TEMA 2º: TABLAS DE CONTINGENCIA BIDIMENSIONALES.

1º.− <u>Distribución de frecuencias observadas.</u>

El único aspecto cuantificable en el análisis cualitativo es el número de individuos que presenta una combinación los niveles de los factores. Estos valores se recogen en tablas de contingencia. (frecuencias observadas de cada combinación).

Ejemplo de tabla de contingencia:

Factores	**Nivel 1º factor B**	**Nivel 2º factor B**	$n_{i.}$
Nivel 1º factor A	n_{11}	n_{12}	$\sum_j n_{ij}$ para i=1
Nivel 2º factor A	n_{21}	n_{22}	$\sum_j n_{ij}$ para i=2
$n_{.j}$	$\sum_i n_{ij}$ para j=1	$\sum_i n_{ij}$ para j=2	$n = \sum_i \sum_j n_{ij}$

Los n_{ij}
representan el número de individuos observados en cada combinación de los niveles de los factores A, B y se consideran como la realización de una V.A. con valores enteros y positivos.

2º.− <u>Modelos muestrales para las frecuencias observadas.</u>

Nuestro objetivo principal es contrastar la independencia entre los factores en estudio. Para ello tendremos en cuenta los modelos de muestreo utilizados para diseñar el experimento que dependerán de la fijación o no de algunos de los totales marginales.

<u>Modelos muestrales más utilizados:</u>

a.− Poisson: los totales marginales y el total muestral varían libremente.

Una tabla generada por este tipo de muestreo está compuesta por V.A. N_{ij}
independientes asociadas a cada casilla y con distribución de Poisson $p(m_{ij})$.

Distribución de probabilidad conjunta para toda la tabla:

$$P\left[N_{ij} = n_{ij}, \forall i, j\right] = \prod_i \prod_j \frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}$$

(producto de las IxJ distribuciones).

b.− Muestreo multinomial completo: se fija de antemano el tamaño de la muestra.

La distribución del vector asociado a la tabla es una multinomial $M(n, \{P_{ij}, i:1...I, j:1...J\})$

$$P\left[N_{ij} = n_{ij}, \forall i, j\right] = \frac{n!}{n_{ij}!} \quad p_{ij}^{n_{ij}}.$$

Además $m_{ij} = E\left[N_{ij}\right] = np_{ij}$

<u>Nota</u>: si la distribución de una tabla de tipo Poisson se condiciona a que $n_{ij} = n$

, el resultado es una tabla con distribución multinomial $M(n, p_{ij} = \frac{m_{ij}}{m_{ij}})$

c.− Muestreo multinomial independiente: fijados de antemano los totales marginales de uno de los factores.

Para generar una tabla de contingencia seleccionamos MAS de tamaños correspondientes a los tamaños asignados a los niveles de uno de los factores y se clasifican a los individuos en cada muestra según los niveles del otro factor.

$$\forall i, \ (N_{ij} : j : 1...J) \qquad M(n_{i.}, \{P_{j/i}, j : 1...J\});$$

donde $P_{i/j}$

es la probabilidad de clasificar a un individuo de la fila i−ésima en la columna j−ésima.

La función de probabilidad conjunta para la tabla de contingencia es:

$$P\left[N_{ij} = n_{ij}, j : 1...J, i : 1...I\right] = \frac{n_{i.}!}{n_{ij}!} P_{j/i}^{n_{ij}}$$

$$m_{ij} = n_{i.} P_{j/i}$$

<u>nota:</u> si la distribución de una tabla es Poisson con frecuencias esperadas m_{ij}

o multinomial con probabilidades $P_{ij} = \frac{m_{ij}}{n}$

entonces la distribución condicionada del vector $(N_{ij}, j : 1...J)$

a que $N_{ij} = n_{i.}$

es multinomial con $P_{j/i} = \frac{m_{ij}}{m_{i.}}$

d.− Muestreo hipergeométrico: fijados los totales marginales de ambos factores.

En este caso la distribución de la tabla seria una hipergeométrica multivariante.

e.− Muestreo binomial negativo: fijadas las frecuencias de las casillas de un nivel de uno de los factores.

3º <u>Diseños muestrales apareado, longitudinal y de control único.</u>

a.− Diseño apareado: consiste en seleccionar pares de individuos de características similares y clasificar a cada elemento del par según una característica.

b.− Diseño longitudinal: clasificamos un conjunto de individuos según un factor y en dos momentos diferentes de tiempo.

c.− Método d control único: se clasifican a los individuos según dos tratamientos diferentes del factor en estudio.

(tanto en b como en c se considera que el individuo más parecido a uno mismo es el propio individuo).

4º <u>Independencia poblacional y muestral.</u>

Consideremos una tabla de contingencia IxJ generada por MMC y sea p_{ij}

la prob. poblacional de que un individuo sea elegido en la casilla (i, j). El conjunto de todas estas probabilidades para toda la tabla de cómo resultado una tabla similar a la de contingencia con sus respectivas marginales definidas.

Para MMC la hipótesis de independencia entre factores es $P_{ij} = P_{i.}P_{.j} \quad o \quad m_{ij} = \frac{m_{i.}m_{.j}}{n}$

.

En el caso de MMI estudiamos la homogeneidad de proporciones independientes.

$$P_{j/1} = P_{j/2} = ... = P_{j/I} = P_{.j} \quad \forall j : 1...J;$$

o

$$m_{ij} = n_{i.}\frac{m_{.j}}{m..} \quad \text{donde } m.. = \sum_i \sum_j m_{ij}$$

En el caso de tablas cuadradas generadas por datos dependientes, las hipótesis a contrastar son la de simetría de proporciones marginales y la de homogeneidad de proporciones marginales, es decir:

$$P_{ij} = P_{ji} \quad i.j : 1...I$$

$$P_{i.} = P_{.i} \quad i, j : 1...I$$

respectivamente.

SIMETRÍA HOMOGENEIDAD.

5º.− <u>Estimación máximo verosímil de las frecuencias esperadas para los modelos muestrales usuales.</u>

Sea $\{n_{ij}, i : 1...I, j : 1...J\}$

el conjunto de frecuencias observadas.

Consideremos el modelo multinomial completo:

La función de máximoverosimilitud será: $L(P_{ij}) = \frac{n!}{\prod_i \prod_j n_{ij}!} \prod_i \prod_j P_{ij}^{n_{ij}}$

Maximizando el segundo factor mediante el método de Lagrange obtenemos las estimaciones: $P_{ij} = \frac{n_{ij}}{n}$

TEMA 3º: INDEPENDENCIA EN TABLAS DE CONTINGENCIA BIDIMENSIONALES.

Contrastación de la hipótesis de independencia en una tabla de contingencia bidimensional.

1º.− <u>Contrastes de independencia exactos.</u>

En caso de muestras pequeñas.

Método:

1º Determinar el espacio muestral del diseño empleado en la tabla observada.(las tablas)

2º Seleccionar de todas las tablas del apartado anterior las que se alejan tanto o más de H0 que la tabla observada en la dirección de H1.

3º Calcular las probabilidades de ocurrencia bajo H0 de dichas tablas.

4º Calcular el p−valor del test. (sumar las probabilidades de dichas tablas)

5º Comparar el p−valor con el nivel de significación α prefijado.

Si p>α aceptamos H0.

Si p α rechazamos H0.

Inconvenientes: el cálculo de la probabilidad exacta de las tablas puede depender de parámetros desconocidos. Se soluciona estimando estos.

Cuando aumenta la muestra o los niveles de los factores el cálculo del p−valor es muy laborioso.

<u>1.1.− Contraste de independencia en el modelo muestral hipergeométrico.</u>

a.−<u>Test exacto de Fisher a una cola de asociación positiva.</u>

Las hipótesis a contrastar son: $H_0 : P_{1/1} = P_{1/2} = P$

$H_1 : P_{1/1} \quad P_{1/2}$

Se calcula el p−valor del test sumando las probabilidades de las tablas cuyo n_{11} sea mayor o igual que el de la tabla observada. Comparamos con α.

b.− <u>Test exacto de Fisher a una cola de asociación negativa.</u>

Las hipótesis a contrastar son: $H_0 : P_{1/1} = P_{1/2} = P$

$$H_1 : P_{1/1} < P_{1/2}$$

Se calcula el p−valor del test sumando las probabilidades de las tablas cuyo n_{11} **sea menor que el de la tabla observada. Comparamos con** α.

c.− <u>Test exacto de Fisher a dos colas.</u>

Las hipótesis a contrastar son: $H_0 : P_{1/1} = P_{1/2} = P_{.1}$

$$H_1 : P_{1/1} \quad P_{1/2}$$

Las tablas que se alejan de H_0
son las que verifican que $\left|D_{11}\right| \quad \left|D_{11obs}\right|$

Donde $D_{11} = n_{11} - m_{11}$

$$D_{11obs} = n_{11obs} - m_{11}$$

La probabilidad de ocurrencia de una tabla es:

$$P = P\left[N_{ij} = n_{ij} / N_{1.} = n_{1.},\ldots,N_{.2} = n_{.2}\right] = \frac{n_{2.}!n_{1.}!n_{.2}!n_{.1}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!};$$

<u>1.2.−Contraste exacto en el modelo muestral multinomial independiente.</u>

a.− <u>Test de homogeneidad de proporciones.</u>

Las hipótesis a contrastar son: $H_0 : P_{1/1} = P_{1/2} = P$

$$H_1 : P_{1/1} \quad P_{1/2}$$

La probabilidad de ocurrencia de una tabla cualquiera es:

$$p\left[N_{ij} = n_{ij} \forall i, j\right] = \frac{n_{1.}!n_{2.}!}{n_{12}!n_{11}!n_{22}!n_{21}!} p^{n_1} (1-p)^{n_2}$$

<u>2º.− Contrastes de independencia asintóticos.</u>

2.1.− <u>Contraste</u> X^2
<u>de bondad de ajuste a una multinomial de parámetros conocidos.</u>

Las hipótesis a contrastar son: $H_0 : P_i = P_i^0$

$$H_1 : P_i \quad P_i^0$$

Pearson propone el siguiente estadístico $X^2 = \sum_{i=1}^{I} \frac{(N_i - nP_i^0)^2}{nP_i^0}$

el cual se distribuye según una X^2

con I−1 grados de libertad y a nivel de confianza α

. Se rechazará la hipótesis si el valor observado es mayor que el valor esperado.

2.2.− <u>Contraste para una multinomial de parámetros estimados.</u>

Las hipótesis a contrastar son: $H_0 : P_i = P_i^0$

$H_1 : P_i \quad P_i^0$

Se propone el siguiente estadístico $X^2 = \sum_{i=1}^{I} \frac{(N_i - nP_i)^2}{nP_i}$

el cual se distribuye según una X^2

con I−p−1 grados de libertad. Se rechazará la hipótesis nula si el valor observado es mayor que el valor esperado.

2.3.− <u>Contraste X2 de independencia.</u>

Las hipótesis a contrastar son: $H_0 : P_{ij} = P_{i.}P_{.j}$

$H_1 : P_{ij} \quad P_{i.}P_{.j}$

El estadístico propuesto para realizar este contraste es el siguiente:

$$X^2 = \sum_{i} \sum_{j} \frac{(N_{ij} - m_{ij})^2}{m_{ij}}$$

teniendo en cuenta que bajo H_0

tenemos $m_{ij} = \frac{n_{i.}n_{.j}}{n}$

.

Dicho estadístico se distribuye según una X^2

con (I−1)(J−1) grados de libertad. Además si el valor observado supera al esperado, rechazaremos H_0

.

2.4.− <u>Contraste X2 de homogeneidad de proporciones.</u>

Las hipótesis a contrastar son: $H_0 : P_{j/1} = P_{j/2} = \ldots = P_{.j}$

$H_1 : P_{j/1} \quad P_{j/2} \quad \ldots \quad P_{.j}$

teniendo en cuenta que bajo

hipótesis nula se verifica: $m_{ij} = n_{i.}.P_{j/i} = n_{i.}P_{.j} = \frac{n_{i.}n_{.j}}{n}$

.

El estadístico es el mismo utilizado en el contraste anterior.

2.5.− <u>Contrate de independencia de razón de verosimilitudes</u>

Las hipótesis a contrastar son: $H_0 : P_{ij} = P_{i.}P_{.j}$

$H_1 : P_{ij} \quad P_{i.}P_{.j}$

.

El estadístico utilizado en este test es el siguiente: $\lambda = \dfrac{(n_{i.}n_{.j})^{n_{ij}}}{n^{n} \quad \prod_{i} \prod_{j} n_{ij}^{n_{ij}}}$

.

Wiks demostró que $G^2 = -2Ln\lambda$
se distribuye según una X^2
con (I−1)(J−1) grados de libertad bajo hipótesis nula. ($G^2 = 2(\sum_{i} \sum_{j} n_{ij} Ln \dfrac{n_{ij}}{m_{ij}})$

.

La hipótesis nula se rechaza si el valor observado del estadístico es mayor que el esperado para un nivel de significación α
prefijado.

<u>Corrección por continuidad.</u>

Corrección de Yates.

El estadístico corregido tiene la siguiente expresión: $X_c^2 = \sum_{i} \sum_{j} \dfrac{(\left| n_{ij} - m_{ji} \right| - \frac{1}{2})^2}{m_{ij}}$

y se distribuye según una X^2
con (I−1)(J−1) grados de libertad.

3º.− <u>Partición de los estadísticos para detectar fuentes de asociación.</u>

3.1.− Partición de tablas IxJ en tablas 2x2 independientes.

(Nota: aprovechando la reproductividad de la X^2
, cualquier V.A. con dicha distribución y v grados de libertad se puede poner como suma de V.A. X^2
independientes cuyos grados de libertad sumen v).

Landcaster e Irving propusieron el siguiente método para descomponer una tabla en subtablas independientes 2x2:

- **1ª subtabla: dos primeras columnas.**
- **2ª subtabla: suma de las dos primeras columnas y la 3ª columna.**
- **3ª subtabla: suma de las tres primeras columnas y la 4ª columna.**
- **...**

La forma general de dichas tablas es:

T_j		
	$\sum_k^J n_{1k} = S_1^j$ $n_{1\,j+1}$ $\sum_k^J n_{2k} = S_2^j$ $n_{2\,j+1}$	S_1^{j+1} S_2^{j+1}
	$S_\cdot^j$ $n_{\cdot\,j+1}$	$S_\cdot^{j+1}$

El G^2 de la tabla original se descompone como suma de los estadísticos de razón de verosimilitudes asociados a cada una de las subtablas construidas, cosa que no ocurre con el X^2 de Pearson.

Kimball propone el siguiente estadístico para la tabla j−ésima:

$$X_j^2 = \frac{n^2 (S_i^j n_{2\,j+1} - n_{1\,j+1} S_2^j)^2}{n_{1\cdot} n_{2\cdot} S_\cdot^j n_{\cdot\,j+1} S_\cdot^{j+1}}$$

El contraste se realiza en cada un de las subtablas y en cualquier caso se realiza a un nivel de significación $\beta = \frac{\alpha}{j-1}$

.

Esta descomposición en tablas independientes no es única. Para comprobar que una partición da lugar a componentes independientes basta con sumar los estadísticos de razón de verosimilitudes de cada subtabla y comprobar que coincide con el asociado a la tabla completa.

Condiciones para obtener tablas independientes:

- **suma de grados de libertad de las subt. = grados de la tabla completa.**
- **cada frecuencia obs. de la tabla original aparece en una sola subtabla.**
- **los totales marginales en la tabla original aparecen en una sola subtabla.**

3.2.− Partición en tablas dependientes.

Suponiendo que haya asociación queremos comprobar si fijado un nivel de un factor hay dependencia con los restantes niveles. En este caso no se verifica que la suma de los G^2 y Kimball de las subtablas sea igual al Pearson de la tabla original. La contrastación se realiza a un nivel de significación $\beta = \frac{\alpha}{2(j-1)}$

4º.− Análisis de residuos.

Si en una tabla de contingencia la hipótesis de independencia se ha visto rechazada, mediante el análisis de residuos podemos detectar los niveles de los factores que pueden ser los causantes de tal asociación.

Residuos estandarizados: $e_{ij} = \frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}}$

La varianza estimada de los residuos: $V(e_{ij}) = (1 - \frac{n_{i.}}{n})(1 - \frac{n_{.j}}{n})$

Residuos ajustados: $d_{ij} = \frac{e_{ij}}{\sqrt{V_{ij}}}$

Se consideran significativos a un nivel de significación α
aquellos que en valor absoluto superen el cuantil correspondiente a una N(0,1).

5º.− Contraste de independencia para variables ordinales.

En variables ordinales es aconsejable aprovechar la información que podamos obtener del orden de los niveles de los factores.

5.1.− Test de linealidad para una tabla IxJ.

Consiste en descomponer el X^2
en dos componentes independientes que permiten contrastar si existe relación lineal significativa entre dos variables ordinales que han sido codificadas.

Consideramos:

x_i : c digos por filas.
y_j : c digos por columnas.

Estimamos los parámetros de la recta de regresión de una variable sobre la otra por mínimos cuadrados, designando una variable como explicativa y la otra como la explicada. El estimador de la pendiente de la recta de regresión, b , nos proporciona la tendencia o tipo de relación entre x e y.

Una vez estimado el parámetro b, se contrasta su significatividad:

$H_0 : b = 0$
$H_1 : b \quad 0$

Fijado un nivel de significación α
, rechazamos la hipótesis nula si: $\frac{b^2}{V(b)}$
es mayor que el valor observado de una $X^2_{1;\alpha}$
.

6º.− *Análisis de tablas cuadradas generadas por datos dependientes.*

Este tipo de tablas están generadas por diseño apareado, longitudinal o de control único.

Se clasifica a los individuos según una característica en dos instantes de tiempo diferentes, bajo dos tratamientos distintos, ...

Se trata de ver si hay cambios significativos en la variable de interés bajo tratamientos distintos o en dos instantes de tiempo determinados.

Las hipótesis de interés son las de simetría y la de homogeneidad.

6.1.− Test de McNemar.

Consideramos una tabla generada según los diseños anteriores.

Las hipótesis a contrastar son: $H_0 : P_{1.} = P_{.1} \qquad P_{12} = P_{21}$

$H_1 : P_{1.} \quad P_{.1}$

Este test se basa en el estadístico X^2

. Bajo H_0

y suponiendo MMC los estimadores máximo verosímiles de las frecuencias esperadas son:

$m_{11} = n_{11}$

$m_{22} = n_{22}$

$m_{12} = m_{21} = \frac{n_{12} + n_{21}}{2}$

Sustituyéndolos en el estadístico obtenemos:

$X^2 = \frac{(n_{12} - n_{21})}{n_{12} + n_{21}}$

el cual se distribuye según una $X^2_{1,1-\alpha}$

Se rechaza si el valor observado es mayor que el valor esperado.

La corrección de Yates también se puede aplicar a dicho estadístico.

6.2.− Test binomial exacto.

a.− A dos colas: las hipótesis a contrastar son: $H_0 : P_{1.} = P_{.1} \qquad P_{12} = P_{21}$

$H_1 : P_{1.} \quad P_{.1}$

Siendo $m = n_{12} + n_{21}$

, la probabilidad de ocurrencia de una tabla cualquiera es: $P_{n_{12}} = \binom{m}{n_{12}} 0.5^m$

. El p−valor del test es $p = \sum_{n_{12} / p_{n_{12}} \; p_{n_{12obs}}} p_{n_{12}}$

b.− A una cola de asociación positiva: $H_0 : P_{12} = P_{21}$

$$H_1 : P_{12} > P_{21}$$

El p−valor del test es $p =$

$$n_{12} \quad n_{12obs} / p_{n12} \quad p_{n12obs}$$

c.− A una cola de asociación negativa: $H_0 : P_{12} = P_{21}$

$$H_1 : P_{12} > P_{21}$$

El p−valor del test es $p =$

$$n_{12} \quad n_{12obs} / p_{n12} \quad p_{n12obs}$$

6.3.− Extensión de Bower al test de McNemar para contrastar simetría en una tabla cuadrada.

Contrastamos la hipótesis de simetría $H_0 : P_{ij} = P_{ji} \quad i \quad j \quad i > j$

$$H_1 : P_{ij} \quad P_{ji}$$

Bajo H_0

y suponiendo MMC, los estimadores máximo verosímiles de las frecuencias esperadas son:

$m_{ii} = n_{ii}$

$m_{ij} = \dfrac{n_{ij} + n_{ji}}{2}$

$i \quad j$

Sustituyendo estos estimadores en el estadístico X^2

se obtiene el siguiente estadístico:

$$X^2 = \quad \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

$i \quad j$

con i<j.

Bajo la hipótesis nula se distribuye según $X^2_{\frac{I(I-1)}{2}}$

. Se rechazará dicha hipótesis si el valor observado es mayor que el valor esperado.

6.4.− Extensión de Stwart y Maxwell para contrastar la hipótesis de homogeneidad de proporciones marginales en una tabla IxI generada por datos dependientes.

Es interesante cuando la hipótesis de simetría es rechazada.

Se propone el siguiente estadístico: $X^2 = d^1 v^{-1} d$

. Donde d es u vector columna formado por cualquiera de las diferencias $d_i = n_{i.} - n_{.i}$

y v es la matriz cuyos elementos son $v_{ii} = n_{i.} + n_{.i} - 2n_{ii}$

,$v_{ij} = -(n_{ij} + n_{ji})$

de forma que bajo H_0
tiene distribución asintótica con (I−1) grados de libertad.

TEMA 4º: MEDIDAS DE ASOCIACIÓN EN TABLAS IxJ

1º.− <u>Introducción.</u>

Cuando la hipótesis de independencia es rechazada podemos plantearnos cual es el grado de asociación y la dirección en que se produce tal.

Las medidas de asociación son parámetros poblacionales que dependen de las probabilidades poblacionales P_{ij}
. Éstas deben ser fácilmente interpretables y deben estar acotadas de manera que los factores indiquen asociación perfecta o falta de asociación. Suelen estar normalizadas tomando valores entre 0 y 1 ó entre −1 y 1, lo cual permite la comparaciones entre tablas de diferentes tamaños.

A veces los valores extremos no se alcanzan aún cuando hay asociación perfecta.

Distinguimos dos tipos de asociación:

- **Estricta perfecta, cada nivel de uno de los factores está asociado a un único nivel del otro factor.(en cada columna hay una única prob. poblacional no nula).**
- **Implícita :**

- **Asociación perfecta implícita de tipo I: en cada fila habrá una sola prob. No nula pero en alguna columna habrá más de una prob. no nula.**
- **Asociación perfecta implícita de tipo II: algún nivel del factor 1º está relacionado con más de un nivel del factor 2º o viceversa.**

Otra propiedades deseables son la simetría y la invarianza.

Podemos clasificarlas según varios criterios:

- **Medidas parciales y globales.**
- **Medidas nominales y ordinales.**
- **Medidas simétricas o asimétricas.**

2º .− <u>Medidas de asociación en tablas 2x2.</u>

2.1.− Funciones del cociente de probabilidad.

Supongamos una tabla generada por MMC con prob. poblacionales P_{ij}

- **<u>Cociente de probabilidad o razón de productos cruzados.</u>**

Ventaja de un suceso: $\dfrac{prob.ocurrencia}{prob.no\ \ ocurrencia}$

Se define el cociente de probabilidad como:

$$\theta = \frac{w_1}{w_2} = \frac{\frac{p_{22}}{p_{21}}}{\frac{p_{12}}{p_{11}}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Propiedades:

- $\theta \quad [0, \quad]$
- **no definido si** p_{12}
  op_{21}
  son 0. Si las dos son cero hay asociación perfecta estricta positiva. Si alguno es nulo hay asociación perfecta implícita de tipo II.
- $\theta = 0$
  cuando p_{11}
  y/op_{22}
  son nulas. Si las dos son nulas hay APEN. Si una de ellas es nula hay AIT II.
- $\theta = 1$
  dependencia entre los factores.
- $\theta > 1$
  asociación positiva.
- $\theta < 1$
  asociación negativa.
- **Invariante frente a cambios de escala en filas y/o columnas.**
- **El cambio de orden en filas o columnas: mismo grado de asociación pero en dirección opuesta.**

El estimador de θ
es: $\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$
con similar interpretación.

Una medida simétrica es $\theta^* = \ln\theta$
que toma valores entre $^{-} \quad , \quad \cdot$

$\theta^* = 0$
hay independencia.

$\theta^* < 0$
asociación negativa.

$\theta^* > 0$
asociación positiva.

En caso de haber ceros muestrales se propone el siguiente estimador: $\theta = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}$

- **Q de Yule.**

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\theta - 1}{\theta + 1}$$

$Q = 0$ independencia

$Q > 0$ asocic + si $\theta > 1$

$Q < 0$ asocic - si $\theta < 1$

$Q = 1$ asocic perf estrc +

$Q = -1$ asocic perf estrc -

valor muestral: $Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$

2.2.− Medidas que son función del coeficiente de correlación.

Codificando con 0 y 1 las variables de una tabla de contingencia podemos utilizar el cuadrado del coeficiente de correlación de Pearson como medida de asociación.

$$]^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1.}p_{2.}p_{.1}p_{.2}}$$

$[0,1]$

pero no permite determinar la dirección

de la asociación. Para ello consideramos el coeficiente de correlación:

$$] = \frac{(p_{11}p_{22} - p_{12}p_{21})}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}}$$

$[-1,1]$

Si vale 0 hay independencia. Si vale −1, asociación perfecta estricta negativa. Si vale 1, asociación perfecta estricta positiva.

Asociación implícita de tipo II no implica que tome valores extremos.

Invariante frente a cambios en orden de filas o columnas.

Cambia de signo si cambiamos el orden de las filas o columnas.

La estimación muestral es: $\phi = \frac{(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$

con la misma interpretación.

2.3.− Medidas para comparar proporciones.

Suponiendo dos factores, uno explicativo y otro explicado, definimos las siguientes medidas asimétricas:

- **Diferencia de proporciones:** $p = p_{1/1} - p_{1/2}$

Vale 0 si hay independencia.

Vale 1 si asociación perfecta estricta +

Vale −1 si −

Entre 0 y 1 asociación +

Entre −1 y 0 asociación −

- **Riesgo relativo:** $R = \dfrac{1 - p_{1/1}}{1 - p_{1/2}}$

$R = 1$ independencia.

$R = 0$ *APIT II*

si $p_{1/1} = 0, p_{2/2} = 0$ asociaci n perfecta estricta negativa.

si $p_{1/2} = 0$, R no est definido.

Estimación muestral de riesgo relativo:

$$R = \frac{n_{11} n_{2.}}{n_{21} n_{1.}}$$

Podemos utilizar una transformación del riesgo relativo como medida de asociación:

$$LnR = Ln\ \frac{n_{11} + 0.5}{n_{1.} + 0.5} - Ln\ \frac{n_{12} + 0.5}{n_{2.} + 0.5}$$

3º .− <u>Medidas de asociación en tablas IxJ.</u>

3.1.− Medidas basadas en X^2 de Pearson.

- **Medida ϕ^2 de Pearson.**

Valor poblacional: $\phi^2 = \dfrac{1}{n} \sum_i \sum_j \dfrac{(P_{ij} - P_{i.} P_{.j})^2}{P_{i.} P_{.j}}$

Valor estimado: $\phi^2 = \dfrac{1}{n} \sum_i \sum_j \dfrac{(n_{ij} - m_{ij})^2}{m_{ij}} = \dfrac{X^2}{n}$

Vale 0 sii independencia.

Asociación perfecta estricta : vale 1

En tablas 2x2 su valor coincide con ρ^2 . Es simétrica y fácil de calcular.

- **Coeficiente de contingencia.**

Valor poblacional: $C = \sqrt{\frac{\phi^2}{\phi^2 + 1}}$

Valor estimado: $C = \sqrt{\frac{X^2 / n}{(X^2 / n) + 1}}$

Si vale cero hay independencia.

No alcanza su valor máximo aún cuando hay asociación perfecta. Este depende del tamaño de la tabla. Para tablas cuadradas el valor máximo que puede tomar es el siguiente: $C_{max} = \sqrt{\frac{I-1}{I}}$

. En la práctica se utiliza el ajustado: $C_A = \frac{C}{C_{max}}$

- **Medida T de Tschuprov.**

Valor poblacional: $T = \sqrt{\frac{\phi^2}{\sqrt{(I-1)(J-1)}}}$

Valor estimado: $T = \sqrt{\frac{X^2}{n\sqrt{(I-1)(J-1)}}}$

ç

Vale 0 cuando hay independencia.

Vale 1 en caso de asociación perfecta estricta en tablas 2x2.

- **V de Cramer.**

Valor poblacional: $V = \sqrt{\frac{\phi^2}{m}}$ con $m = min\{(I-1),(J-1)\}$

Valor estimado: $V = \sqrt{\frac{X^2}{nm}}$

Vale 0 sii independencia.

En asociación perfecta alcanza su valor máximo.

En tablas cuadradas su valor coincide con T

En tablas 2x2 $V = |^2$

3.2.− Medidas de reducción proporcional del error.

Consideremos los factores A y B. Quiero determinar en qué nivel del factor B clasificar a un individuo elegido al azar. Esta predicción se puede hacer de dos formas:

− Arbitrariamente, si consideramos el nivel del factor A en que se clasifica el individuo. (P1 = prob. de cometer error prediciendo arbitrariamente)

− Predecir el nivel de B, teniendo en cuenta el nivel de A en que está clasificado. (P2 = prob. de cometer error prediciendo de esta forma).

Si A y B son independientes entonces P1 = P2.

Si existe asociación, P1 > P2.

Definimos la medida de la siguiente forma: $$RPE = \frac{P_1 - P_2}{P_1}$$

Interpretación de estas medidas.

− Están entre [0,1]

- **Si los factores son independientes P1=P2 y la medida vale 0.**
- **Si la medida vale 0 puede existir asociación entre los factores.**
- **Si los factores están asociados, P1>P2 y la medida está entre 0 y 1.**
- **Si los factores están perfectamente asociados vale 1.**
- **Son medidas asimétricas y se definen simétricas de la misma forma.**

- **Medida Lambda de Goodman y Kruskal.**

Poblacionalmente toma el valor:

$$\lambda_{B/A} = \frac{(1 - P_{.m}) - (1 - \sum_{i=1}^{I} P_{im})}{1 - P_{.m}}$$

donde $P_{.m} = maxP_{.j}$

$P_{im} = maxP_{ij}$

Propiedades:

- **Indeterminado si P.m=1.**
- **Está entre 0 y 1.**
- **Si A y B son independientes entonces vale 0.**
- **Si vale 0: ausencia de capacidad predictiva de A para B.**
- **Si vale 1: asociación perfecta estricta o implícita de tipo I.**
- **Invariante frente a permutación de filas o columnas.**

Valor muestral:

$$\lambda_{B/A} = \frac{\sum_{i=1}^{I} n_{im} - n_{.m}}{n - n_{.m}}$$

Hay una simétrica para cuando no sea posible determinar qué factor es el explicativo y cual es el explicado.

$$\lambda = \frac{\sum_{i=1}^{I} n_{im} + \sum_{j=1}^{J} n_{mj} - n_{m.} - n_{.m}}{2n - n_{m.} - n_{.m}}$$

3.3.− Medidas de asociación en tablas IxJ de tipo ordinal−ordinal basadas en concordancia discordancia.

Un par de individuos se dice concordante si el individuo que se encuentra clasificado en un nivel superior de uno de los factores, es clasificado también en un nivel superior para el segundo factor.

Un par se dice discordante si el individuo que se clasifica en el nivel superior de un factor, está clasificado en un nivel inferior para el segundo factor.

Un par se dice ligado si ambos tienen igual clasificación en ambos factores.

Dado un par elegido aleatoriamente la probabilidad de concordancia es: $2 \sum_{k>i} \sum_{l>j} P_{ij} P_{kl}$

.

La probabilidad de discordancia es: $2 \sum_{k>i} \sum_{l<j} P_{ij} P_{kl}$

.

En una tabla de contingencia se definen dichas probabilidades como:

Concordancia: $2 \sum_{i} \sum_{j} P_{ij} \sum_{k>i} \sum_{l>j} P_{kl}$

=PD

Discordancia: $2 \sum_{i} \sum_{j} P_{ij} \sum_{k>i} \sum_{l<j} P_{kl}$

=PD

Número de pares concordantes: $\sum_{i} \sum_{j} n_{ij} \sum_{k>i} \sum_{l>j} n_{kl}$

=C

Número de pares discordantes: $\sum_{i} \sum_{j} n_{ij} \sum_{k>i} \sum_{l<j} n_{kl}$

=D

Número de pares ligados por el factor A: $\sum_{i=1}^{I} \binom{n_{i.}}{2}$

=TA

Número de pares ligados por el factor B: $\sum_{j=1}^{J} \binom{n_{.j}}{2}$

=TB

TAB= $\sum_i \sum_j \frac{n_{ij}}{2}$

El total de pares es C+D+TA+TB−TAB

Gamma de Goodman y Kruskal.

$$\gamma = \frac{C - D}{C + D}$$

Está entre −1 y 1.

Vale 0 en caso de independencia.

Si hay asociación perfecta negativa vale −1.

Si hay asociación perfecta + vale 1.

Es simétrica.

Cuanto más cercano sea el valor a 1 mayor asociación +.

En tablas 2x2 coincide con Q de Yule.

D de Somers.

Es asimétrica. Considerando el factor B como respuesta: $d_{B/A} = \frac{C - D}{C + D + B_0}$

B0= nº pares ligados por B.

Interpretación igual que el anterior.

Versión simétrica: $d_{B/A} = \frac{C - D}{C + D + \frac{1}{2}(A_0 + B_0)}$

Medida $\rfloor$
de Wilson.

$$\rfloor = \frac{C - D}{C + D + A_0 + B_0}$$

Si A0=0 coincide con $d_{B/A}$

Si B0=0 coincide con $d_{A/B}$

Si ambos son cero coincide con Gamma.

3.4.− Medidas basadas en correlación por rangos.

Para variables de tipo ordinal, el rango es el lugar que ocupa el valor de la variable entre todos ordenados de menor a mayor.

- **Coeficiente de correlación por rangos de Spearman.**

Ai , Bi son las modalidades de las variables y xi , yi son los rangos asociados.

$$1 = 1 - \frac{6 \quad d_i^{\,2}}{n^3 - n} con(d_i = x_i - y_i)$$

varia entre −1 y 1.

Si hay concordancia perfecta entonces vale 1.

Si hay discordancia vale −1.

- **Coeficiente de correlación por rangos de Kendall.**

Hay dos grupos:

1º Γ_a
de Kendall:

Ambos factores tienen el mismo número de categorías y los mismos totales marginales.

$$\Gamma_a = \frac{(C - D)}{\frac{n(n-1)}{2}}$$

Su valor está entre −1 y 1.

Si hay independencia su valor es 0.

Si asociación perfecta + vale 1.

Si asociación perfecta − vale −1.

2º $\Gamma_b = \frac{C - D}{\sqrt{(C + D + A_0)(C + D + B_0)}}$

$$\Gamma_b = \frac{2m(C - D)}{n^2(m-1)}$$

con m = min (I, J).

- **Kappa de Cohen.**

Se utiliza en tablas cuadradas generadas por datos dependientes. Mide el grado de acuerdo entre los casos 1 y 2.

Sea Pii la probabilidad de acuerdo. $\theta_1 = \sum_i P_{ii}$

proporción de casos en los que hay acuerdo. Si hubiese independencia: $\theta_1 = \sum_i P_{i.}P_{.i}$

(proporción de casos en que el acuerdo es casual).

$$K = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$

Su valor muestral es:

$$K = \frac{n \sum_i n_{ii} - \sum_i n_{i.} n_{.i}}{n^2 - \sum_i n_{i.} n_{.i}}$$

4º.− <u>Inferencia en las medidas de asociación.</u>

A partir de las medidas muestrales vamos a intentar establecer el valor de las medidas poblacionales mediante inferencia.

Una medida de asociación es una función f sobre un espacio de vectores asociados a una tabla de contingencia IxJ.

Nuestro objetivo es proporcionar, basándonos en la información muestral, un intervalo de confianza para el valor poblacional de la medida. Mediante el Th. Central del límite, bajo muestreo multinomial completo, el vector de proporciones muestrales tiene distribución asintótica normal multivar.

con vector de medidas P. Aunque nos interesa f(p). Utilizamos el método delta.

$$\sigma^2(f(p)) = \frac{1}{n}\left(\sum_i p_i \phi_i^2 - \left(\sum_i p_i \phi_i\right)^2\right)$$

donde ϕ_i

es el vector formado por las derivadas parciales de f respecto a cada Pij.

El intervalo de confianza es: $(f(p) \quad z_{1-\frac{\alpha}{2}} \sigma(f(p)))$

18

Datos Cualitativos.