_k^A = \psi_{k'}^B)$, presentando la propiedad de simetría. Además incluye la **restricción de constancia diagonal**, que asume razones *odds* iguales para la tabla de probabilidades esperadas, y en concreto $\theta_{12} = \theta_{13} = \theta_{23} = \exp(2\delta) = 23.400$. Por tanto, la constante de proporcionalidad es $CP = \phi_k = \psi_k^2 = 2.900$.

2) En este modelo también puede definirse la probabilidad de acuerdo corregido del azar a través de la suma de los productos de las probabilidades latentes para ambos observadores, los cuales se asumen iguales: $p_e^{\eta} = \sum_{k=1}^{K} \psi_k^2 = (.4261)^2 + (.3025)^2 + (.2714)^2 = .3467.$

Y por ende puede calcularse la probabilidad de acuerdo sistemático mediante la ecuación RCA,

$$\hat{\eta} = \frac{p_o - p_e^{\eta}}{1 - p_e^{\eta}} = .5707$$

siendo su interpretación esencialmente similar a la de modelos anteriores.

7.2.5. El Modelo mixtura QIU

Un interesante modelo, que ha generado mucha controversia, es el Modelo mixtura QIU (modelo mixtura de cuasi-independencia uniforme), que respecto del Modelo básico mixtura QI prescinde de los parámetros del efecto de los jueces *A* y *B*. Formulado como modelo loglineal, se define mediante

$$\log(m_{kk'}) = \lambda + \xi_k \qquad \text{(Ec. 7.11)}$$

La estimación de una medida de acuerdo asociada con el Modelo mixtura QIU produce como resultado una medida equivalente a la medida descriptiva σ de Bennet y otros (1950) y reformulado después por Maxwell (1977) y por Brennan y Prediger (1981); por este motivo, se designa asimismo como σ. Aplicando la Ecuación 7.3, esta medida de acuerdo resulta igual a

$$\sigma = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi)+1} \right] = .5793$$

y su complemento es $1-\sigma = .4207$. La Tabla 7.9 presenta las probabilidades observadas, esperadas, condicionales y latentes para este modelo.

Tabla 7.9. Resultados del Modelo mixtura QIU

Frecuencias observadas					*Probabilidades observadas*			
	k1	*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**61.000**	**7.667**	**7.667**	76.334	**.3720**	**.0467**	**.0467**	.4655
k2	**7.667**	**26.000**	**7.667**	41.334	**.0467**	**.1585**	**.0467**	.2520
k3	**7.667**	**7.667**	**31.000**	46.334	**.0467**	**.0353**	**.1890**	.2825
Total	76.334	41.334	46.334	164.000	.4655	.2520	.2825	1.0000
Prob. condicionales X1					*Prob. condicionales X2*			
k1	**.3252**	**.0000**	**.0000**	.3252	**.0467**	**.0467**	**.0467**	.1402
k2	**.0000**	**.1118**	**.0000**	.1118	**.0467**	**.0467**	**.0467**	.1402
k3	**.0000**	**.0000**	**.1423**	.1423	**.0467**	**.0467**	**.0467**	.1402
Total	.3253	.1118	.1423	***.5793***	.1402	.1402	.1402	***.4207***
Prob. clase latente X1 (ϕ_k) $(\sigma = .5793)$					*Prob. clase latente X2* $(\psi_k^A = \psi_{k'}^B)$ $(1-\sigma = .4207)$			
A	**.5614**	**.1930**	**.2456**	1.0000	**.3333**	**.3333**	**.3333**	1.0000
B	**.5614**	**.1930**	**.2456**	1.0000	**.3333**	**.3333**	**.3333**	1.0000

La Tabla 7.10 presenta el ajuste del Modelo mixtura QIU y sus correspondientes parámetros.

Tabla 7.10. Ajuste del Modelo mixtura QIU

Ajuste Modelo QIU	L^2	*gl residual*	*Probabilidad*	*BIC*
	40.059	5	.000	14.560
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta_k)$	*7.9565*	*3.3913*	*4.0435*	
$\exp(\xi_k)$	*6.9565*	*2.3913*	*3.0435*	
ϕ_k	*.5614*	*.1930*	*.2456*	$\sigma=.5793$
ψ_k^A	*.3333*	*.3333*	*.3333*	$1-\sigma=.4207$
$\psi_{k'}^B$	*.3333*	*.3333*	*.3333*	

Algunas de las características distintivas de este modelo se especifican a continuación.

1) Las dos restricciones básicas son la **restricción de homogeneidad marginal**, por la que los marginales se asumen homogéneos para ambos evaluadores ($\psi_k^A=\psi_{k'}^B$) que presentan la propiedad de simetría; y, en segundo lugar, la **restricción de equiprobabilidad marginal**, en la que los marginales de *A* y de *B* de la subtabla aleatoria de probabilidades condicionales se asumen iguales para todas las categorías. Esta restricción implica igualmente que las probabilidades conjuntas de la subtabla citada sean también iguales. La restricción es pertinente en la práctica para aquellas ocasiones donde puede asumirse que los observadores responden por azar con igual probabilidad a cualquier categoría de la tabla.

2) Con este modelo es posible calcular la probabilidad esperada por azar mediante la suma de los productos de las probabilidades latentes de la subtabla aleatoria para ambos jueces,

$$p_e^{\sigma} = \sum_k \psi_k^2 = .3333$$

y por tanto puede calcularse la probabilidad de acuerdo sistemático aplicando la ecuación general RCA, que coincide exactamente con el valor de la medida descriptiva σ tratadas en capítulos anteriores.

$$\hat{\sigma} = \frac{p_o - p_e^{\sigma}}{1 - p_e^{\sigma}} = .5793$$

7.2.6. El Modelo mixtura QIHX

Además de las restricciones anteriormente abordadas, otra que resulta a veces de interés incorporar es **la homogeneidad de las probabilidades de clase latente**. Esta restricción no puede formularse con un modelo loglineal; sólo es posible mediante un modelo mixtura. Es la que se asume cuando se utiliza el coeficiente κ de Cohen. La Tabla 7.11 expone los resultados fundamentales del modelo.

Aplicando la Ecuación 7.3, la medida de acuerdo equivalente –aunque no exactamente igual– a κ resulta

$$\kappa = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi)+1} \right] = .5590$$

siendo su complemento es $1-\kappa = .4410$.

Tabla 7.11. Resultados del Modelo mixtura QIHX

Frecuencias observadas					*Probabilidades observadas*			
k1		*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**61.910**	**10.460**	**7.597**	78968	**.3714**	**.0638**	**.0463**	.4815
k2	**10.460**	**34.059**	**4.739**	49.258	**.0638**	**.2077**	**.0289**	.3004
k3	**7.597**	**4.739**	**23.439**	35.439	**.0463**	**.0289**	**.1429**	.2180
Total	78.968	49.258	35.775	164.000	.4655	.3004	.2180	1.0000
Prob. condicionales X1					*Prob. condicionales X2*			
k1	**.2692**	**.0000**	**.0000**	.2692	**.1022**	**.0638**	**.0463**	.2123
k2	**.0000**	**.1679**	**.0000**	.1679	**.0638**	**.0398**	**.0289**	.1325
k3	**.0000**	**.0000**	**.1219**	.1219	**.0463**	**.0289**	**.0210**	.0962
Total	.2692	.1679	.1219	***.5590***	.2123	.1325	.0962	***.4410***
Prob. clase latente X1 (ϕ_k) $(\kappa = .5590)$					*Prob. clase latente X2* $(\psi_k^A = \psi_{k'}^B)$ $(1-\kappa = .4410)$			
A	**.4815**	**.3004**	**.2181**	1.0000	**.4815**	**.3004**	**.2181**	1.0000
B	**.4815**	**.3004**	**.2181**	1.0000	**.4815**	**.3004**	**.2181**	1.0000

El ajuste del Modelo mixtura QIHX y sus parámetros esenciales se muestran en la Tabla 7.12.

Tabla 7.12 Ajuste del Modelo mixtura QIHX

Ajuste Modelo QIU	L^2	*gl residual*	*Probabilidad*	*BIC*
	37.611	5	.000	12.112
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta_k)$	*2.6324*	*4.2201*	*5.8106*	
$\exp(\xi_k)$	*1.6324*	*3.2201*	*4.8116*	
ϕ_k	*.4815*	*.3004*	*.2181*	$\kappa = .5590$
ψ_k^A	*.4815*	*.3004*	*.2181*	$1-\kappa = .4410$
$\psi_{k'}^B$	*.4815*	*.3004*	*.2181*	

Las propiedades de este modelo enlazan con las restricciones que se asumen cuando se utilizan coeficientes descriptivos tales como la κ de Cohen (1960); así lo han demostrado Guggenmoos-Holtzman y Vonk (1998). Entre las propiedades del Modelo mixture QIHX destacamos:

1) Las dos restricciones básicas que contiene son la **restricción de homogeneidad marginal**, por la cual las probabilidades marginales de ambos jueces se asumen iguales y presentan la propiedad de simetría, y la **restricción de homogeneidad de las probabilidades de clase latente,** en la que los marginales de *A* y *B* se asumen iguales para las dos subtablas (acuerdo sistemático y acuerdo aleatorio/desacuerdo), y por tanto $\phi_k = \psi_k^A = \psi_{k'}^B$.

2) Sin embargo, debido a la restricción que implica a las probabilidades latentes de ambas subtablas, con este modelo no es posible obtener un

índice de acuerdo utilizando la fórmula general RCA, ya que $\sum \psi_k^A \psi_{k'}^B \neq .5590$. No obstante, la ventaja de utilizar este modelo se fundamenta sobre la constatación de que las restricciones utilizadas no son compatibles con los datos empíricos utilizados en el Ejemplo 6.1; así pues, no sería correcto emplear un índice tipo *kappa* con tales datos.

7.2.7. El Modelo mixtura QICU

De la familia de modelos de cuasi-independencia que se han utilizado hasta este capítulo, sólo se ajusta el modelo básico (Modelo mixtura QI).

Las restricciones que se han empleado con los modelos sometidos a prueba son un conjunto cerrado de **restricciones de homogeneidad** que cabe asumir con este tipo de modelos, tales como la constancia diagonal, la homogeneidad marginal, la uniformidad marginal y la homogeneidad de las probabilidades latentes. Asimismo, las restricciones de homogeneidad se pueden combinar con otras relativas a la asociación entre variables, por ello llamadas **restricciones de asociación**, usualmente entre jueces (véase Agresti, 1992, 2002). Cabe además cualquier combinación de ambos tipos de restricciones para componer un modelo determinado.

1) El Modelo mixtura QICU es simplemente una de las muchas combinaciones posibles compuesta de un Modelo mixtura QIC con una restricción de asociación uniforme. Aplicando la Ecuación 7.3, la

medida de acuerdo resultante –que llamaremos genéricamente ϵ – de aplicar el Modelo mixtura QICU es

$$\epsilon = \sum_{k=1}^{K} \left[p_{kk} - \frac{p_{kk}}{\exp(\xi)+1} \right] = .4833$$

y su complemento es $1-\epsilon = .5167$. La Tabla 7.13 expone las probabilidades observadas, esperadas, condicionales y latentes para este modelo.

Tabla 7.13. Resultados del Modelo mixtura QICU

Frecuencias observadas					*Probabilidades observadas*			
	k1	*k2*	*k3*	Total	*k1*	*k2*	*k3*	Total
k1	**61**	**26**	**5**	92	**.3720**	**.1585**	**.0305**	.5610
k2	**4**	**26**	**3**	33	**.0244**	**.1585**	**.0183**	.2012
k3	**1**	**7**	**31**	39	**.0061**	**.0427**	**.1890**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Frecuencias esperadas					*Probabilidades esperadas*			
k1	**62.121**	**24.948**	**4.932**	92	**.3788**	**.1521**	**.0301**	.5610
k2	**2.811**	**26.000**	**4.189**	33	**.0172**	**.1585**	**.0255**	.2012
k3	**1.068**	**8.052**	**29.879**	39	**.0065**	**.0491**	**.1822**	.2378
Total	66	59	39	164	.4025	.3597	.2378	1.000
Prob. condicionales X1					*P* *Prob. condicionales X2*			
k1	**.2544**	**.000**	**.000**	.2544	**.1244**	**.1521**	**.0301**	.3066
k2	**.000**	**.1065**	**.000**	.1065	**.0171**	**.0520**	**.0255**	.0946
k3	**.000**	**.000**	**.1224**	.1224	**.0065**	**.0491**	**.0598**	.1194
Total	.2544	.1065	.1224	1.000	.1480	.2532	.1154	1.000
Prob. clase latente X1 (Φ_k) $(\epsilon = .4833)$					*Prob. clase latente X2* $(\psi_k^A = \psi_{k'}^B)$ $(1-\epsilon = .5167)$			
A	**.5264**	**.2203**	**.2532**	1.000	**.5933**	**.1833**	**.2234**	1.000
B	**.5264**	**.2203**	**.2532**	1.000	**.2865**	**.4902**	**.2234**	1.000

El ajuste del Modelo mixtura QICU se muestra a continuación en la Tabla 7.14. Obsérvese que se ajusta óptimamente y su BIC es inferior al deparado por el Modelo mixtura básico QI, por lo que resulta ser el mejor modelo de la familia.

Tabla 7.14. Ajuste del Modelo mixtura QICU

Ajuste Modelo QIU	L^2	*gl residual*	*Probabilidad*	*BIC*
	1.074	2	.585	-9.130
Parámetros	*positiva*	*neutral*	*negativa*	
$\exp(\delta_k)$				3.046
$\exp(\xi_k)$				2.046
$\exp(\theta_k)$				2.482
ϕ_k	*.5264*	*.2203*	*.2532*	$\epsilon = .4833$
ψ_k^A	*.5933*	*.2833*	*.2234*	$1-\epsilon = .5167$
$\psi_{k'}^B$	*.2865*	*.4902*	*.2234*	

Es interesante destacar algunas propiedades características de este modelo:

1) Además de la restricción de constancia diagonal, el Modelo mixtura QICU incluye la restricción de asociación uniforme, que implica que las razones *odds* locales –es decir, las que implican categorías contiguas– para la subtabla de acuerdo aleatorio y desacuerdo son homogéneas e iguales al parámetro de asociación uniforme θ; que es $\exp(\theta_{12}) = \exp(\theta_{23}) = 2.482$ en escala exponencial. Recuérdese que en el Modelo mixtura QIC la restricción de constancia diagonal implicaba que todas las razones *odds* diagonales eran iguales a la unidad $\exp(\theta_{12}) = \exp(\theta_{13}) = \exp(\theta_{23}) = 1.$.

2) Sin embargo, debido a la combinación de ambos tipos de restricción, con este modelo tampoco es posible obtener un índice de acuerdo

utilizando la fórmula general RCA, ya que $\sum \psi_k^A \psi_{k'}^B \neq .4833$. El modelo resultante alberga no obstante una interesante interpretación, que parte de las propiedades del Modelo mixtura QIC, sobre el que se contraponen las propiedades de la asociación uniforme. El excelente ajuste del modelo implica que las restricciones utilizadas son perfectamente compatibles con los datos empíricos de este ejemplo.

Capítulo 8

Estimación del sesgo entre jueces

8.1. Introducción

El **efecto de sesgo** de un observador respecto a otro (u otros) ocurre cuando sus promedios (si la medida es numérica) o sus probabilidades marginales (si la medida es categórica) difieren debido a su interpretación diferencial de la escala de valoración o bien a su percepción única y divergente de los ítemes. En general es mayor conforme aumenta la diferencia entre medias o la heterogeneidad de sus respectivas distribuciones marginales. En cierta medida vinculado al efecto de sesgo se encuentra el **efecto de prevalencia**, que ocurre en presencia de una proporción global extrema de resultados para una

determinada respuesta o categoría de respuesta y, en la práctica, representa la proporción de casos positivos de la población. Ambos efectos se han demostrado en numerosos trabajos (Spitznagel y Helzer, 1985; Feinstein y Cicchetti, 1990; Byrt, Bishop y Carlin 1993; Agresti, Ghosh y Bini, 1995; Lantz y Nebenzahl, 1996 y Hoehler 2000).

Dos jueces u observadores interpretan un determinado ítem de una escala de forma diferente o tienen reacciones específicas a determinados estímulos de tal modo que las valoraciones obtenidas reflejan en cierta medida características de los jueces, además de determinadas características de los estímulos. Aunque para muchos investigadores el análisis del sesgo carece de interés substantivo, y por tanto contribuye al error de medida, el impacto del sesgo en los resultados de un estudio puede ser importante en una investigación y depende de numerosos factores, incluyendo la naturaleza del constructo que se evalúa, el grado en el que los jueces son entrenados para interpretar de forma similar el significado de los estímulos y otras características típicas del diseño del estudio. Como consecuencia de esta complejidad, puede ser difícil para los investigadores saber si el sesgo es probablemente un problema de sus datos y cuál es el impacto de esta fuente de error en sus resultados.

El análisis del sesgo entre jueces depende también del nivel de medida de las variables. Si la medida es numérica, entonces es común utilizar la teoría de la generalizabilidad y los componentes de la varianza para analizar el sesgo (Hoyt, 2000). Si la medida es categórica, en cambio, es usual utilizar alguno de una serie de procedimientos descriptivos basados en la hipótesis de la homogeneidad o la simetría.

8.2. Detección del sesgo con datos numéricos

Cuando varios jueces valoran varios ítemes en alguna característica psicológica, el sesgo de la valoración puede afectar (1) a la media de las valoraciones, (2) a la varianza de las valoraciones o (3) a la covarianza de las valoraciones con las de otra característica de los mismos jueces (Hoyt, 2000). Un ejemplo típico es la valoración que se hace de los exámenes de alguna asignatura por los estudiantes universitarios. Si algunos profesores consideran que 5 es una puntuación baja (a pesar de tratarse de una puntuación aceptable) mientras otros no, las puntuaciones de los primeros serán sesgadas y sus alumnos estarán por encima del promedio y el resultado es un **sesgo específico del juez**. En cambio, si algunos profesores se dejan influir por atributos de los alumnos (atractivo físico, participación en clase, etc.), no relacionados con la ejecución, para valorar a los alumnos, las puntuaciones están sesgadas debido a las diferentes impresiones que los profesores tienen de cada alumno y el resultado es un **sesgo específico de la díada.** Mientras que el primero es fácil de estimar y de corregir (al comparar diferentes profesores), porque es general para los profesores implicados, el segundo es aún más complejo de estimación y de corrección, ya que el sesgo de un determinado juez es específico del alumno objeto de evaluación.

La teoría de la generalizabilidad (Cronbach y otros, 1972; Shavelson y Webb, 1991; Brennan, 2001a) es un instrumento analítico de extrema utilidad para estudiar el sesgo entre observadores, porque permite el examen simultáneo del impacto de múltiples fuentes de error (y sus interacciones) sobre los datos. Puesto que el análisis se centra en la estimación de la varianza explicada por efectos del modelo (en lugar de probar sus significación

estadística), produce información útil sobre la importancia relativa de varias fuentes de error y su impacto sobre la calidad de la valoración. Así, cuando se valora una determinada variable, el modelo univariante que se asume para la varianza de las valoraciones es el siguiente (Lakes y Hoyt, 2008):

$$\sigma^2_{IJ} = \sigma^2_I + \sigma^2_J + \sigma^2_D + \sigma^2_R \qquad \text{(Ec. 8.1)}$$

donde la varianza total de la valoración del ítem *I* por el juez *J* (σ^2_{IJ}) es la suma de un conjunto de componentes de varianza,

- para los efectos de Ítem (σ^2_I), que se definen como la desviación de la media (promediando el conjunto de jueces) de la media global (promediando todos los ítemes y jueces). El efecto refleja cómo se percibe consensuadamente el estímulo;

- para los efectos de Juez (σ^2_J), que es la desviación de la media de los jueces (promediando el conjunto de los ítemes) de la media global. El efecto refleja en qué medida los jueces difieren en sus percepciones generalizadas de los estímulos;

- para los efectos de Díada (σ^2_D), que aparece cuando el juez *j* valora el ítem *i* por encima o por debajo de lo que se pronosticaría conociendo el efecto de tal juez y de tal ítem. Cuando solo hay una observación en cada par juez-ítem, los efectos de díada se confunden con la varianza de error. Es decir, el efecto de díada no es estimable si las observaciones

no son replicadas.

- un efecto residual o de error (σ_E^2), que es la varianza residual una vez eliminados los restantes efectos.

Todos los efectos se asumen ortogonales, y por tanto la covarianza entre efectos es nula. La magnitud de los efectos de juez y de díada indican en qué grado dos observadore se espera que sistemáticamente desacuerden en sus valoración del mismo ítem y por tanto se entiende que es varianza debida a **errores entre jueces** (*interrater errors*); por el contrario, el término de error residual representa la varianza debida a **errores dentro de jueces** (*intrarater errors*), o sea, varianza no atribuíble a ítemes o a errores sistemáticos (replicables) por parte de los jueces.

8.3 Detección del sesgo con datos categóricos

Una situación diferente se enfrenta cuando se desea detectar el sesgo con medidas categóricas, particularmente las que proceden de una escala nominal. Entre los métodos más utilizados para detectar el sesgo entre observadores destacan aquellos que se basan en probar si se cumplen las hipótesis de homogeneidad marginal y de simetría. La razón que justifica este hecho radica en la propia definición de sesgo. Como se definió anteriormente, el sesgo de un observador se valora respecto de otro observador y se refiere a las discrepancias entre sus distribuciones marginales, por lo que disminuye en la medida que las distribuciones marginales se hacen equivalentes. La ausencia de

sesgo implica que $p_{i+} = p_{+i}$ para todo i (Agresti, 2002).

La mayoría de los métodos para detectar el sesgo (al igual que los empleados para evaluarlo) se han definido para tablas de dimensión 2×2. En este trabajo se tratan también algunos métodos usualmente utilizados para tablas de mayor dimensión, que son en general una extensión de los anteriores.

8.3.1 Ejemplo 8.1

Con el objetivo de ilustrar los procedimientos para la detección y medida del sesgo entre jueces nos serviremos de un ejemplo de Dillon y Mullani (1984) en el que dos observadores registraron un conjunto de 164 respuestas cognitivas elicitadas en un estudio de comunicación persuasiva sobre una escala de $K = 3$ categorías de respuesta ("positiva", "neutral" y "negativa") Este mismo ejemplo fue tratado en los Capítulos 6 (véase Tabla 6.1) y 7 (véase Tabla 7.1) lo empleamos de nuevo aquí con propósitos comparativos. La Tabla 8.1 presenta de nuevo los datos empíricos. Nótese la igualdad de la diferencia en valor absoluto de los marginales de fila y columna para la categoría "positiva" (92 – 66 = 26) y para la categoría "neutral" (59 – 33 = 26), y la no diferencia para la categoría "negativa"(39 – 39 = 0).

Tabla 8.1. Frecuencias y sus probabilidades del Ejemplo 8.1

Juez A	*Juez B*			
	positiva	*neutral*	*negativa*	Marginales
positiva	**61**(.372)	**26**(.159)	**5**(.030)	92 (.561)
neutral	**4** (.025)	**26**(.159)	**3**(.018)	33 (.201)
negativa	**1** (.006)	**7**(.043)	**31** (.189)	39(.238)
Marginales	66(.402)	59(.360)	39(.238)	164(1.000)

Para ilustrar algunos de los índices tratados más adelante, es conveniente colapsar esta tabla de acuerdo 3x3 en una tabla 2x2 más simple, por ejemplo para comparar la categoría "positiva" con las restantes. El resultado es la Tabla 8.2 siguiente.

Tabla 8.2. Frecuencias y probabilidades del Ejemplo 8.1 (2x2)

Juez A	*Juez B*		
	positiva	*otras*	Marginales
positiva	**61** (.372)	**31** (.189)	92 (.561)
otras	**5** (.030)	**67** (.409)	72 (.439)
Marginales	66 (.402)	98(.598)	164(1.000)

8.3.2. Procedimientos para detectar y probar el sesgo en el enfoque clásico

Dada una Tabla de acuerdo como la Tabla 8.1, donde los jueces clasifican a los sujetos o ítemes según una variable de interés con *K* niveles, para evaluar si las valoraciones de los jueces u observadores son iguales o divergentes se pueden aplicar pruebas que se basan en una distribución χ^2. Un resultado significativo implicaría que las frecuencias o probabilidades marginales no son homogéneas. Una exhaustiva revisión de la literatura (Benavente, Ato y López, 2006) acerca de las pruebas estadísticas utilizadas para detectar el sesgo entre observadores nos ha permitido destacar las pruebas siguientes:

- **La prueba binomial exacta** (Siegel y Castellan, 1988), se obtiene calculando primero la proporción

$$P = \frac{n_{12}}{n_{12} + n_{21}} \qquad \text{(Ec. 8.1)}$$

con el objeto de probar si $P = .500$. Las hipótesis estadísticas que se plantean son pues $H_0: P = (1 - P)$ y $H_1: P \neq (1 - P)$. Para los datos de la Tabla 8.2, $P = .861$ y $1 - P = .139$ y la hipótesis nula se rechaza con $P < .001$.

- **La prueba de McNemar** (McNemar, 1947) para tablas 2×2 utiliza una distribución de χ^2 y se calcula utilizando

$$X^2_{gl} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \qquad \text{(Ec. 8.2)}$$

Algunos autores recomiendan una versión de la prueba de McNemar con una corrección de la continuidad cuando los valores de n_{12} y/o n_{21} son pequeños (por ejemplo, $n_{12} + n_{21} < 10$), calculado con la fórmula:

$$X^2_{gl} = \frac{[(n_{[12]} - n_{21}) - 1]^2}{(n_{12} + n_{21})} \qquad \text{(Ec. 8.3)}$$

que de igual modo se rechaza la hipótesis nula con $P < .001$. Para tablas de acuerdo de mayor dimensión $(K \times K)$, una forma sencilla de calcular el sesgo a través de la prueba de McNemar se describe en Bishop, Fienberg y Holland (1975) y consiste en aplicar la Ecuación 8.1 pero definiendo como cantidades básicas las siguientes: n_S es igual a la suma de las frecuencias de las casillas del triángulo superior (las que se encuentran por encima de la diagonal principal), y n_I es igual a la suma de las frecuencias de las casillas del triángulo inferior (las que se encuentran por debajo de la diagonal principal). Para una tabla 3×3 como la Tabla 8.1, por ejemplo, $n_S = n_{12} + n_{13} + n_{23} = 34$ y $n_I = n_{21} + n_{31} + n_{32} = 12$, y la aplicación de la Ecuación 8.1 produce una $P = .739$ con probabilidad $P = .002$. Puesto que resulta significativa se puede afirmar que no existe homogeneidad marginal.

- La **prueba de Bowker** (Bowker, 1948; Bishop, Fienberg y Holland, 1975; Krampe y Kuhnt, 2007) para una tabla cuadrada es una extensión de la de McNemar y consiste en probar la hipótesis de simetría mediante $H_0: p_{ij} = p_{ji}$ y $H_1: p_{ij} \neq p_{ji}$. Se distribuye según $\chi^2_{K(K-1)/2}$ y viene dado por la siguiente ecuación (con la corrección de la continuidad):

$$X^2_{K(K-1)/2} = \frac{\sum [(n_{ij} - n_{ji}) - 1]^2}{\sum (n_{ij} + n_{ji})} \qquad \text{(Ec. 8.4)}$$

Al aplicar la Ecuación. 8.4 a los datos de la Tabla 8.2 observamos que X_3^2 = 9.978; P = .018, y por tanto cabe también concluir que existe evidencia clara de sesgo entre jueces.

- La **prueba de Stuart - Maxwell** (Stuart, 1955; Maxwell, 1961 y Everitt, 1992) verifica si existe homogeneidad marginal en tablas de dimensión $K \times K$ para todas las categorías de forma simultánea. Se interpreta como una χ^2 con $K-1$ grados de libertad. Para tablas 2×2, los resultados obtenidos con la prueba de Stuart - Maxwell y la prueba de McNemar son idénticos. El cálculo es algo complejo, basado en álgebra de matrices, pero puede obtenerse una aproximación aceptable para tablas de acuerdo pequeñas. Por ejemplo, para una tabla 3×3 la aproximación es la siguiente (Everitt, 1992):

$$X^2_{K-1} = \frac{\bar{n}_{23} d_1^2 + \bar{n}_{13} d_2^2 + \bar{n}_{12} d_3^2}{2(\bar{n}_{12}\bar{n}_{23} + \bar{n}_{12}\bar{n}_{13} + \bar{n}_{13}\bar{n}_{23})} \qquad \text{(Ec. 8.5)}$$

donde $\bar{n}_{ij}$ es el promedio de las casillas simétricas, o sea, $n_{ij}=1/2\,(n_{ij}+n_{ji})$ y d_i es la diferencia entre los marginales de fila y columna, esto es, $d_i=(n_{i+}-n_{+i})$. Para los datos de la Tabla 8.2, la prueba Stuart -Maxwell adopta una valor de X_2^2 = 20.030, P < .001. Se concluye por tanto de nuevo que los datos empíricos no son compatibles con la hipótesis de homogeneidad marginal.

Existen otros procedimientos específicos desarrollados para medir el sesgo, que alcanzan un valor concreto pero, con excepción del índice de sesgo de Ludbrock (2004), no informan acerca de su significación estadística. Estos índices se han desarrollado sólo para el enfoque clásico (o descriptivo). Una revisión de la literatura nos ha permitido destacar los siguientes:

- El **índice de simetría en el desacuerdo** (*symmetry in disagreement index*) para tablas 2×2 (Lanz y Nebenzahl, 1996) se calcula del modo siguiente:

$$S_D=\frac{(n_{12}-n_{21})}{N} \qquad \text{(Ec. 8.6)}$$

  y se distribuye según χ^2 con un grado de libertad. Puede adoptar valores desde -1 a +1. Sin embargo, no existe una prueba estadística asociada a este índice.

- El **índice de sesgo o BI** (*bias index*) para tablas 2×2 fue propuesto por Byrt, Bishop y Carlin (1993), presenta un rango de valores que va

de 0 a +1 y se obtiene mediante:

$$BI_{BBC} = \frac{|n_{12} - n_{21}|}{N} \qquad \text{(Ec. 8.7)}$$

Ludbrook (2004) propuso además una forma de evaluar indirectamente el índice BI basado en el uso de una prueba exacta no condicional sobre las diferencias entre proporciones, en la que la hipótesis nula es $H_0 = p_1 - p_2 = 0$. Las proporciones binomiales son n_{12}/N y n_{21}/N. Las frecuencias binomiales correspondientes son $n_{12}/(N - n_{12})$ y $n_{21}/(N - n_{21})$ respectivamente.

- Ludbrook (2002) extendió la evaluación del índice BI_{BBC} a tablas de dimensión $K > 2$ aplicando la Ecuación 8.7, definiendo n_S como la suma de todas las casillas que hay por encima de la diagonal principal ($\sum S$: sumatorio de todos los elementos del triángulo superior) y n_I como la suma de todos los elementos (frecuencias) que hay por debajo de la diagonal principal ($\sum I$: suma de todos los elementos del triángulo inferior). Las frecuencias binomiales correspondientes son $\sum S/(N_{\sum S})$ y $\sum I/(N_{\sum I})$, siguiendo las recomendaciones del texto clásico de Bishop, Fienberg y Holland (1975). Al aplicar la Ecuación 8.2 a los datos de la Tabla 8.2 junto con el método propuesto por Ludbrook para tablas 3×3 (que se ha complementado con el procedimiento de comparación múltiple de Holm, como se documenta

en el trabajo citado), obtenemos un índice de sesgo de $BI_L = 0.134$ con probabilidad $P = .00084$, lo que implica diferencias estadísticamente significativas entre proporciones y por tanto se asume la existencia de sesgo entre observadores.

- **El índice PABAK** (*Prevalence and Bias Adjusted Kappa*)**,** propuesto asímismo por Byrt, Bishop y Carlin, 1993, produce un valor de *kappa* corregido de sesgo y prevalencia para tablas de dimensión 2×2, tal y como se comentó en la sección 2.4 de este trabajo. Esencialmente toma el valor empírico de kappa y calcula una kappa 'equivalente' con una población con 50/50 de prevalencia y ausencia de sesgo (homogeneidad marginal) . El índice se obtiene aplicando la siguiente fórmula:

$$PABAK = 2P_o - 1 = \kappa(1 - PI^2 + BI^2) + PI^2 - BI^2 \qquad \text{(Ec. 8.8)}$$

donde BI = $n_{12} - n_{21}$ y PI = $n_{11} - n_{22}$. Los valores del índice PABAK varían de -1 a +1, igual que el índice *kappa*. La diferencia entre los valores reportados por los índices *kappa* y PABAK nos aporta un valor de sesgo (en este caso también controlando la prevalencia). Sin embargo, un importante inconveniente es que sólo se puede aplicar a tablas de dimensión 2×2.

- Arstein y Poesio (2005) proponen para tablas de cualquier dimensión un **índice de sesgo diferencial**, basado en la diferencia entre el acuerdo esperado por azar para el índice π (Scott, 1955) y el acuerdo esperado por azar para el índice κ (Cohen, 1960) tal como se muestra

a continuación:

$$B = \sum_i p_{ei}^{\pi} - \sum_i p_{ei}^{\kappa} \qquad \text{(Ec. 8.9)}$$

basándose en el hecho de que el primero asume homogeneidad marginal mientras que el segundo no. Al aplicar la Ecuación 8.9 a los datos de Dillon y Mullani (Tabla 8.2) se obtiene una diferencia de B = = .0164.

En consecuencia, puede concluirse que, de los índices descriptivos propuestos en la literatura, el índice de sesgo entre jueces más apropiado para datos categóricos parece ser el índice de sesgo BI de Ludbrock (2002), que es válido para tablas de cualquier dimensión y permite utilizar una prueba estadística para probar su significación.

8.3.3. Procedimientos para detectar y probar el sesgo en el enfoque del modelado loglineal

Algunos de los procedimientos basados en el enfoque clásico utilizan pruebas estadísticas de hipótesis nula mediante el contraste de hipótesis estadísticas. Una de las alternativas actualmente más consistentes es el enfoque del modelado estadístico, donde el concepto de modelo pasa a jugar un papel primordial (Ato y otros, 2005; Ato y Vallejo, 2007). Para detectar el sesgo mediante modelado estadístico se aplican modelos loglineales para probar si el

modelo de homogeneidad marginal es consistente con los datos empíricos.

Hay dos formas básicas de probar el modelo de homogeneidad marginal, una forma indirecta y otra directa. La **forma indirecta** se basa en una estrategia de ajuste condicional, donde asumiendo que se cumple el modelo de cuasi-simetría, el modelo de homogeneidad marginal es equivalente al modelo de simetría (Causinus, 1965):

Cuasi-Simetría (QS) + Homogeneidad marginal (HM) = Simetría (S)(Ec. 8.10)

Al aplicar la Ecuación. 8.10 a los datos de la Tabla 8.1 se obtiene para el modelo de simetría una desvianza de $L^2(5)=22.585; P=.000$, y para el modelo de cuasi-simetría una desvianza de $L^2(7)=0.182; P=.6693$. Puesto que el modelo de cuasi-simetría se ajusta aceptablemente, la diferencia entre los modelos de simetría y cuasi-simetría nos proporciona una prueba aproximada del modelo de homogeneidad marginal, que en este caso alcanza una desvianza residual de $L^2(2)=22.403(2); P=.000$, lo que evidencia un alto grado de desajuste, y por tanto se concluye que los datos empíricos no son congruentes con el modelo de homogeneidad marginal.

Una **forma directa** de probar el modelo de homogeneidad marginal es a través del ajuste propio del modelo. Esta prueba es compleja, puesto que implica la aplicación de **modelos marginales** (véase Bergsma, 1998; Vermunt, Rodrigo y Ato, 2001), pero puede obtenerse utilizando una versión experimental del programa LEM (Vermunt, 1997). Tras aplicarlo a los datos de la Tabla 8.1 se obtiene una desvianza de $L^2(2)=22.081; P=.000$, lo que

conduce de nuevo a la misma conclusión anterior de que los datos empíricos no son consistentes con el modelo. Nótese además que la diferencia entre ambas pruebas es prácticamente insignificante, aunque en ciertos casos puede producir conclusiones estadísticas diferentes.

8.3.4. Procedimientos para detectar y probar el sesgo en el enfoque de modelado mixtura

Varias familias de modelos de acuerdo pueden utilizarse en un enfoque mixtura con una variable latente con dos clases. La familia de modelos de cuasi-independencia (QI), tratada por Schuster y Smith (2002) y Ato, Benavente y López (2006) está integrada entre otros por los seis modelos siguientes tratados en el capítulo anterior: Modelo básico (QI), Modelo QI constante (QIC), Modelo QI homogéneo (QIH), Modelo QI constante y homogéneo (QICH), Modelo QI uniforme (QIU) y Modelo QI homogéneo en las clases latentes (QIHX). Cada uno de ellos se caracteriza por satisfacer un conjunto de restricciones y se identifica con medidas de acuerdo ya existentes en la literatura. Así, el Modelo QI se corresponde con la medida de acuerdo Δ propuesta por Martin y Femia (2004), aunque derivada desde otro planteamiento diferente, el Modelo QIC con la medida de acuerdo propuesta por Tanner y Young (1985) con modelos loglineales y con la medida α propuesta por Aickin (1990),el Modelo QIH se corresponde con el modelo de observadores homogéneos de Schuster y Smith (2002) y con la medida λ (Ato, Benavente y López, 2006), el Modelo QICH es equivalente al índice

descriptivo π de Scott (1955), el Modelo QIU es similar al índice descriptivo σ de Bennet y otros (1954) y el Modelo QIHX se corresponde en parte con el índice descriptivo κ de Cohen (1960). El ajuste de tales modelos, con sus restricciones características y las probabilidades de las dos clases latentes (*X1*, para acuerdo sistemático y *X2* para acuerdo aleatorio y desacuerdo), para los datos empíricos de la Tabla 8.1 se obtuvo con una versión experimental del programa LEM (Vermunt, 1997) y se presentó en las correspondientes tablas del capítulo anterior. El flujo completo del programa LEM utilizado se muestra en la Salida 8.1 al final de este capítulo. La Tabla 8.3 muestra los resultados del ajuste de los modelos citados con los datos de Dillon y Mullani (1984).

Tabla 8.3.
Ajuste de modelos mixtura con la familia de Modelos QI utilizando una variable latente (X) con 2 clases

Modelo	*Proporción latente X1*	*Proporción latente X2*	*Ajuste del modelo*
QI	.567	.433	$L_1^2 = .1800 ; P > .100$
QIC	.620	.380	$L_3^2 = 10.100 ; P < .100$
QIH	.506	.494	$L_3^2 = 22.600 ; P < .100$
QICH	.571	.429	$L_5^2 = 40.100 ; P < .100$
QIU	.579	.421	$L_5^2 = 43.000 ; P < .100$
QIHX	.360	.640	$L_5^2 = 36.900 ; P < .100$

Nota: los asteriscos representan los modelos aceptablemente ajustados.

En consecuencia, no existe una medida específica para medir el sesgo en los modelos mixtura, más allá de la prueba de homogeneidad marginal

comentada para los modelos loglineales. Sin embargo, una ampliación del número de variables latentes permitiría una interpretación más racional de la descomposición del acuerdo y el desacuerdo, y desarrollar además una medida más apropiada del sesgo entre jueces basada en modelos mixtura. Téngase en cuenta que, de las dos clases latentes que se utilizan en un modelo mixtura básico, la interpretación de la primera (*X1*), que representa la proporción de acuerdo sistemático es directa e inambigua y no plantea problema alguno. En cambio, la interpretación de la segunda clase (*X2*), que representa la proporción de acuerdo aleatorio y desacuerdo es sin embargo ambigua y en el fondo apenas tiene utilidad alguna porque mezcla componentes de distinta naturaleza.

Una solución que se propuso para interpretar la descomposición del acuerdo y del desacuerdo de forma más lógica (Ato, López y Benavente, 2008) consistió en descomponer la clase *X2* del modelo mixtura básico para aislar los componentes de acuerdo aleatorio y de desacuerdo. Tres condiciones deberían cumplirse para realizar adecuadamente tal descomposición: en primer lugar, el resultado no debe afectar al ajuste de los modelos de la familia; en segundo lugar, no debe ser afectada la proporción de acuerdo sistemático de la clase *X1* y en tercer lugar, la suma de las proporciones de acuerdo sistemático y aleatorio debe ser igual a la proporción total de acuerdo observado. Para preservar el equilibrio de los pesos utilizados, se requiere ampliar el modelo original incluyendo dos variables latentes con dos clases cada una (en lugar de una variable latente con tres clases) y seleccionar los pesos cuidadosamente con el objeto de lograr una solución satisfactoria. El flujo del programa requerido para obtener la solución se muestra en la Salida 8.2 al final de este mismo capítulo. La primera variable latente (*X*) evalúa el acuerdo (donde la

primera clase, *X1*, representa el acuerdo sistemático y la segunda clase, *X2*, el acuerdo aleatorio) y la segunda variable latente (*Y*) evalúa el desacuerdo. Puesto que nos interesaba analizar el desacuerdo desde la perspectiva del sesgo entre jueces, se optó por definir las dos clases de Y ponderando la magnitud de las frecuencias en los triángulos superior (clase *Y1*) e inferior (clase *Y2*) con la finalidad de promover el desarrollo de una medida de sesgo. La Tabla 8.4 resume el ajuste de la familia de Modelos QI para dos variables latentes con dos clases cada una que se obtuvo con el programa LEM utilizando los datos de la Tabla 8.1. Solo se ajustó aceptablemente el Modelo QI, y por tanto es el único que es susceptible de interpretación. Nótese la notable similaridad aparente en los flujos de las Salidas 8.1 y 8.2 que se muestran al final de este capítulo, similaridad que se manifiesta también en los resultados del ajuste del Modelo QI con una o dos variables latentes.

Tabla 8.4. Ajuste de modelos mixtura con la familia de Modelos QI utilizando 2 variables latentes con 2 clases cada una

Modelo	*Proporción C.latente X1Y1*	*Proporción C.latente X1Y2*	*Proporción C.latente X2Y1*	*Proporción C. latente X2Y2*	*Ajuste*
QI	.567	.153	.209	.071	L_1^2=.180 *
QIC	.620	.099	.200	.081	
QIH	.506	.213	.140	.140	
QICH	.571	,149	.140	.140	
QIU	.579	.140	.140	.140	
QIHX	.360	.360	.140	.140	

Nota: los asteriscos representan los modelos aceptablemente ajustados.

Con esta ampliación a cuatro clases, la primera clase (combinación *X1Y1*)

representa la proporción de acuerdo sistemático, que corresponde a una subpoblación de ítemes en la que los jueces clasifican del mismo modo los objetos para los que no existe duda alguna. La segunda clase (combinación *X1Y2*) representa la proporción de acuerdo aleatorio en la que los observadores coinciden por azar en la clasificación de los objetos dudosos. La tercera clase (combinación *X2Y1*) representa la proporción de desacuerdo esperado en las casillas del triángulo superior y la cuarta clase (combinación *X2Y2*) representa la proporción de desacuerdo esperado en las casillas del triángulo inferior. Nótese, en comparación con la Tabla 8.3, que, en cada uno de los modelos de la familia QI, la clase latente de acuerdo sistemático (*X1Y1*) produce el mismo resultado que la clase *X1* de la Tabla 8.3 y que el ajuste es también exactamente el mismo que se obtuvo con la familia de modelos definida para una variable latente con dos clases, ya que se trata esencialmente del mismo modelo. Además, la suma de las proporciones de acuerdo sistemático y acuerdo aleatorio es igual a la probabilidad empírica observada de acuerdo, que para los datos del ejemplo es igual a la proporción de los elementos diagonales P (61 + 26 + 31)/164 = .7195. Del mismo modo, la suma de las proporciones de desacuerdo en el triángulo superior e inferior es igual a la probabilidad empírica observada de desacuerdo, que por definición es el complemento de la probabilidad de acuerdo observado, es decir, 1 – .7195 = .2805. Obviamente, para cada uno de los modelos ajustados la suma de las combinaciones *X1Y2*, *X2Y1* y *X2Y2* es igual al acuerdo aleatorio y desacuerdo que se obtiene en el modelo mixtura básico con una dos clases. En consecuencia, se preserva la naturaleza del modelo mixtura original pero se amplía el número de clases con el objeto de distinguir el acuerdo aleatorio del desacuerdo. A su vez, el desacuerdo se formula de forma que posibilite la prueba de la hipótesis de

homogeneidad marginal mediante la estimación separada del desacuerdo en ambas porciones triangulares.

8.4. Un índice de sesgo basado en modelos mixtura

La revisión de los procedimientos más relevantes actualmente existentes para la detección y medición del sesgo entre observadores nos permite llegar a la conclusión de que en la actualidad los investigadores aplicados de las ciencias del comportamiento y de las ciencias sociales no disponen de herramientas satisfactorias para obtener estimaciones fiables e insesgadas del sesgo entre evaluadores. Como se infiere de la sección anterior, se han propuesto distintas alternativas para detectar y medir el sesgo, pero la mayoría de ellas se basan en los datos brutos de una tabla de acuerdo y aplican procedimientos estadísticos globales que permiten responder a hipótesis concretas, pero no abordan dos aspectos fundamentales que justifican la existencia del sesgo entre observadores, a saber, la descomposición del grado de acuerdo y desacuerdo entre observadores y la separación de sesgo y error de medida en componentes mutuamente independientes. El primer aspecto ha recibido mucha atención en años recientes (Schuster, 2002; Schuster y Smith, 2002; Martín y Femia, 2004, Ato, Benavente y López, 2006), pero el segundo sigue siendo uno de los aspectos olvidados de la investigación aplicada. La definición del desacuerdo en el contexto de dos variables latentes, distinguiendo para la segunda variable latente una clase con los elementos del triángulo superior y otra clase con los elementos del triángulo inferior de la tabla de contingencia, permite sin

embargo desarrollar una medida de sesgo basada en los modelos mixtura.

Para los modelos de la familia QI que no tienen restricciones de homogeneidad o uniformidad marginal (en concreto, para el Modelo QI, que se asocia con la medida de acuerdo Δ de Martín y Femia, 2006, y el Modelo QIC, que se asocia con la medida de acuerdo α de Aickin, 1990), puede también definirse un índice de sesgo de similares características a la generalización del índice BI propuesta por Ludbrook (2004), pero obtenido mediante el ajuste de modelos mixtura, en lugar de hacerlo mediante un enfoque descriptivo simple. El índice de sesgo (ϵ) que proponemos aquí se define del mismo modo que el índice de Ludbrook (2002, 2004), es decir, mediante una diferencia en valor absoluto de las frecuencias triangulares superior e inferior de la tabla de acuerdo, y se obtiene, una vez conocidas las proporciones de las cuatro clases del modelo mixtura ampliado a dos variables latentes, calculando también la diferencia en valores absolutos entre las proporciones de las dos clases latentes que evalúan el desacuerdo (en concreto, las clases *X2Y1* y *X2Y2*).

La Tabla 8.5 presenta a su vez cuatro ejemplos ficticios para sendas tablas de contingencia 4×4 en las que se ajustan los Modelos QI y QIC de la familia de modelos de cuasi-independencia sin alterar el total muestral (N = 223), con el objeto de valorar detalladamente las diferencias entre ambos índices.

La primera tabla corresponde a los datos empíricos de la Tabla 8.1 y representa un caso de sesgo moderado. La proporción de sesgo respecto de la suma de las probabilidades latentes del desacuerdo se reporta también junto con el índice ϵ y se obtiene mediante la razón P = .051/(.181 + .232) = .120.

Los dos índices se comportan de forma similar, pero el índice BI es siempre mayor que el índice ϵ tanto con el Modelo QI como con el Modelo QIC.

La segunda tabla es una modificación de la primera obtenida con los mismos datos, pero la práctica totalidad de las frecuencias del triángulo superior se han trasladado al triángulo inferior con el objeto de representar una situación con un mayor grado de sesgo. En este caso, las proporciones de sesgo son de .890 y .850 respectivamente para los Modelos QI y QIC. Nuevamente, los índices son muy similares pero es ligeramente más alto el índice BI respecto a los dos índices ϵ particularmente para los Modelos QI y QIC.

La tercera tabla utiliza los mismos datos empíricos pero calculando un promedio entre los elementos equivalentes del triángulo superior e inferior de la tabla para representar una situación donde se cumpla simetría y homogeneidad marginal con sesgo nulo. Ambos índices detectan la situación y reportan que el sesgo es inexistente.

La cuarta tabla es similar a la anterior pero se ha realizado una permutación de los elementos del triángulo inferior para no hacerlos coincidir simétricamente con los elementos del triángulo superior forzando así que no se cumpla simetría y homogeneidad marginal con presencia de sesgo no nulo. En este caso las proporciones de sesgo son de .100 (Modelo QI) y .120 (Modelo QIC). En consecuencia, el índice BI se muestra insensible a una permutación de los índices y reporta ausencia de sesgo mientras que el índice ϵ detecta la presencia de sesgo produciendo un índice no nulo.

Tabla 8.5.
Comparación entre el índice BI de Ludbrook (2002, 2004) y el índice ε de Ato, López y Benavente (2008)

Datos empíricos	*Modelo*	*Proporciones de clase latente*	*BI*	*ϵ*	*Ajuste del modelo*
40 6 4 15 **4 25 1 5** **5 2 21 9** **17 13 12 45**	*QI*	*11*: .368; *12*: .219 *21*: .181; *22*: .232	.054	.051 (*P* = .120)	$L^2(5) = 1.560$ *P* = 910*
	QIC	*11*: .444; *12*: .143 *21*: .182; *22*: .231	.054	.049 (*P* = .130)	$L^2(8) = 18.350$ *P*=.020
40 1 0 0 **9 25 1 0** **8 2 21 1** **32 18 20 45**	*QI*	*11*: .543; *12*: .045 *21*: .023; *22*: .390	.386	.367 (*P* = .890)	$L^2(5) = 7.330$ *P* =.200*
	QIC	*11*:.531; *12*: .057 *21*: .031; 22: .381	.386	.350 (*P* = .850)	$L^2(8) = 11.750$ *P* =.160*
40 5 5 16 **5 25 1 9** **5 1 21 10** **16 9 10 45**	*QI*	*11*: .372; *12*: .215 *21*: .206; *22*: .206	.000	.000 (*P* = .000)	$L^2(5) = 2.490$ *P* =.830*
	QIC	*11*: .440; *12*: .147 *21*: .206; *22*: .206	.000	.000 (*P* = .000)	$L^2(8) = 18.470$ *P* =.000
40 5 5 16 **9 25 1 9** **10 16 21 10** **5 5 1 45**	*QI*	*11*: .466; *12*: .122 *21*: .228; 22: .185	.000	.043 (*P* = .100)	$L^2(5) = 9.56\ 0$ *P* =.100*
	QIC	*11*: .466; *12*: .122 *21*: .231; *22*: .181	.000	.050 (*P* = .120)	$L^2(8) = 10.210$ *P* =.140*

En consecuencia, varias ventajas pueden en principio derivarse de la utilización del índice de sesgo ϵ basado en modelos mixtura que proponemos aquí, en comparación con el índice descriptivo BI de Ludbrook (2002, 2004).

- En primer lugar, a diferencia del índice BI, ϵ es un índice basado en un modelo mixtura con dos variables latentes obtenido a partir de las

proporciones de las clases latentes y por tanto su interpretación depende del ajuste del modelo subyacente. Desde esta perspectiva, si para los datos de una tabla de contingencia no se ajustara ninguno de los dos Modelos QI o QIC de la familia de modelos de cuasi-independencia, no podría de hecho definirse una medida de sesgo apropiada. Obviamente, para los modelos que requieren satisfacer homogeneidad marginal no tiene ningún sentido definir medidas de sesgo. Por esta razón, el cálculo de las medidas de sesgo se realiza exclusivamente con los Modelos QI y QIC de la familia de modelos de cuasi-independencia.

- En segundo lugar, a diferencia del índice BI, que se obtiene mediante la diferencia absoluta de las proporciones empíricas de los triángulos superior e inferior de la tabla de contingencia, el índice ϵ se obtiene mediante la diferencia absoluta de las probabilidades esperadas de las dos clases latentes que valoran el desacuerdo, y como consecuencia, representan magnitudes corregidas de efectos no controlados. En general, el índice ϵ produce valores de sesgo más pequeños que el índice BI. Así, en la primera tabla de contingencia del Cuadro 3 se muestran los datos del ejemplo de la Tabla 1, que presenta un grado leve de sesgo, donde BI = .054 mientras que ϵ_1 = .051 para el Modelo QI y ϵ_2 = .049 para el Modelo QIC (aunque este modelo no es en esencia interpretable porque no obtiene un ajuste aceptable). En la segunda tabla se muestra un caso con grado extremo de sesgo, donde BI = .386 mientras que ϵ_1 = .367 para el Modelo QI y ϵ_2 = .350 para el Modelo QIC. Ambos modelos son interpretables porque el

ajuste que se obtiene es aceptable.

- En tercer lugar, una característica indeseable del índice BI de Ludbrook (2004) es que tanto el acuerdo como el desacuerdo (y por ende el índice de sesgo) es invariante ante una permutación de los elementos dentro de su triángulo (superior o inferior). En cambio, el índice ϵ no es invariante ante una permutación de los elementos, que de hecho puede cambiar tanto la proporción de acuerdo sistemático como el propio índice de sesgo. En la tercera tabla de contingencia de la Tabla 8.3, que se obtiene forzando la igualdad de los triángulos superior e inferior para representar simetría y homogeneidad marginal perfecta, el índice BI es cero y ϵ_1, ϵ_2 para los diferentes modelos es también cero, pero la proporción de acuerdo sistemático es .372 para el Modelo QI y .440 para el Modelo QIC (aunque este modelo tampoco es interpretable). Nótese además que la suma de las proporciones correspondientes a las clases *X1Y1* y *X1Y2* es en ambos casos igual a *P* = .587). Por el contrario, en la cuarta tabla de contingencia, que es una simple permutación de los datos de la tabla anterior que afecta únicamente a las frecuencias del triángulo inferior, el índice BI es también cero, mientras que ϵ_1=.043 para el Modelo QI y ϵ_2=.050 para el Modelo QIC, y ambos modelos son además interpretables. Nótese además que en los dos modelos la proporción de acuerdo sistemático es igual a 0.466 debido al equilibrio obtenido en la permutación de los elementos.

- En cuarto lugar, la interpretación del índice de sesgo basado en modelos depende del ajuste del modelo mixtura correspondiente y su valoración puede realizarse calculando la proporción

- Y finalmente, mientras que el índice BI no es fácilmente generalizable para más de dos jueces u observadores (aunque siempre es posible formularlo para cualesquier combinaciones de 2 jueces), el índice ϵ puede ser formulado para cualquier número de jueces, aunque en este caso se requiere una minuciosa preparación del flujo de programa para obtener resultados válidos.

8.5. Un estudio de simulación

Con el objeto de estudiar con mayor detalle las diferencias entre el índice BI de Ludbrook (1984) y el índice ϵ propuesto por Ato, López y Benavente (2008), se procedió a realizar un estudio de simulación con tablas de acuerdo 3×3. Con esta finalidad, nuestro equipo de investigación desarrolló una herramienta de software que permitía generar un gran número de tablas de acuerdo ficticias y posibilitara controlar las siguientes variables en el proceso:

1) la prevalencia de las categorías de respuesta;
2) los marginales de fila;
3) los marginales de columna;
4) el porcentaje de acuerdo en la diagonal;

5) el tamaño muestral total;
6) el número de muestras.

La Figura 8.1 es la primera pantalla de la herramienta de software para simulación de tablas de acuerdo 3×3 que presenta todas las variables que fueron objeto de manipulación. La Figura 8.2 es una salida estándar con la generación de varias tablas de acuerdo para un tamaño muestral total de N = 200 y donde se han hecho variar los marginales de fila (en concreto, 40/30/20, dejando un margen de error de 10), los marginales de columna (en concreto, 20/30/40, dejando así un margen de error de 10), el porcentaje de acuerdo en la diagonal de la tabla de acuerdo (30) y el número de muestras a generar (5).

El estudio de simulación se limitó a Modelos de clase latente de cuasi-independencia (QI) y cuasi-independencia constante (QIC) con tamaño muestral de N = 100 que fueron ajustados con el criterio de estimación por máxima verosimilitud de un máximo de 5000 iteraciones, para cada uno de los cuales se generaron al azar un total de 357000 muestras con diferentes tablas de acuerdo 3×3, de cuyo conjunto total correspondían 71400 muestras para cada uno de los niveles de prevalencia establecidos (5 niveles de prevalencia, en incrementos de 10, desde 50 a 90). En consecuencia, el número de muestras con tablas de acuerdo generadas para el estudio de simulación que realizamos fue en total de 714000 muestras.

Los resultados obtenidos presentaban sin embargo algunos problemas que hicieron necesario realizar un proceso de depuración posterior con el objeto de seleccionar muestras válidas para el análisis. En primer lugar, al ser las muestras generadas al azar y al tratarse de tablas de acuerdo 3×3 (es decir,

con nueve frecuencias de celdilla) era bastante probable que se presentaran numerosas repeticiones. En consecuencia, el primer paso del proceso depurativo fue eliminar todos los casos repetidos. En segundo lugar, algunos de los modelos probados no se ajustaban apropiadamente, o sea, no alcanzaban el nivel de probabilidad mínimo requerido ($P > .10$) y por tanto tenían que ser descartados, ya que por convención (Vermunt, 1997) un modelo de clase latente no ajustado no es susceptible de interpretación. Y en tercer lugar, algunos modelos superaron el criterio de 5000 iteraciones requeridas para obtener una solución máximo-verosímil satisfactoria. Puesto que tales modelos son sospechosos de presentar ciertos problemas de estimación en los límites del espacio paramétrico, se aceptaron únicamente modelos que fueron ajustados con un máximo de 3500 iteraciones.

Las Tablas 8.6 a 8.9 presentan una síntesis de los resultados fundamentales del proceso de simulación. Para cada uno de los Modelos (QI ó QIC) y para cada uno de los niveles de prevalencia (con valores 50 a 90, en incrementos de 10) se calcularon todas las posibles diferencias en valor absoluto para cada uno de los marginales de fila y columna y se sumaron en conjunto, obtenido diferencias que oscilaban entre 0 (los marginales de fila y columna son iguales) a 140 (los marginales de fila y columna extremos), en incrementos de 20. Por ejemplo, en el caso de marginales de fila 10/80/10 y de columna 80/10/10, la diferencia entre marginales sumaba en total 70 + 70 + 0 = 140. Para cada una de las combinaciones de prevalencia (PREVAL) y diferencia entre marginales (DIFMARG) se han calculado estadísticos descriptivos básicos (media, mediana, desviación típica y tamaño muestral) obteniendo en valores absolutos las diferencias entre las frecuencias del triángulo superior y las del triángulo inferior (estadístico DIFER) y del mismo modo para las diferencias entre las

proporciones de las dos clases latentes de desacuerdo con el modelo ajustado correspondiente (estadístico DILAT).

Para los Modelos QI que fueron seleccionados (Tablas 8.6 y 8.7) se observa una disminución en el número de muestras válidas conforme aumenta la prevalencia, lo que sin duda es un resultado esperable dado que el proceso de ajuste (y también el número de iteraciones requerido) suele ser más complejo. Asimismo se observan valores relativamente constantes muy similares entre sí para cada uno de los niveles de prevalencia (por ejemplo, examinando las medias o las medianas en función de cada una de las diferencias entre marginales), lo que en el caso del estadístico DIFER es también lógico, puesto que a medida que aumenta la diferencia entre los marginales aumenta paralelamente la medida del sesgo. De igual modo sucede con las desviaciones típicas de cada uno de los estadísticos. En todos los casos, las diferencias con el estadístico DILAT, tanto para el caso de la media, la mediana o la desviación típica, son mínimas, lo que en cierta medida prueba que ambos estadísticos están midiendo lo mismo, aunque con un enfoque radicalmente diferente. A destacar en particular la medida de sesgo que presenta DIFER en el caso de una diferencia nula entre marginales (DIFMARG = 0), para cuyo caso se asume sesgo igual a cero. Sin embargo, en todos los modelos probados, y para todos los niveles de prevalencia, DIFER aporta valores inesperadamente distintos de cero, que van disminuyendo progresivamente conforme aumenta la prevalencia, mientras que DILAT muestra valores exactamente igual a cero, como es esperable.

Las Figuras 8.3 y 8.4 son gráficos de barras múltiples que destacan en particular las principales diferencias entre ambas medidas de sesgo.

En primer lugar, la Figura 8.3 presenta las diferencias en valores absolutos entre las medidas de sesgo DIFER y DILAT en función de las diferencias entre marginales (DIFMARG) para todos los niveles de prevalencia (PREVAL) con el Modelo QI. Recuérdese que en una tabla de acuerdo 3×3 el Modelo QI tiene un ajuste casi perfecto, porque solo deja 1 grado de libertad residual y por tanto es muy cercano al modelo saturado. Como se observa en el gráfico, las diferencias son mínimas (prácticamente despreciables) para todos niveles de prevalencia y para todas las diferencias marginales posibles. La excepción es cuando la diferencia entre marginales es nula, en cuyo caso las diferencias entre ambos estadísticos se dispara cuando la prevalencia es mínima (PREVAL = 50) y disminuye moderadamente conforme aumenta la prevalencia.

En segundo lugar, la Figura 8.4 presenta de nuevo las diferencias en valor absoluto entre los estadísticos DIFER y DILAT en función de las diferencias entre marginales y los niveles de prevalencia. Obviamente, en una tabla de acuerdo 3×3 como la utilizada en el estudio de simulación el Modelo QIC está más alejado del modelo saturado que el Modelo QI, porque deja 3 grados de libertad residuales. Obsérvese que las diferencias son sensiblemente mayores que en el Modelo QI, aunque también se presenta de nuevo la excepción de la diferencia nula entre marginales.

Tabla 8.6. Estadísticos descriptivos (x100) de los Modelos QI con diferente prevalencia en función de las diferencias entre marginales

difmarg	*Estadístico*	*QI-50*		*QI-60*		*QI-70*		*QI-80*		*QI-90*	
		difer	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*
0	Media	1.263	.000	.979	.000	.663	.000	.545	.000	.384	.000
	Mediana	1.000	.000	1.000	.000	1.000	.000	1.000	.000	1.000	.000
	D.Típica	1.061	.000	.792	.000	.608	.000	.525	.000	.507	.000
	N	2494	2494	2055	2055	1977	1977	1363	1363	711	711
20	Media	8.586	8.582	8.617	8.602	8.744	8.758	8.899	8.913	9.204	9.224
	Mediana	8.000	8.100	8.000	8.140	9.000	8.560	9.000	8.800	9.000	9.010
	D.Típica	3.434	2.991	3.120	2.902	2.816	2.701	2.455	2.394	1.808	1.748
	N	8688	8688	7336	7336	7039	7039	4630	4630	2305	2305
40	Media	16.010	15.956	16.019	15.966	16.330	16.315	16.487	16.494	15.722	16.723
	Mediana	17.000	17.360	17.000	17.480	18.000	18.040	18.000	18.380	9.000	9.010
	D.Típica	7.232	7.181	7.191	7.161	6.855	6.853	6.762	6.757	6.701	6.702
	N	11488	11488	9539	9539	8421	8421	4643	4643	2689	2689
60	Media	24.288	24.353	24.417	24.415	24.821	24.827	25.072	25.088	25.437	25.442
	Mediana	27.000	27.060	27.000	27.400	28.000	27.760	28.000	28.080	19.000	19.010
	D.Típica	9.676	9.570	9.460	9.394	8.922	8.901	8.604	8.599	8.852	8.872
	N	11100	11100	8579	8579	6682	6682	3227	3227	2232	2232

Tabla 8.7. Estadísticos descriptivos (x100) de los Modelos QI con diferente prevalencia en función de las diferencias entre marginales

difmarg	*Estadístico*	*QI-50*		*QI-60*		*QI-70*		*QI-80*		*QI-90*	
		difer	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*
80	Media	32.115	32.115	32.150	32.126	32.753	32.765	33.005	33.015	33.263	33.263
	Mediana	37.000	37.000	37.000	37.520	38.000	37.750	38.000	38.001	39.000	39.000
	D.Típica	12.997	12.978	13.025	13.048	12.399	12.396	12.253	12.558	12.194	12.192
	N	8479	8479	5994	5994	4149	4149	1897	1897	1519	1519
100	Media	39.968	39.947	40.232	40.218	41.112	41.120	41.361	41.361	41.647	41.644
	Mediana	47.000	47.000	48.000	47.540	48.000	47.640	48.000	48.010	49.000	49.000
	D.Típica	15.752	15.752	15.546	15.534	14.751	14.745	14.524	14.520	14.347	14.340
	N	4793	4793	3101	3101	1999	1999	869	869	888	888
120	Media	46.833	46.753	46.613	46.604	48.627	48.611	48.894	48.883	48.988	48.983
	Mediana	57.000	57.000	58.000	57.600	58.000	57.640	58.000	58.001	59.000	59.010
	D.Típica	19.910	19.926	19.923	19.955	18.407	18.409	18.256	18.253	18.374	18.368
	N	1833	1833	1121	1121	743	743	300	300	376	376
140	Media	50.832	50.794	50.644	50.659	52.139	52.107	52.388	52.390	52.291	52.298
	Mediana	66.000	66.110	67.000	67.015	67.000	67.001	68.000	68.010	69.000	69.010
	D.Típica	22.941	22.969	23.390	23.404	22.466	22.456	22.456	22.451	22.491	22.584
	N	304	304	188	188	121	121	52	52	79	79

Tabla 8.8. Estadísticos descriptivos (x100) de los Modelos QIC con diferente prevalencia en función de las diferencias entre marginales

difmarg	*Estadístico*	*QIC-50*		*QIC-60*		*QIC-70*		*QIC-80*		*QIC-90*	
		difer	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*
0	Media	1.141	.000	1.076	.000	.799	.000	.631	.000	.421	.000.
	Mediana	1.000	.000	1.000	.000	1.000	.000	1.000	.000	.000	.000
	D.Típica	.936	.000	.876	.000	.715	.000	.617	.000	.501	.000
	N	2397	2397	2218	2218	2072	2072	1561	1561	723	723
20	Media	8.535	8.405	8.579	8.408	8.697	8.758	8.808	8.563	9.142	8.862
	Mediana	8.000	7.580	8.000	7.540	8.000	8.560	9.000	8.680	9.000	8.180
	D.Típica	3.398	2.986	3.230	2.952	2.932	2.701	2.595	2.655	1.909	2.148
	N	8266	8266	7640	7640	7358	7358	5548	5548	2277	2277
40	Media	15.926	15.579	15.929	15.569	16.046	16.315	16.153	15.791	16.410	16.205
	Mediana	17.000	17.040	17.000	17.120	18.000	18.040	18.000	17.460	19.000	18.760
	D.Típica	7.327	7.416	7.233	7.413	7.116	6.853	7.035	7.233	6.960	7.803
	N	10876	10876	9902	9902	9479	9479	6616	6616	2574	2574
60	Media	24.134	23.799	24.093	23.750	24.450	24.827	24.651	24.370	25.026	25.907
	Mediana	27.000	26.140	27.000	26.180	27.000	27.760	28.000	27.440	29.000	28.880
	D.Típica	7.784	9.879	9.651	9.839	9.270	8.901	8.994	9.171	8.598	8.709
	N	10607	10607	9427	9427	8585	8585	5514	5514	2102	2102

Tabla 8.9. Estadísticos descriptivos (x100) de los Modelos QIC con diferente prevalencia en función de las diferencias entre marginales

difmarg	Estadístico	*QIC-50*		*QIC-60*		*QIC-70*		*QIC-80*		*QIC-90*	
		difer	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*	*difer*	*dilat*
80	Media	31.749	31.364	31.671	31.259	32.247	32.765	32.429	32.190	32.890	32.788
	Mediana	36.000	35.580	37.000	36.200	37.000	37.750	38.000	37.920	39.000	39.000
	D.Típica	13.296	13.576	13.423	13.732	12.849	12.369	12.753	12.986	12.499	12.636
	N	8530	8530	7227	7227	6255	6255	3715	3715	1434	1434
100	Media	39.728	39.327	39.437	39.087	40.519	41.120	40.979	40.814	41.374	41.296
	Mediana	47.000	45.400	47.000	46.780	47.000	47.640	48.000	48.040	49.000	49.000
	D.Típica	15.979	16.228	16.193	16.452	15.180	14.745	14.838	14.040	14.582	14.676
	N	5413	5413	4264	4264	3488	3488	2063	2063	818	818
120	Media	46.323	45.919	46.352	45.991	47.989	48.611	48.422	48.257	48.725	48.618
	Mediana	57.000	56.060	58.000	56.740	57.000	57.640	58.000	58.020	59.000	59.000
	D.Típica	20.042	20.355	20.241	20.581	18.870	18.409	18.604	18.774	18.569	18.699
	N	2449	2449	1836	1836	1442	1442	856	856	340	340
140	Media	48.141	47.721	49.723	49.318	51.133	52.107	51.619	51.481	52.503	52.426
	Mediana	65.000	64.440	67.000	66.100	67.000	67.001	68.000	68.000	69.000	69.000
	D.Típica	24.734	24.866	23.856	24.141	23.016	22.456	22.854	22.990	22.586	22.654
	N	491	491	365	365	290	290	175	175	63	63

Figura 8.1.
Herramienta de software para simulación Monte-Carlo de tablas de acuerdo 3x3

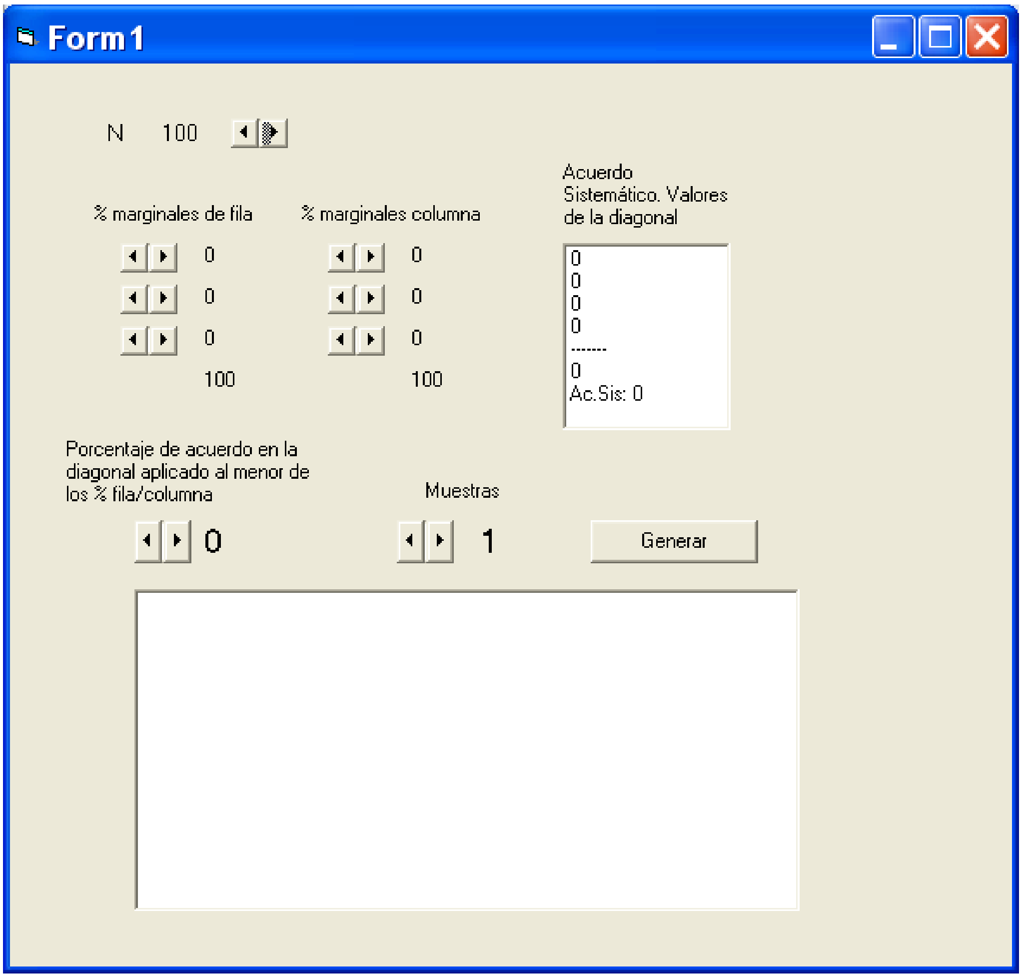


Figura 8.2.
Salida típica de la herramienta de software para simulación Monte-Carlo para N=200 controlando marginales de fila/columna, porcentaje de acuerdo y tamaño muestral

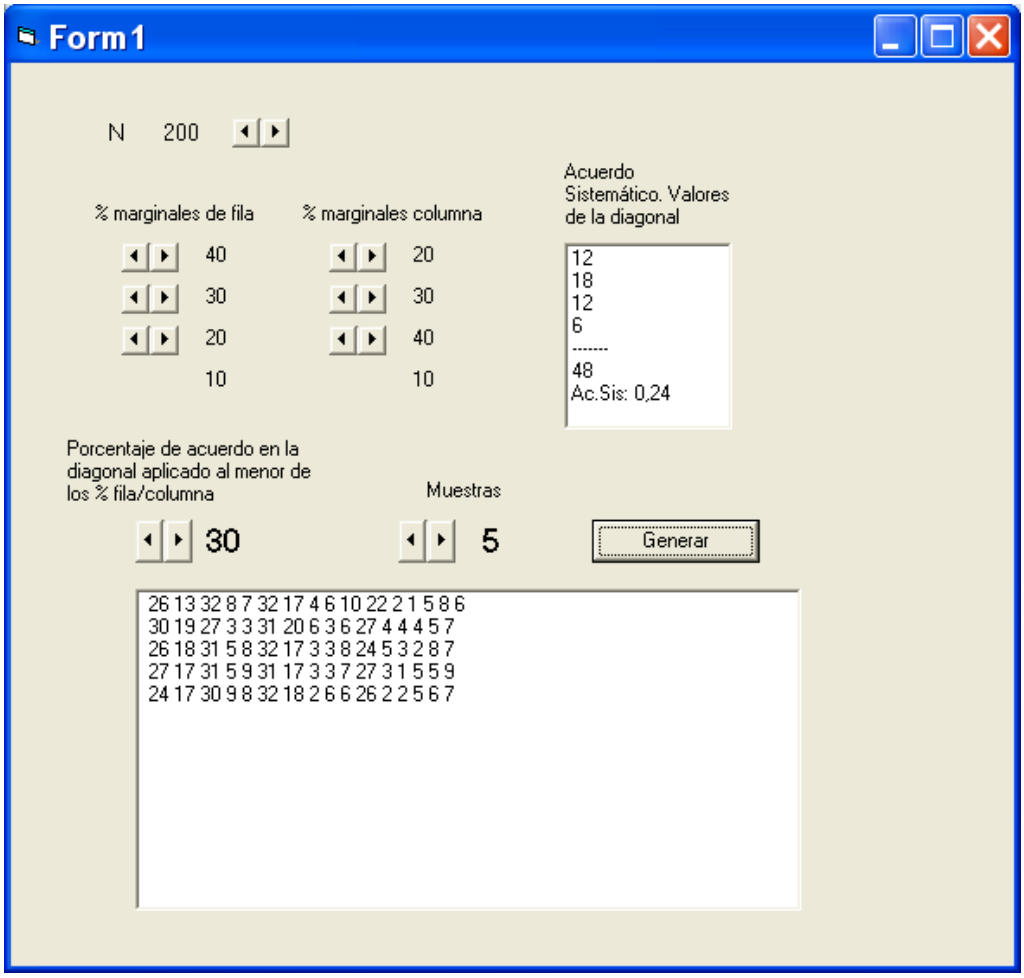


Figura 8.3.
Diferencias entre el índice BI y el índice ε para el Modelo QI en función de las diferencias entre marginales y de la prevalencia

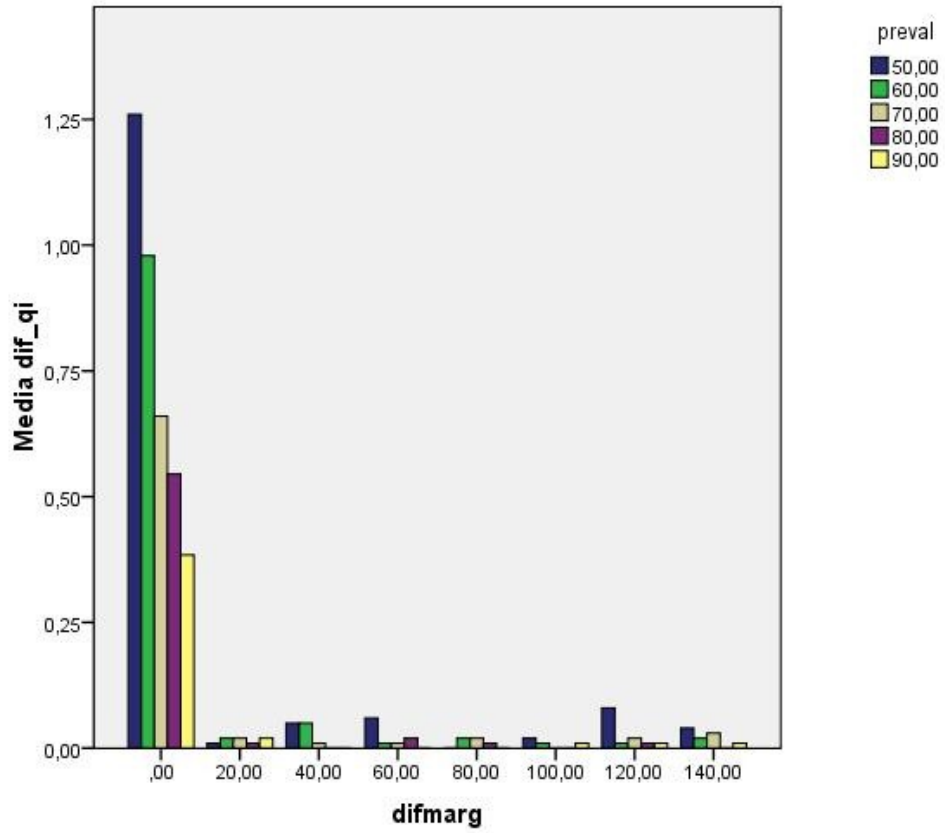
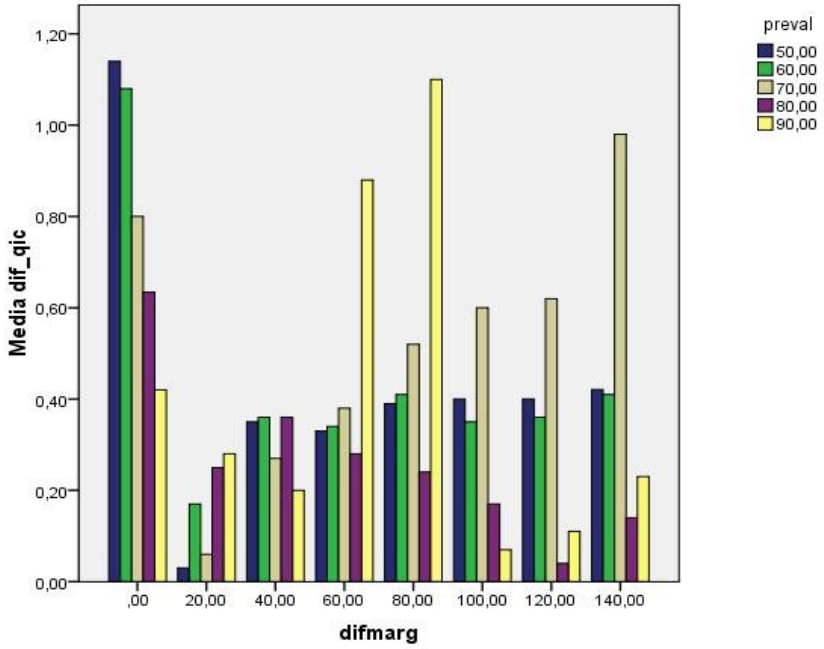


Figura 8.4.
Diferencias entre el índice BI y el índice ε para el Modelo QIC en función de las diferencias entre marginales y de la prevalencia



SALIDA 8.1

FLUJO DEL PROGRAMA LEM PARA AJUSTAR LA FAMILIA DE MODELOS MIXTURA DE CUASI-INDEPENDENCIA CON UNA VARIABLE LATENTE

Para emplear este flujo del programa LEM (Vermunt, 1997) deben mantenerse como están todas las lineas que no están marcadas con asterisco y desmarcar, para cada modelo que se quiera someter a prueba, todas las líneas que corresponden a tal modelo. Por ejemplo, si se desea probar el Modelo QI (esto es, para estimar una versión de *delta* de Martín y Femia, 2006, o el modelo *Heterogeneous raters* de Schuster,) deben desmarcarse las líneas 6 a 9 del flujo. Los datos que se utilizan aquí son los del Ejemplo 8.1 (Tabla 8.1) de Dillon y Mullani (1984). Nótese además que hay tres procedimientos alternativos para ajustar el Modelo QIHX (estimador de *kappa*), el primero de los cuales emplea una aproximación marginal (Bergsma,1998; Vermunt, Rodrigo y Ato, 2001) y los dos restantes una aproximación loglineal clásica. Para más detalles sobre el proceso de ajuste e interpretación de los modelos véase Ato, Benavente y López (2006).

```
**********************************************************
lat 1
man 2
dim 2 3 3
lab X A B
*QI
*mod {spe(A,1a) spe(B,1a) fac(XAB,3) wei(XAB)}          *Delta (heterogeneous raters)
*des [
*1 0 0 0 2 0 0 0 3   0 0 0 0 0 0 0 0 0
*]
*QIC
```

```
*mod {spe(A,1a) spe(B,1a) cov(XAB,1) wei(XAB)}              *Alpha
*des [
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0  * delta
*]
*QIH
 *mod {spe(A,B,1a) fac(XAB,3) wei(XAB)}                    *Lambda (homogeneous raters)
 *des [
 *1 0 0 0 2 0 0 0 3   0 0 0 0 0 0 0 0 0
 *]
 *QICH
 *mod {spe(A,B,1a) cov(XAB,1) wei(XAB)}                    *Pi Scott
 *des [
 *1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0  * delta
 *]
 *QIU
 *mod {cov(XAB,3) wei(XAB)}                                *RE Maxwell
 *des [
 *1 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 * delta1
 *0 0 0 0 1 0 0 0 0   0 0 0 0 0 0 0 0 0 * delta2
 *0 0 0 0 0 0 0 0 1   0 0 0 0 0 0 0 0 0 * delta3
 *]
 *QIHX (1)
 *mod {mar(XA,XB,b,2) cov(A,B,2) cov(XAB,3) wei(XAB)}  * Kappa marginal
 *des [
 *0 0 0  0 0 0  1 0 -1  -1  0 1          * Restricción marginal (1)
 *0 0 0  0 0 0  0 1 -1   0 -1 1          * Restricción marginal (2)
 *1 0 0  0 1 0                           * Homogeneidad marginal (1)
 *1 0 0  0 1 0                           * Homogeneidad marginal (2)
 *1 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0  * Delta1 Homogeneidad diagonal (1)
 *0 0 0 0 1 0 0 0 0   0 0 0 0 0 0 0 0 0  * Delta2 Homogeneidad diagonal (2)
 *0 0 0 0 0 0 0 0 1   0 0 0 0 0 0 0 0 0  * Delta3 Homogeneidad diagonal (3)
 *]
*QIHX (2)
*mod {cov(XAB,2) cov(XAB,1) wei(XAB)}                     * Kappa loglineal
*des [
*0 0 0 0 1 0 0 0 0   0 1 0 1 2 1 0 1 0
*0 0 0 0 0 0 0 0 1   0 0 1 0 0 1 1 1 2
*1 0 0 0 1 0 0 0 1   0 0 0 0 0 0 0 0 0
*]
*QIHX (3)
*mod {fac(XAB,XAB,XAB,2) cov(XAB,1) wei(XAB)}           * Kappa loglineal
*des [
*0 0 0 0 0 0 0 0 0  1 1 1 2 2 2 0 0 0
*0 0 0 0 0 0 0 0 0  1 2 0 1 2 0 1 2 0
*1 0 0 0 2 0 0 0 0  0 0 0 0 0 0 0 0 0
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0
```

```
 *]
*QIUAU
*mod {spe(A,1a) spe(B,1a) cov(AB,1) cov(XAB,1) wei(XAB)} * QIU + Asociación Uniforme
*des [
*1 0 -1 0 0 0 -1 0 1                          * Asociación uniforme
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0 0       * Delta
*]


sta wei(XAB)[1 0 0 0 1 0 0 0 1   1 1 1 1 1 1 1 1 1]          * Pesos variable latente X

dat [61 26 5 4 26 3 1 7 31]                                   * Datos de Dillon y Mullani (1984)
see 2009                                                      * Semilla de aleatorización
cri .00000000000000000000000000000001                         * Criterio de ajuste
*******************************************************************************
```

SALIDA 8.2

FLUJO DEL PROGRAMA LEM PARA AJUSTAR LA FAMILIA DE MODELOS MIXTURA DE CUASI-INDEPENDENCIA CON DOS VARIABLES LATENTES

Para emplear este flujo del programa LEM (Vermunt, 1997) deben mantenerse como están todas las lineas que no están marcadas con asterisco y desmarcar, para cada modelo que se quiera someter a prueba, todas las líneas que corresponden a tal modelo. Por ejemplo, si se desea probar el Modelo QI (o sea, para estimar una versión de *delta* de Martín y Femia, 2004, o el modelo *Heterogeneous raters* de Schuster y Smith, 2002) deben desmarcarse las líneas 6 a 10 del flujo. Los datos que se utilizan aquí son los del Ejemplo 8.1 (Tabla 8.1) de Dillon y Mullani (1984). Nótese también que hay tres procedimientos alternativos para ajustar el Modelo QIHX (estimador de *kappa*), el primero de los cuales emplea una aproximación marginal (Bergsma,1998; Vermunt, Rodrigo y Ato, 2001) y los dos restantes una aproximación loglineal clásica. Para más detalles sobre el proceso de ajuste, véase Ato, López y Benavente (2007).

```
**************************************************************************
lat 2
man 2
dim 2 2 3 3
lab X Y A B
*QI
*mod {spe(A,1a) spe(B,1a) fac(XYAB,3) wei(XAB) wei(YAB)} *Delta (Heterogeneous
raters)
*des[
*1 0 0 0 2 0 0 0 3  0 0 0 0 0 0 0 0 0
*0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0
*]
*QIC
*mod {spe(A,1a) spe(B,1a) cov(XYAB,1) wei(XAB) wei(YAB)}          *Alpha
```

```
*des [
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0
*0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0
*]
*QIH
*mod {spe(A,B,1a) fac(XYAB,3) wei(XAB) wei(YAB)}          *Homogeneous raters
*des[
*1 0 0 0 2 0 0 0 3  0 0 0 0 0 0 0 0 0
*0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0
*]
*QICH
*mod {spe(A,B,1a) cov(XYAB,1) wei(XAB) wei(YAB)}          *Pi Scott
*des [
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0
*0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0
*]
*QIU
*mod {fac(XYAB,3) wei(XAB) wei(YAB)}                      *RE Maxwell
*des [
*1 0 0 0 2 0 0 0 3  0 0 0 0 0 0 0 0 0
*0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0
*]
*QIHX
*mod {mar(XA,XB,b,2) cov(A,B,2) cov(XAB,3) wei(XAB) wei(YAB)}  * kappa (marginal)
*des [
*1 0 -1 -1 0 1  0 0 0 0 0 0 *1 0 -1 -1  0 1          * Restricción marginal (1)
*0 1 -1 0 -1 1  0 0 0 0 0 0 *0 1 -1  0 -1 1          * Restricción marginal (2)
*1 0 0  0 1 0                                        * Homogeneidad marginal (1)
*1 0 0  0 1 0                                        * Homogeneidad marginal (2)
*1 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0                * Delta1 diagonal (1)
*0 0 0 0 1 0 0 0 0  0 0 0 0 0 0 0 0 0                * Delta2 diagonal (2)
*0 0 0 0 0 0 0 0 1  0 0 0 0 0 0 0 0 0                * Delta3 diagonal (3)
*]
*QIHX (2)
*mod {cov(XAB,2) cov(XAB,1) wei(XAB) wei(YAB)}            * Kappa (loglineal)
*des [
*0 0 0 0 1 0 0 0 0  0 1 0 1 2 1 0 1 0
*0 0 0 0 0 0 0 0 1  0 0 1 0 0 1 1 1 2
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0
*]
*QIHX (3)
*mod {fac(XAB,XAB,XAB,2) cov(XAB,1) wei(XAB) wei(YAB)}    * Kappa (loglineal)
*des [
*0 0 0 0 0 0 0 0 0  1 1 1 2 2 2 0 0 0
*0 0 0 0 0 0 0 0 0  1 2 0 1 2 0 1 2 0
*1 0 0 0 2 0 0 0 0  0 0 0 0 0 0 0 0 0
```

```
*1 0 0 0 1 0 0 0 1  0 0 0 0 0 0 0 0 0
*]
 sta wei(XAB) [1 0 0 0 1 0 0 0 1  0 1 1 1 0 1 1 1 0]                    * Pesos para X
 sta wei(YAB) [1 1 1 0 1 1 0 0 1  1 0 0 1 1 0 1 1 1]                    * Pesos para Y
dat [61 26 5 4 26 3 1 7 31]                                  * Datos de Dillon y Mullani (1984)
see 2009                                                      *Semilla de aleatorización
cri .0000000000000000000000000000000000000001                * Criterio de ajuste
*****************************************************************************
```

Conclusiones

El análisis de la literatura de investigación sobre las medidas de acuerdo entre jueces realizado en este trabajo ha partido distinguir entre dos grandes grupos de medidas de acuerdo: medidas descriptivas y medidas basadas en modelos.

Las **medidas descriptivas** definen coeficientes basados en la aplicación de algún principio general (i.e., el principio RCA de acuerdo corregido del azar, o el principio de la distancia euclidiana) que pueden aplicarse con variables numéricas, categóricas o ambas en el contexto del muestreo tipo III (donde se asume que se fijan simultáneamente los marginales y los jueces con similar grado de experiencia). Hay una gran variedad de medidas descriptivas, pero excluyendo el coeficiente *kappa*, que parece ser un procedimiento universalmente aceptado, los demás son en la práctica exclusivos de un área de investigación concreta y apenas se aplican en contextos aplicados diferentes.

La investigación seguida con las medidas descriptivas revela en el fondo el interés por generalizar los coeficientes de acuerdo para hacerlos universalmente aplicables. Así, algunos de los coeficientes originalmente definidos con variables numéricas para dos jueces u observadores (por ejemplo, el coeficiente de correlación de concordancia propuesto por Lin en 1989) han progresado después hacia su generalización a más de dos jueces (Barnhart, Haber y Song, 2002) y más recientemente a variables categóricas (King y Chinchilli, 2001; Lin, 2007) e incluso a medidas repetidas (Carrasco, King y Chinchilli, 2009; Chinchilli y otros, 1996; King, Chinchilli y Carrasco, 2007; Williamson, Manatunga y Lipsitz, 2000). Aunque el recorrido todavía es muy corto, básicamente lo mismo ha sucedido con el coeficiente r_{WG} (LeBreton y Senter, 2008).

La misma secuencia de acontecimientos ha ocurrido también con las medidas descriptivas de acuerdo para variables categóricas como el coeficiente *kappa*, que en su origen se definieron además para dos jueces y variables nominales (Cohen, 1960) y progresivamente se ha intentado su generalización a más de dos jueces (Fleiss, 1971) y a variables ordinales (Cohen, 1968), donde parece haberse estancado. De hecho, el coeficiente *iota* (Janson y Olson, 2001, 2004) se propuso inicialmente para ampliar los límites del coeficiente *kappa* mediante el empleo del principio de las medidas de distancia de Berry y Mielke (1988, 2007), y permitió definir una medida de amplia generalización para datos numéricos y categóricos, apropiada para cualquier número de jueces, e incluso para conjuntos diferentes de jueces y para más de una variable de respuesta.

En este trabajo se ha demostrado que, a pesar de los diferentes orígenes de

las medidas descriptivas citadas, para cada una de las cuales existe una fórmula RCA convencional y una estimación de su varianza, todas ellas pueden obtenerse igualmente en la práctica utilizando una forma peculiar del modelo ANOVA en dos sentidos mixto, con los Ítemes u Objetos como efecto aleatorio y los Jueces u Observadores como efecto fijo, derivando, a partir de una estimación atípica de las esperanzas de las medias cuadráticas, un coeficiente de correlación intraclase que representa la medida descriptiva en cuestión. Se requiere para ello un procedimiento computacional más complejo que la simple aplicación de una fórmula para obtener el estimador y su varianza mediante el enfoque descriptivo, tal y como se ha detallado en los Capítulos 4 y 5 de este trabajo y para los que hemos desarrollado plantillas de cálculo con el objeto de facilitar la comprensión de las cantidades básicas necesarias para el procedimiento de cálculo y su posterior programación. Para alcanzar este objetivo con variables no numéricas se ha recurrido a la formulación del ANOVA para variables categóricas (CATANOVA) propuesta por Light y Margolin (1971) y Margolin y Light (1974) a partir de la idea genial de Gini (1939) de definir una suma de cuadrados a partir de la distancia entre elementos, en lugar de hacerlo a partir de diferencias respecto de una media, puesto que una media no puede obtenerse con variables categóricas. Esta formulación, que se aplica también al coeficiente *iota*, ha tenido no obstante escasa continuidad en la literatura hasta el momento, si exceptuamos los trabajos de Onokogu (1985a,b) y de Singh (1993, 1996).

La utilización del ANOVA clásico y del CATANOVA en dos sentidos sin interacción, seguida de la estimación de los componentes de la varianza y el desarrollo de un coeficiente de correlación intraclase, representa un procedimiento estable y asequible al investigador aplicado para estimar el

acuerdo entre jueces con cualquiera de los coeficientes *pi* y *kappa* generalizados, con el coeficiente *iota* y con el coeficiente de correlación de concordancia. Sin embargo, queda mucho por hacer en esta línea de investigación, de lo que tendrá que encargarse la investigación futura en este campo de investigación. Aunque parte del trabajo está siendo realizado por los investigadores involucrados con el coeficiente de correlación de contingencia con variables numéricas, está prácticamente sin desarrollar para variables categóricas. Destacamos en concreto las siguientes líneas de actuación futuras, que no se han contemplado en este trabajo, entre las más relevantes:

- En primer lugar, en todos los casos tratados aquí se asume que un conjunto de I ítemes/objetos es valorado por un conjunto de J jueces/observadores en una ocasión única $(M=1)$. No se ha contemplado la replicación en las medidas, ya que con el modelo ANOVA/CATANOVA en dos sentidos sin interacción pasarían a formar parte del componente de error. Pero una situación de investigación cotidiana solicita de los jueces que valoren un mismo ítem en dos administraciones paralelas o en ocasiones diferentes distantes en el tiempo para garantizar la independencia de las medidas. Obviamente, si se incluye más de una réplica $(M \geq 2)$ por combinación ítem $\times$ juez es preciso ampliar el modelo ANOVA incluyendo el componente interactivo ítem $\times$ juez, lo que lleva a formular modelos ANOVA/CATANOVA en dos sentidos con interacción.

- En segundo lugar, tampoco se ha contemplado la inclusión de medidas repetidas. Las medidas repetidas introducen muchas complicaciones en los modelos ANOVA y solo pueden ser rigurosamente estimadas con modelos ANOVA mixtos. Además, no ha habido ningún intento hasta el momento presente por incluir medidas repetidas en los modelos CATANOVA. Sin embargo, el contexto longitudinal es una interesante línea de investigación que, pese a sus dificultades, no debe ser en ningún momento abandonada.

- En tercer lugar, a pesar de su interés, poco se ha experimentado con la inclusión de covariantes y variables de agrupamiento, que pueden aportar información muy útil para profundizar en las diferencias individuales entre jueces y/o ítemes. La inclusión de variables de grupo amplía el marco del modelo ANOVA/CATANOVA en dos sentidos con o sin interacción en un modelo ANOVA/CATANOVA en tres sentidos.

Cabe señalar que existen otras opciones integradoras para obtener coeficientes de acuerdo además de la que se propone en este trabajo, que en cierta medida son soluciones bastante similares, aunque no utilizan la estimación mínimo cuadrática, y por tanto son considerablemente más complejas. Una de las propuestas más interesantes consiste en emplear modelos de ecuaciones de estimación generalizada (*Generalized Estimating Equations, GEE*) que admiten variables numéricas y categóricas, variables de agrupamiento y covariantes, y pueden incluir medidas replicadas (Barnhart, Haber y Song, 2002; Barnhart, Song y Haber, 2005; Klar, Lipsitz e Ibrahim,

2000; Lin, Hedayat y Wu, 2007). Otra propuesta reciente utiliza los modelos lineales mixtos generalizados (*Generalized Linear Mixed Models*) con una estructura de efectos aleatorios cruzados para ítemes y jueces, así como empleando una función de enlace logit o probit para variables de respuesta nominal o una función de enlace identidad para variables de respuesta numéricas (Baayen, Davidson y Bates, 2008; Nelson y Edwards, 2008; McCulloch y Searle, 2001).

Por el contrario, las **medidas basadas en modelos estadísticos** obtienen coeficientes de acuerdo a partir de la especificación y el ajuste de un modelo estadístico particular con variables categóricas. Aunque los modelos loglineales fueron la propuesta inicial (Agresti, 1992; Bergan, 1980; Tanner y Young, 1985,ab) y los modelos mixtura fueron una propuesta posterior (Guggenmoos-Holtzman y Vonk, 1998; Schuster, 2002; Schuster y Smith, 2002; Uebersax, 1990), ambos modelos son en lo fundamental similares. Sin embargo, son preferibles los modelos mixtura a los loglineales por el potencial explicativo que representan las clases latentes. Cualquier modelo loglineal o mixtura es miembro de una familia que contiene un modelo básico y un conjunto de restricciones utilizadas con el objeto de simplificarlo. Las dos familias de mayor interés para evaluar el acuerdo entre jueces son la familia de modelos de cuasi-independencia y la familia de modelos de cuasi-simetría. Esta tesis se ha limitado a la familia de modelos de cuasi-independencia.

Dados unos datos empíricos de una tabla de acuerdo, el modo de operar es siempre el mismo y consiste en probar varios modelos de la familia, cuyas propiedades dependen de las restricciones utilizadas respecto del modelo básico, y ajustarlos con estimación por máxima verosimilitud para determinar

qué modelo es el más apropiado, aplicando las reglas estándar para interpretar los modelos loglineales (razón entre desvianza y grados de libertad aproximadamente igual a uno o menor, probabilidad de la desvianza menor de 0.10 y criterio AIC o BIC más bajo de todos los modelos) y finalmente el ajuste condicional de modelos para decidir el modelo óptimo.

Una vez seleccionado el mejor modelo se calcula el coeficiente de acuerdo, que se define como la proporción de casos del conjunto de ítemes donde ambos jueces acuerdan de forma sistemática, sin error. El complemento del coeficiente de acuerdo es la proporción de casos del conjunto donde los jueces acuerdan por azar o están en desacuerdo. La ventaja de esta interpretación es que, en primer lugar, depende del modelo que se ajuste y por lo tanto es diferente para diferentes modelos de la misma familia, y en segundo lugar, revela un mayor grado de pureza en la definición e interpretación del coeficiente de acuerdo.

Se requiere una destreza especial para desarrollar el flujo requerido para operar con modelos mixtura, que varía en función de la dimensión de la tabla de acuerdo, pero es posible ajustarlos con 2 jueces para un número manejable de categorías (desde 2 hasta no más de 7 ó 9, como en las escalas Likert). Un ejemplo completo para una tabla de acuerdo 3×3 se propone en la Salida 8.1.

Mientras que la proporción de acuerdo sistemático tiene una interpretación clara, su complemento representa un "cajón de sastre" que mezcla acuerdo aleatorio y desacuerdo. La necesidad de delimitar ambas fuentes de variación, que es imposible con una variable latente con dos clases, condujo a proponer una ampliación del número de variables latentes (de 1 a 2) y del número de

clases (de 2 a 4). El resultado ha permitido una interpretación desagregada del acuerdo sistemático y aleatorio (primera variable latente) y del desacuerdo (segunda variable latente). Y puesto que el desacuerdo entre jueces es básicamente una fuente de sesgo, propusimos definir una medida de sesgo basada en modelos mediante la diferencia entre los valores estimados de los triángulos simétricos de la tabla de acuerdo.

A partir de una revisión de los índices de sesgo propuestos en la literatura sobre el acuerdo entre jueces, puede deducirse que el índice más sencillo (ya que se expresa mediante una diferencia absoluta de las sumas de las casillas triangulares superior e inferior) y más general (porque puede ser formulado para tablas de acuerdo con cualquier número de categorías) es el índice BI propuesto por Ludbrook (2002; 2004), basado en los trabajos anteriores de Byrt, Bishop y Carlin (1993) y de Lanz y Nebenzahl (1996). Otros índices de sesgo alternativos al índice BI presentan expresiones muy similares a aquél y proceden todos de una formulación propuesta en el clásico trabajo de Bishop, Fienberg y Holland (1975) o bien basada en el supuesto de homogeneidad marginal (Agresti, 2002). En el trabajo de 2004, el propio Ludbrook afirma: "*I am no doubt that the BI is currently the best way to detect bias between two raters and I have argued to this effect*" (p. 115). Pero el comportamiento del índice BI, a pesar de la confianza de su autor, presenta al menos problemas en cierta medida indeseables. En primer lugar, por su propia definición como una diferencia entre frecuencias de celdilla triangulares, el índice BI puede reportar valores diferentes de cero aún en presencia de homogeneidad marginal. Por ejemplo, sea la tabla de acuerdo siguiente, para $N = 100$, con 2 observadores y 3 categorías de respuesta:

	a	*b*	*c*	Marginal
a	**9**	**0**	**1**	10
b	**1**	**18**	**1**	20
c	**0**	**2**	**68**	70
Marginal	10	20	70	100

Obsérvese que la diferencia en valor absoluto entre los triángulos superior e inferior es $|2 – 3| = 1$, pero los marginales son exactamente iguales, por lo que el sesgo debería ser nulo. Y en segundo lugar, y probablemente consecuencia de lo anterior, el índice BI es invariante ante una permutación de las frecuencias triangulares de fila y columna. Por ejemplo, para la tabla de acuerdo anterior, el resultado siguiente

	a	*b*	*c*	Marginal
a	**9**	**2**	**0**	11
b	**0**	**18**	**1**	19
c	**1**	**1**	**68**	70
Marginal	10	21	69	100

produce una tabla de acuerdo donde el índice BI es el mismo (puesto que solo se han permutado las frecuencias triangulares superior e inferior) pero ahora la tabla no presenta homogeneidad marginal.

Para solucionar los problemas del índice BI, se ha propuesto recientemente una ampliación del modelo mixtura con una variable latente y dos clases, tratado en Schuster (2002) y Schuster y Smidt (2002), a dos variables latentes

con dos clases cada una (Ato, López y Benavente, 2008). Esta ampliación nos posibilita delimitar el acuerdo sistemático y el acuerdo aleatorio por un lado del desacuerdo por otro, para conseguir una interpretación más refinada del acuerdo global, y paralelamente definir una medida de sesgo basada en modelos mixtura, que hemos denominado índice ϵ.

Un estudio de simulación Monte Carlo con una herramienta de software diseñada para dar respuesta a esta propuesta ha conducido a la conclusión general de que el índice ϵ tiene un comportamiento más apropiado que el índice BI de Ludbrook, por las razones siguientes:

- ϵ es un **índice basado en modelos**, en contraposición al índice BI, que es un **índice descriptivo**. Un índice de sesgo basado en modelos es un índice que depende del ajuste del modelo, es decir, de que el modelo sea una representación aceptable de los datos empíricos. El índice descriptivo, en cambio, se obtiene de forma directa con los datos empíricos y por tanto es el índice que correspondería al modelo perfecto o saturado. En consecuencia, el índice ϵ será tanto más similar al índice BI cuanto más cercano al modelo saturado sea el modelo más apropiado, y será tanto más dispar cuanto más se aleje de aquél. La justificación de este razonamiento puede seguirse comparando las Figuras 8.3 y 8.4, donde se comparan los dos índices con dos modelos mixtura que discrepan en el ajuste.

- A pesar de que ϵ es un índice que se define de forma similar al índice BI (de hecho es la diferencia entre las frecuencias de casillas

triangulares como puede comprobarse en la Salida 8.2), a diferencia de aquél respeta el supuesto de homogeneidad de los marginales de fila y columna en que se basa la definición de un índice de sesgo, y produce valores nulos en presencia de homogeneidad marginal y valores no nulos en presencia de heterogeneidad marginal.

- Una importante diferencia entre los índices BI y ϵ que no se ha documentado en este trabajo, es que mientras el primero solo puede definirse para 2 jueces u observadores, o a lo sumo con combinaciones pareadas en el caso de más de 2 jueces, el índice ϵ puede ser formulado para cualquier número de observadores, conduciendo a una interpretación global del grado de sesgo que es absolutamente impensable desde un enfoque eminentemente descriptivo.

- Y finalmente, el índice ϵ, como índice basado en un modelo mixtura, no solo permite la inclusión de covariantes o variables de grupo, sino que también hace posible, mediante el oportuno análisis de las probabilidades condicionales de las clases latentes del desacuerdo, una valoración más profunda de las casillas simétricas sobre las que se basa la medida de sesgo a fin de comprender de forma más cabal la naturaleza del desacuerdo existente entre jueces.

Referencias

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics, 44,* 539-548.

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1*, 201-218.

Agresti, A. (2002). *Categorical data analysis*. 2nd edition. Hoboken, NJ: Wiley.

Agresti, A., Ghosh, A. y Bini, M. (1995). Raking kappa: describing potential impact of marginal distributions on measures of agreement. *Biometrical Journal, 7,* 811-820.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa.

Biometrics, 46, 293-302.

Arstein, R. y Poesio, M. (2005). *Kappa= Alpha (or Beta)*. CS Technical report CSM-437. Essex, UK: University of Essex.

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

Ato, M. y. Vallejo, G. (2007). *Diseños experimentales en Psicología.* Madrid: Pirámide.

Ato, M.; Benavente, A. y López, J. J. (2006). Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. *Psicothema, 18,* 638-645.

Ato, M.; López, J. J. Y Benavente, A. (2008). Un índice de sesgo para modelos mixtura. *Psicothema, 20*, 918-923.

Ato, M., Benavente, A., Rabadán, R. y López, J. J. (2004). Modelos con mezcla de distribuciones para evaluar el acuerdo entre observadores. *Metodología de las Ciencias del Comportamiento, V. E., 2004,* 47-54.

Baayen, R. H.; Davidson, D. M. y Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 29,* 390-412.

Bakeman, R. y Gottman, J. M. (1987). *Observing interaction: an introduction to sequential analysis* (Traducción al español: Morata, 1989). Cambridge, MA: Cambridge University Press.

Barnhart, H. X. y Williamson, J. M. (2001). Modelling concordance correlation via GEE to evaluate reproducibility. *Biometrics ,57*, 931-940.

Barnhart, H. X. y Williamson, J. M. (2002). Weighted least-squares approach for comparing correlated kappa. *Biometrics, 58*, 1012-1019.

Barnhard, H. X., Haber, M. y Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics 58,* 1020-1027.

Barnhart, H. X., Song, J. y Haber, J. (2005). Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine, 24*, 1371-1384.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 3-11.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83,* 762-765.

Bartko, J. J. (1991). Measurement and reliability: statistical thinking considerations. *Schizofrenia Bulletin, 17*, 483-489.

Benavente, A.; Ato, M. y López, J. J. (2006). Procedimientos para detectar y medir el sesgo entre observadores. *Anales de Psicología, 22*, 161-167.

Bennet, E. M., Alpert, R. y Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-318.

Bergan, J. R. (1980). Measuring observer agreement using the quasi-independence concept. *Journal of Educational Measurement, 17*, 59-69.

Bergsma, W. (1998). *Marginal models for categorical data*. Tilburg: University of Tilburg.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal*

of Mental Deficiency, 83, 460- 472.

Berry, K. J., y Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.

Bishop, Y. M. M., Fienberg, S. E. y Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: The MIT Press.

Blackman, N.J-M. y Koval, J.J. (1993). Estimating rater agreement in 2 x 2 tables: chance and intraclass correlation. *Applied Psychological Measurement, 17,* 211-223.

Bliese, P. D. (2000). Within-group agreement, non.independence and reliability: implications for data aggregation and analysis. In K.J. Klein y S.W. Kozlowski, eds.: *Multilevel theory, research and methods in organizations* (pp. 349-381). San Francisco, CA: Jossey-Bass.

Bloch, D. A. y Kraemer, H.C. (1989). 2x2 kappa coefficients: measures of agreement or association. *Biometrics, 45,* 269-287.

Bowker, A. H. (1948) A test for symmetry in contingency tables. *Journal of American Statistical Association, 43,* 572-574.

Brennan, R. L. (2001a). *Generalizability theory (2nd Edition)*. New York: Springer.

Brennan, R. L. (2001b). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*, 6-18.

Brennan, R. L. (2001c). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38,*

295-317.

Brennan, R. L. y Prediger, D. J. (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Brennan, R. L., Harris, D. J., y Hanson, B. A. (1987). *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts*. ACT Research Report 87-7. Iowa City: American College Testing.

Brown, R. D. y Hauenstein, N.M.A. (2005). Interrater agreement reconsidered: an alternative to the r_{WG} indices. *Organizational Research Methods, 8,* 165-184.

Burke, M. J.; Finkelstein, L. M. y Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2,* 49-68.

Byrt, T.; Bishop, J. y Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology, 46*, 423- 429.

Carrasco, J. L. (2004). *Concordança nous procediments i aplicacions*. Tesis doctoral en Biometria i Estadistica. Universidad de Barcelona.

Carrasco J. L. y Jover L. (2003). Estimating the Generalized Concordance Correlation Coefficient through Variance Components. *Biometrics. 59,* 849-858.

Carrasco, J. L., King. T. S. y Chinchilli, V. M. (2009). The concordance correlation coefficient for repeated measures estimated by variance components. *Journal of Biopharmaceutical Statistics, 19*, 90-105.

Causinus, H. (1965). *Contribution à l'analyse statistique des tableaux de correlation*. Anals Faculté de Sciences University de Toulouse, 29, 77-182.

Chinchilli, V. M., Martel, J. K., Kumanyika, S. y Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measured designs. Biometrics, 52: 341-353.

Christensen, R. (1997). *Log-linear models and logistic regression*. 2nd Edition. New York: Springer-Verlag.

Cicchetti, D.V. y Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology, 11,* 101-109.

Cichetti, D.V. y Feinstein, A .R. (1990). High agreement but low Kappa. II: resolving the paradoxes. *Journal of Clinical Epidemiology, 43,* 551-558.

Cicchetti, D.V. y Fleiss, J. L. (1977). Comparison of the nu1l distributions of weighted kappa and the *C* ordinal statistic. *Applied Psychological Measurement, 1,* 195-201.

Clogg, C. C. y Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. London: Sage Publications.

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

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

Collis, G. M. (1985). Kappa, measures of marginal symmetry and intraclass correlations. *Educational and Psychological Measurement,5,* 55-62.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88,* 322-328.

Conger, A. J. (1984). Statistical considerations. In M. Hersen, L. Michelson, y A. S. Bellack (Eds.), *Issues in psychotherapy research* (pp. 285- 309). New York: Plenum Press.

Coull, B.A. y Agresti, A. (2003).Generalized log-linear model with random effects, with application to smoothing contingency tables. *Statistical Modelling, 3*, 251-271.

Cox, C. P. (1997). *On the comparison of two proportions using a linear model analysis of 0-1 scorings*. TR97-6. Ames, IA: Department of Statistics.

Craig, R. J. (1981). Generalization of Scott's index of intercoder agreement. *Public Opinion Quarterly, 45,* 260-264.

Crocker, L. y Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Cronbach, L.J.; Rajaratnam, N. y Gleser, N.C. (1963). Theory of generalizability: a liberation of reliability theory. *British Journal of Statistical Psychology, 16*, 137-163.

Cronbach, L. J., Gleser, G.C., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: John Wiley and sons.

Darroch, J.N. y McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics, 28,* 371-388.

Dawid, A. P. Y Skene, A. M. (1979). Maximum likelihood estimation of

observer error-rates using the EM algorithm. *Applied Statistics, 28*, 20-28.

De Mast, J. (2007). Agreement and kappa-type indices. *The American Statistician, 61,* 148-153.

De Vet, H.C.W. (1998). Observer reliability and agreement. En Armitage, P. y Colton, T., eds. *Encyclopaedia of Biostatistics, vol. 4* (pp. 3123-3128). Chichester, UK: Wiley.

De Vet, H.C.W.; Terwee, C.B.; Knol, D.L. y Bouter, L.M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59,* 1033-1039.

Dillon, W. R. y Mullani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19*, 438-458.

Dunn, G. (1989). *Statistical evaluation of measurement errors: Design and analysis of reliability studies.* London: Arnold.

Dunn, G. (2004). *Statistical evaluation of measurement errors: Design and analysis of reliability studies.* 2nd Edition. London: Arnold.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika ,16,* 407-424.

Evers, M., y Namboodiri, N. K. (1979). On the design matrix strategy in the analysis of categorical data. K. F. Schuessler (ed.), *Sociological Methodology* 86-111. San Francisco: Jossey Bass.

Feinstein, A. y Cichetti, D. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 71-76.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh:Oliver & Boyd.

Fleenor, J.W.; Fleenor, J. B. y Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: a methodological comparison. *Journal of Business Psychology, 10,* 367-380.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. 2nd Edition. London, UK: Chapman and Hall.

Fleiss, J. L. (1986). *Design and analysis of clinical experiments*. New York: John Wiley & Sons.

Fleiss, J.L. y Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613-619.

Fleiss, J. L., Cohen, J. y Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Fleiss, J. L., Levin, B. y Paik, M. C. (2003). *Statistical methods for rates and proportions.* 3rd Edition. Wiley.

Fleiss, J. L., Nee, J. C. M. y Landis, J. R. (1979). The large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin, 86*, 974- 977.

Feinstein, A. R. y Cicchetti, D. V. (1990). High agreement but low kappa I: The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-548.

Galton, F. (1886). "Family likeness in stature". *Proceedings of the Royal Society 40,* 42-73.

Galton, F. (1887). On recent designs for anthropometric instruments. *Journal of the Anthropological Institute*, *16,* 2-8.

Galton, F. (1892). *Fingerprints.* London, UK: Macmillan.

Gini, C. (1939). *Variabilita e Concentrazione.* Vol 1. *Memorie di metodologia statistica.* Milan, IT: Giuffre.

Gleser, G. C.; Cronbach. L.J. y Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. P*sychometrika, 30,* 395-418.

Goodman, L. A. y Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-764.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association, 74*, 537-552.

Goodman L. A. y Kruskal W. (1979). *Measures of association for cross classifications*. New York, NY: Springer-Verlag.

Graham, P. (1995). Modelling covariante effects in observer agreement studies: the case of nominal agreement. *Statistics in Medicine, 14*, 299-310.

Guggenmoos-Holtzmann, I. (1993). How reliable are change-corrected measures of agreement. *Statistics in Medicine, 12*, 2.191-2.205.

Guggenmoos-Holtzmann, I. y Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine, 17*, 797-812.

Gwet, K. (2001)a. *Statistical tables for inter-rater agreement coefficients*. StaTaxis Publishing Company.

Gwet, K. (2001b). *Handbook of Inter-Rater Reliability. How to Estimate the Level of Agreement Between Two or Multiple Raters*. Gaithersburg, MD: Stataxis Publishing Company.

Gwet, K. (2002a). Computing inter-rater reliability with the SAS system. *Statistical Methods for Inter-Rater Reliability Assessment Series, 3,* 1-16.

Gwet, K. (2002b). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series, 2,* 1-9.

Gwet, K. (2002c). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment Series, 1,* 1-6

Gwet, K. (2002d). An inquiry into the adequacy of inter-rater reliability assessment methods and the validity if associated standard errors. Retrieved April 11, 2006, from STATAXIS Consulting Web site: http://www.stataxis.com/articles/papers/Inquiry.pdf

Gwet, K. (2008). Computing inter-rater agreement and its variance in presence of high agreement. *British Journal of Mathematical and Statistical*

Psychology, 61, 29-48.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance*. NY: Dryden.

Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology, 53*, 499-503.

Holley, J. W. y Guilford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement, 24*, 749-753.

Hoyt, W.T. (2000). Rater bias in psychological research: when is it a problem and what an we do about it? *Psychological Methods, 5*, 64-86.

Hsu, L. M. y Field, R. (2003). Interrater agreement measures: comments on $kappa_n$,Cohen's kappa, Scott's π and Aickin's α. *Understanding Statistics, 2*, 205-219.

Hubert, L. J. (1978). A general formula for the variance of Cohen's weighted kappa. *Psychological Bulletin, 85,* 183-184.

James, L.R.; Demaree, R.G. y Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85-98.

James, L. R.; Demaree, R. G. y Wolf, G. (1984). r_{WG} : an assessment of within-group interrater agreement. *Journal of Applied Psychology, 78,* 306-309.

Janson, H. y Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological*

Measurement, 61, 277-289.

Janson, H. y U. Olsson (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement, 64,* 62-70.

Janson, S. y Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*, 255-269.

Jackson, K. M.; Sher, K. J.; Gotham, H. J. Y Wood, P. K. (2001). Transitioning into and out of large-effectg drinking in young adulthood. *Journal of Abnormal Psychology, 110,* 378-391.

Johnson, S. M. y Bolstad, O. D. (1973). Methodological issues in naturalistic observations: some problems and solutions for field research. En L.A. Haemerlynck, L.C. Handy y E.J. Mash, eds.: *Behavior change: methodology, concepts and practice. Proceedings of the Fourth Banff International Conference on Behavior Modification* (pp. 7-67). Champaign, ILL: Research Press.

King, T. S. Y Chinchilli, V. M. (2001a). A generalized concordance correlation coeficient for continuous and categorical data. *Statistics in Medicine, 20*, 2131-2147.

King, T. S. Y Chinchilli, V. M. (2001b). Robust estimators of the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics, 11,* 83 -105.

King, T.S.; Chinchilli, V.M. y Carrasco, J.L. A repeated measures concordance correlation coefficient. *Statistics in Medicine, 26,* 2095-3113.

Kirk, R.E. (1995). *Experimental design: Procedures for the behavioral sciences* (3ª ed.). Pacific Grove, CA: Brooks/Cole.

Klar, N.; Lipsitz, S.R. e Ibrahim, J.G. (2000). An estimating equations approach for modelling kappa. *Biometrical Journal, 42*, 45-58.

Krampe, A. y Kuhnt, T. (2007). Bowker's tests for summetry and modification within the algebraic framework. *Computational Statistics and Data Analysis, 51,* 4124-4142.

Krippendorff, K. (1970). Bivariate agreement coefficient for reliability data. In F. Borgatta y G.W. Bohrnsteadt, eds.: *Sociological Methodology 1970* (pp. 139-150). San Francisco, CA: Jossey-Bass.

Kozlowski, S.W.J. y Hattrup, K. (1992). A disagreement about within-group agreement: *Journal of Applied Psychology, 77*, 161-167.

Kozlowski, S.W.J. y Hults, B.M. (1987). An exploration of climates for technical updating and performance. *Personnel Psychology, 40,* 539-563.

Kraemer H. C. (1979) Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44*, 461-472.

Kraemer H. C. y Bloch D. A. (1988). Kappa coefficients in epidemiology: an appraisal of a reappraisal. *Journal of Clinical Epidemiology, 41*, 959-68.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. In E. F. Borgatta (Ed.). *Sociological methodology 1970* (pp. 139-150). San Francisco: Jossey-Bass.

Lakes, K.D. y Hoyt, W. T. (2008). What sources contribute to variance in observer ratings? Using generalizability theory to assess construct validity

of psychological measures. *Infant and Child Development, 17,* 269-284.

Landis, J. R. y Koch, G. G. (1977a). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Landis, J. R. y Koch, G. G. (1977b). An application of hierarchical kappa-type statistics in the assesment of majority agreement among multiple observers *Biometrics, 33,* 363-374.

Landis, J. R. y Koch, G. G. (1977c). A one way components of variance model for categorical data. *Biometrics, 33,* 671-679.

Lantz, C.A. y Nebenzahl, E. (1996). Behavior and interpretation of the statistics: resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*, 431-434.

LeBreton, J. M. y Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852.

Li, M. F, y Lautenschlager, G. (1997). Generalizability theory applied to categorical data. *Educational and Psychological Measurement, 57,* 813-822.

Light, R. J. y Margolin, B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association, 66*, 534-544.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*, 255-268.

Lin, L. (2000). A note on the concordance correlation coefficient. *Biometrics 56,* 324-325.

Lin, L. (2008). Overview of agreement statistics for medical devices. *Journal of Biopharmaceutical Statistics, 18,* 126-144.

Lin, L., Hedayat, A. S., Sinha, B. y Yang, M. (2002). Statistical methods in assessing agreement: Models, issues and tools. J*ournal of American Statistical Association, 97*, 257-270.

Lin, L., Hedayat, A. S. y Wu, W. (2007). A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics, 17*, 629-652.

Loken, W. y Rovine, M. J. (2006). Peirce's 19th century mixture model approach to rater agreement. *The American Statistician, 60,* 158-161.

Looney, S.W. y Hagan, J. L. (2008). Statistical methods for assessing biomarkers and analyzing biomarker data. En Rao, C.R., Miller, J.P. y Rao, D.C. (eds.): *Handbook of Statistics, vol 27: Epidemiology and Medical Statistics* (pp. 109-147). Amsterdam: Elsevier.

López, J. J. y Ato, M. (2008). *MEVACO: A Windows program for rater agreement evaluation*. Poster presentado en el III European Congress of Methodology. Oviedo.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology, 29*, 529-536.

Ludbrook, J. (2004). Detecting systematic bias between two raters. *Clinical and Experimental Pharmacology and Physiology, 31*, 113-115.

MacCulloch, C. E. y Searle, S.R. (2001). *Generalized, linear and mixed models*. New York, NY: Wiley.

Maclure, M, y Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistics. *American Journal of Epidemiology, 126,* 161-169.

Marcoulides, G. A. (2000). Generalizability theory. En Tinsley, H.E.A. Y Brown, S.D., eds.: *Handbook of Applied Multivariate Statistics and Mathematical modeling* (pp. 527-551). San Diego, CA: Academic Press.

Margolin, B. H. y Light, R. J. (1974). An analysis of variance for categorical data, II: small sample comparisons with chi square and other competitors. *Journal of the American Statistical Association*, 69, 755-764.

Martin, A. y Femia, P. (2004). Delta: a new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology, 57*, 1-19.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130*, 79-83.

Maxwell, A. E. y Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *The British Journal of Mathematical and Statistical Psychology, 21,* 105-116.

McGraw, K. O. y Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153-157.

Mielke, P. W. e Iyer, H. K. (1982). Permutation techniques for analyzing multiresponse data from randomized block experiments. *Communications in Statistics: Theory and Methods, 11* , 1427-1437.

Milliken, G. A. y Johnson, D. E. (2009). *Analysis of messy data* (vol. 1). 2nd Edition. Boca Ratón, CA: Chapman & Hall.

Nelson, K. P. y Edwards, D. (2008). On population-based measures of agreement for binary classifications. *The Canadian Journal of Statistics, 36*, 411-426.

Nelson, J.C. y Pepe, M.S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research, 9,* 475-496.

Nickerson, C.A.E. (1997). Comment on "A concordance correlation coefficient to evaluate reproducibility". *Biometrics, 53,* 103-1507.

Onokogu, I. B. (1985a). An analysis of variance for nominal data. *Biometrical Journal, 27,* 375-384.

Onukogu, I. B. (1985b). Reasoning by analogy from ANOVA to CATANOVA. *Biometrical Journal, 27,* 839-349.

Pearson, K. (1901). Mathematical distributions to the theory of evolution. *Philosophical Transactions of the Royal Society of London (Series A) 197*, 385-497.

Quenouille, M. H. (1949.) Approximate tests of correlation in time series. *Journal of The Royal Statistical Society, Series B,11, 68-84.*

Rae, G. (1988). The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. *Educational and Psychological Measurement, 48,* 367-374.

Robieson, W. Z. (1999). On weighted kappa and concordance correlation coefficient. Ph.D. Thesis. Chicago, IL: Graduate College, University of

Illinois

Schmidt, F.L. y Hunter, J.E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74,* 368-370.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology, 55*, 289-303.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement, 64,* 243-253.

Schuster, C. y Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7,* 384-395.

Singh, B. (1993). On the analysis of variance method for nominal data. *Sankhya: The Indian Journal of Statistics, 55,* 40-47.

Singh, B. (1996). On CATANOVA method for analysis of two-way classified nominal data. *Shankhya: The Indian Journal of Statistics, 58,* 379-388.

Scott, W. A. (1955). Reliability or content analysis: the case of nominal scale coding. *Public Opinion Quarterly, 19,* 321-325.

Shavelson, R. J. y Webb. N. M. (1991). *Generalizability Theory: A Primer.* Newbury Park, CA: Sage Publications.

Shoukri, M. M. (2004). *Measures of interobserver agreement.* Boca Raton, FL: CRC Press.

Shoukri M. M., Chaudhary M. A., Mohamed G. H. (2002). Evaluating normal approximation Confidence Intervals for measures of 2x2 associations with

applications to Twin data. *Biometrical Journal, 44*, 1-14.

Shrout, P.E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*, 301-317.

Shrout, P. E. y Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Siegel, S. y Castellan, N.J. (1988). *Non parametric statistics for the behavioral sciences*. 2nd Edition. New York, NY: McGraw Hill.

Spitzer, R. L.,Cohen, J., Fleiss, J. L., y Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry, 17*, 83–87.

Spitznagel, E. L. y Helzer, J. E. (1985). A proposed solution to the base rate problem in the Kappa statistic. *Archives of General Psychiatry, 42*, 725-728.

Stemler, S.E. (2004). A comparison of consensus, consistency and measurement approaches to estimatinf interrater reliability. *Practical Assessment, Research & Evaluation online*, 9(4). Recuperado el 6 de abril de 2009 de <http://PAREonline.net/getvn.asp?v=9&n=4>

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 40*,105-110.

Suen, H. K. y Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Lawrence Erlbaum.

Tanner, M. A. y Young, M. A. (1985a). Modelling ordinal scale disagreement. *Psychological Bulletin, 98*, 408-415.

Tanner, M. A. y Young, M. A. (1985b). Modelling agreement among raters. *Journal of the American Statistical Association, 80*, 175-180.

Tinsley, H. E. A. y Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 22*, 358-376.

Tinsley, H. E. A. y Weiss, D. J. (2000). Interrater reliability and agreement. In Tinsley, H.E.A. & Brown, S.D., eds.: *Handbook of Multivariate Statistics ad Mathematical Modeling*, ch. 4 (pp. 95-124). San Diego, CA: Academic Press.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (Abstract). *Annals of Mathematical Statistics, 29*, 614.

Tukey, J. W. (1959). A quick, compact two-sample test to Duckworth's specifications. *Technometrics 1* , 31—48.

Tukey, J. W. (1959). Aproximate confidence limits for most estimates. Unpublished manuscript.

Uebersax, J.S. (2003). Statistical Methods for Rater Agreement. Documento de <http://ourworld.compuserve.com/homepages/juebersax/agree.com> recuperado el 30/6/2003.

Uebersax, J. S. y Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine, 9*, 559-572.

Vangeneugden, T.; Laenen, A;. Geys, H.; Renard, D. and Molenberghs, G. (2005). Applying concepts of generalizability gheory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics, 61,* 295-304.

Vermunt, J. K. (1996a). *Log-linear event history analysis: a general approach with missing data, unobserved heterogeneity, and latent variables*. Tilburg. NE: Tilburg University Press.

Vermunt, J. K. (1996b). Causal log-linear modelling with latent variables and missing data. U. Engel and J. Reinecke (eds.*), Analysis of change: advanced techniques in panel data analysis*, 35-60. Berlin/New York: Walter de Gruyter.

Vermunt, J. K. (1997). *LEM: a general program for the analysis of categorical data*. Technical Report. Tilburg: Tilburg University.

Vermunt, J. K.; Rodrigo, M. F. y Ato, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research, 30*, 170-196.

Von Eye, A. (2002). *Configural Frequency Analysis - Methods, Models, and Applications*. Mahwah,NJ: Lawrence Erlbaum.

Von Eye, A. (2004). Base models for Configural Frequency Analysis. *Psychology Science*, *46*, 150 -170.

Von Eye, A. (2005). *Workshop on "Categorical data analysis"*. Free University of Berlin, Germany, Department of Education.

Von Eye, A. y Mun, E. Y. (2005). *Analyzing Rater Agreement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Von Eye, A., y Schuster, C. (2000). The odds of resilience. *Child Development, 71,* 563-566.

Von Eye, A., y Schuster, C. (2000). Log-linear models for rater agreement.

Multiciência, 4, 38-56.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.

Williamson, J. M. y Manatunga, A. K. (1997). Assessing interrater agreement from dependent data. *Biometrics, 53*, 707-714.

Winer, B. J. (1971). *Statistical principles in experimental design.* 2nd Edition. New York: McGraw-Hill.

Winer, B. J., Brown, D. R. y Michels, K. M. (1991). S*tatistical Principles in Experimental Design.* 3rd Edition. New York, NY: McGraw -Hill.

Wikipedia (entrada *inter-rater reliability*).

Williamson, J. M.; Manatunga, A.K. y Lipsitz, S.R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics, 1*, 191-202.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 374-378.