Análisis de Relaciones

Penna – Cobos – Vázquez Ferrero – Ulagnero

Análisis de Datos Categóricos

En cualquier área del conocimiento es muy común encontrar situaciones donde los datos recogidos son observaciones de **variables categóricas** cuyos niveles o categorías son empleados en la discriminación o identificación de las unidades muestrales en estudio.

En esta Unidad se pretende introducir el análisis de datos categóricos, el cual sólo se restringirá a la presentación del análisis de **tablas de contingencia.**

Una variable categórica es una característica para la cual la escala de medida consiste de un conjunto de categorías.

Dentro de la escala categórica se distinguen tres tipos principales de variables:

- *Nominales*: son aquellas cuyos niveles no están naturalmente ordenados, por ejemplo género, raza, etc.
- *Ordinales*: son aquellas cuyas distintas categorías tienen un orden natural, por ejemplo diagnóstico de una enfermedad (seguro, probable, improbable), etc.
- *De intervalo*: son aquellas variables de tipo numérico que tienen una distancia entre dos niveles, por ejemplo edad de los individuos (entre 15-20, 20-25 y 25-30 años), etc.

Análisis de Datos Categóricos (cont.)

Ordenando en forma decreciente los tipos de variables enunciados en función de la cantidad de información que proveen, se tiene: 1º de intervalo, 2º ordinal, 3º nominal.

Los métodos diseñados para un tipo de variable pueden ser usados para una de nivel superior. Así, una técnica para variables "ordinales" puede ser usada para una "de intervalo" pero no para una nominal.

Una variable puede ser nominal, ordinal o de intervalo, según lo que se mida o cómo se lo mida.

Por ejemplo, la variable educación es nominal, si se refiere al tipo de educación: pública o privada; ordinal si mide el nivel de educación: preescolar, primario, secundario, terciario o universitario, mientras que es de intervalo si se cuantifica la cantidad de años de educación formal: 0, 1, 2,..., etc.

Cuando las UA extraídas de una población son clasificados de acuerdo a, por lo menos, dos características observadas en ellos, se dice que los mismos están estudiándose en forma bivariada, esto es, por medio de dos variables aleatorias. Para analizar esa información se puede construir, entre otras cosas, una tabla de contingencia.

Análisis de Datos Categóricos (cont.)

Una tabla de contingencia se obtiene cuando el conjunto de unidades de análisis, son clasificadas de acuerdo a uno o más criterios.

Para el análisis de tablas de contingencia es necesario indagar primeramente en la clasificación de las variables que la definen. Ellas pueden ser:

VARIABLES DE *RESPUESTA* O *DEPENDIENTES*

→ Son aleatorias y describen lo que fue observado en las unidades muestrales

VARIABLES DE *CLASIFICACIÓN* O *INDEPENDIENTES*

→ Son fijas por condicionamiento y las combinaciones de sus niveles definen estratos, poblaciones o subpoblaciones a las cuales las unidades muestrales pertenecen

CONCEPTOS PREVIOS

Como se mencionó, el muestreo (cuando es probabilístico) muchas veces tiene como objetivo inferir propiedades de la población a partir de los datos obtenidos de una muestra. Estadísticamente, se pretende conocer los parámetros de la distribución de la variable de interés y es el muestreo el encargado de proveer dicha información.

Luego, considerando las propiedades "clásicas" de los buenos estimadores, los estadísticos muestrales sirven para estimar parámetros que caracterizan a la población.

Recordemos que:

	Estimadores (n)	Parámetros (N)
Media Aritmética →	$\bar{X}$	μ
Varianza →	S^2	σ^2
Desviación estándar →	S	σ
Mediana →	$\tilde{X}$	Θ
…	…	…
Proporción →	$\hat{p}$	Π

Definición de Hipótesis Estadística

Una hipótesis estadística es un supuesto que se hace sobre uno o varios parámetros. Ejemplos de dichos supuestos incluyen el que la media de una población tenga un determinado valor, o que los valores de una variable presenten menor dispersión en torno al valor medio en una población comparada con la dispersión en otra, etc.

Evidentemente, la forma más directa de comprobar tales hipótesis sería estudiando todas y cada una de las UA de la población. Sin embargo esto no siempre es posible (la población podría ser infinita), por lo que el contraste de hipótesis ha de basarse en una muestra aleatoria de la población en estudio. Al no estudiarse la población entera, nunca podremos estar completamente seguros de sí la hipótesis realizada es verdadera o falsa. Es decir, siempre existe la probabilidad de llegar a una conclusión equivocada.

El primer paso, en un ensayo de hipótesis, es la formulación de la hipótesis estadística que se quiere aceptar o rechazar y que se formula con el propósito de rechazarla para así probar el argumento deseado.

Definición de Hipótesis Estadística (cont.)

Por ejemplo, para demostrar que… :

- una metodología de enseñanza es mejor que otra, se plantea la hipótesis de que son iguales, es decir, que cualquier diferencia observada es debida únicamente a variaciones en el muestreo.
- un dado está "cargado" (no existe igual probabilidad de que salgan: 1, 2, … ó 6) se plantea la hipótesis de que no está cargado (es decir, la probabilidad de que aparezca cualquier número es 1/6) y a continuación se estudia si los datos de la muestra llevan a un rechazo de esa hipótesis.

Por este motivo, la hipótesis de partida que se quiere contrastar se llama **hipótesis nula**, y se representa por $\mathbf{H_0}$. La hipótesis nula es, por lo tanto, la hipótesis que se rechaza o no se rechaza como consecuencia del contraste de hipótesis. Por otra parte, la hipótesis que se acepta cuando se rechaza H_0 es la **hipótesis alternativa**, denotada por $\mathbf{H_1}$. Es decir: si se acepta H_0 se rechaza H_1 y si se rechaza H_0 se acepta H_1.

VALOR *P* (*P*-VALUE O P_V) PARA UNA PRUEBA DE HIPÓTESIS

Ya que nos adentramos en prueba de hipótesis, definiremos (en un sentido amplio) una medida de la "credibilidad" de la hipótesis nula, llamada *p-value*. Cuanto más pequeño es el valor *p*, menos probable es que H_0 sea verdadera y por ello, si es menor que el nivel de significación, H_0 se rechaza:

a) En un contraste unilateral izquierdo, el *p-value* corresponde a la probabilidad que el estadístico del contraste tome valores menores que su valor calculado.

b) En un contraste unilateral derecho, el *p-value* corresponde a la probabilidad que el estadístico del contraste tome valores mayores que su valor calculado.

c) En un contraste bilateral, el *p-value* se calcula como la suma de los *p* valores suponiendo una prueba unilateral derecha y una izquierda.

La regla de decisión para rechazar (o no) H_0, bajo el enfoque del p_v, es:

- **Si el *p-value* $< \alpha \Rightarrow$ se rechaza H_0**
- **Si el *p-value* $\geq \alpha \Rightarrow$ no se rechaza H_0**

TABLAS DE CONTINGENCIA A UN CRITERIO DE CLASIFICACIÓN

Si se toma una muestra aleatoria simple de 100 establecimientos educativos y se los clasifica, según un determinado criterio de calidad, en “alto”, “medio” o “bajo”, se obtiene una tabla con un único criterio de clasificación con tres niveles:

Criterio de calidad			
Alto	*Medio*	*Bajo*	**Total**
80	15	5	**100**

En el caso de que se dispusiera de alguna hipótesis sobre la distribución de la variable categórica *criterio de calidad*, estos resultados podrían utilizarse para someterla a prueba.

Por ejemplo, si las especificaciones del grupo de establecimientos educativos, del cual se extrajo la muestra, dicen que las proporciones para las categorías de criterio de calidad son las siguientes:

Criterio de calidad		
Alto	*Medio*	*Bajo*
0,95	0,03	0,02

Tablas de Contingencia a un criterio de clasificación (cont.)

Podría ser de interés probar si las frecuencias observadas son consistentes con las establecidas por las especificaciones, o no. Este tipo de análisis se conoce como **prueba de bondad de ajuste**.

La prueba de bondad de ajuste radica en la *comparación de las frecuencias observadas con aquellas esperadas (por un modelo)* mediante un *estadístico* conveniente. Cuando se realiza una prueba de bondad de ajuste, se establece como H_0 que las frecuencias observadas son consistentes con las frecuencias esperadas.

$$\mathbf{H_0}: \pi_i = \hat{p}_i \ (i = 1, 2, \dots, k) \text{ vs. } \mathbf{H_1}: \text{ alguna igualdad no se cumple}$$

Para la construcción del estadístico de prueba o variable pivotal se estiman las frecuencias esperadas cuando la H_0 es cierta:

$$\chi^2_{H_0} = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i} \sim \chi^2_{(k-p-1)}$$

k = nº de clases o categorías
p = nº de parámetros
fo = frecuencia observada
fe = frecuencia esperada

Tablas de Contingencia a un criterio de clasificación (cont.)

Si consideramos el ejemplo de criterio de calidad, las frecuencias observadas y esperadas (calculadas como: $fe_i = n \times \hat{p}_i$) son las siguientes:

Frecuencias	*Alto*	*Medio*	*Bajo*	**Total**
Observadas	80	15	5	**100**
Esperadas	95	3	2	**100**

En este caso, los grados de libertad del estadístico χ^2 están dados por la diferencia entre el *número de categorías y el número de parámetros que deben estimarse para calcular dichas frecuencias* (suponiendo cierta H_0) *menos uno*; es decir: $3 - 0 - 1 = 2$

Supongamos que se considera oportuno trabajar con un nivel de significación (α) del 1%. La hipótesis a ser planteada es la siguiente:

$\mathbf{H_0}$: $\pi_{Alto} = 0{,}95;\ \pi_{Medio} = 0{,}03;\ \pi_{Bajo} = 0{,}02$ vs. $\mathbf{H_1}$: alguna igualdad no se cumple

TABLAS DE CONTINGENCIA A UN CRITERIO DE CLASIFICACIÓN (CONT.)

$$\chi^2_{H_0} = \frac{(80-95)^2}{95} + \frac{(15-3)^2}{3} + \frac{(5-2)^2}{2} = 2{,}37 + 48 + 4{,}5 = \mathbf{54,87}$$

TABLA 3

gl \ α	0,995	0,99	0,975	0,95	0,90	0,75	0,50	0,25	0,10	0,05	0,025	0,01	0,005
1	0,000039	0,00016	0,00098	0,0039	0,02	0,10	0,46	1,32	2,71	3,84	5,02	6,63	7,88
2	0,010	0,020	0,051	0,103	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,989	1,239	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,30
8	1,344	1,647	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,10	22,00
9	1,735	2,090	2,70	3,33	4,17	5,90	8,34	11,39					
10	2,160	2,560	3,25	3,94	4,87								

Prueba de Hipótesis

$\chi^2_{H_0} = 54{,}87 \Rightarrow$ para 2 gl, $p_v \cong 0{,}005 < \alpha = 0{,}01 \Rightarrow$ se rechaza H_0, es decir que las frecuencias observadas no son consistentes con lo establecido por las especificaciones del criterio de calidad.

Tablas de contingencia a dos criterios de clasificación: Marginales Libres

Siguiendo el ejemplo anterior, una situación diferente se podría encontrar si se distinguieran, además, establecimientos educativos pequeños y grandes. Supóngase que las frecuencias obtenidas al clasificar 100 establecimientos, por tamaño y criterio de calidad, fueron:

	Criterio de calidad			
Tamaño	*Alto*	*Medio*	*Bajo*	**Total**
Pequeño	16	9	5	**30**
Grande	50	12	8	**70**
Total	**66**	**21**	**13**	**100**

En este caso, como en el anterior, la tabla es en sí misma una herramienta descriptiva de la distribución de frecuencias y permite visualizar comportamientos que pueden ser de interés.

Lo usual, es que no se disponga de una distribución de frecuencias teórica en dos vías, por lo que se establece la **hipótesis de independencia**. Esta hipótesis establece que "*el criterio de calidad de los establecimientos es el mismo, independientemente del tamaño*".

Tablas de contingencia a dos criterios de clasificación: Marginales Libres (cont.)

Si la hipótesis de independencia no fuera cierta, entonces, se concluiría que la el tamaño de los establecimientos está asociado al criterio de calidad. El análisis de esta hipótesis se conoce como **prueba χ^2 para la hipótesis de independencia**.

$$H_0: \pi_{ij} = \pi_{i.} \times \pi_{.j} \ (i = 1, 2, \dots, f\,; j = 1, 2, \dots, c) \ \text{ vs. } H_1: \text{alguna igualdad no se cumple}$$

Para la construcción del estadístico de prueba se estiman las frecuencias esperadas cuando la H_0 cierta:

$$\chi^2_{H_0} = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} \sim \chi^2_{(f-1)(c-1)}$$

f = nº de filas
c = nº de columnas
fo = frecuencia observada
fe = frecuencia esperada

Si la hipótesis nula se refiere a la independencia entre las dos variables de *respuesta* que conforman la tabla, esto implica que la distribución conjunta de las mismas puede obtenerse a partir del producto de las distribuciones marginales. Pero, a diferencia de la *Prueba de Bondad de Ajuste*, las proporciones bajo H_0 no son conocidas y deben estimarse.

Tablas de contingencia a dos criterios de clasificación: Marginales Libres (cont.)

Frecuencias observadas

Tamaño	Criterio de calidad: *Alto*	*Medio*	*Bajo*	**Total**
Pequeño	x_{11}	x_{12}	x_{13}	$\mathbf{x_{1\cdot}}$
Grande	x_{21}	x_{22}	x_{23}	$\mathbf{x_{2\cdot}}$
Total	$\mathbf{x_{\cdot 1}}$	$\mathbf{x_{\cdot 2}}$	$\mathbf{x_{\cdot 3}}$	**n**

Frecuencias esperadas

Tamaño	Criterio de calidad: *Alto*	*Medio*	*Bajo*	**Total**
Pequeño	$\frac{x_{\cdot 1} \times x_{1\cdot}}{n}$	$\frac{x_{\cdot 2} \times x_{1\cdot}}{n}$	$\frac{x_{\cdot 3} \times x_{1\cdot}}{n}$	$\mathbf{x_{1\cdot}}$
Grande	$\frac{x_{\cdot 1} \times x_{2\cdot}}{n}$	$\frac{x_{\cdot 2} \times x_{2\cdot}}{n}$	$\frac{x_{\cdot 3} \times x_{2\cdot}}{n}$	$\mathbf{x_{2\cdot}}$
Total	$\mathbf{x_{\cdot 1}}$	$\mathbf{x_{\cdot 2}}$	$\mathbf{x_{\cdot 3}}$	**n**

Tamaño	Criterio de calidad: *Alto*	*Medio*	*Bajo*	**Total**
Pequeño	16	9	5	**30**
Grande	50	12	8	**70**
Total	**66**	**21**	**13**	**100**

Tamaño	Criterio de calidad: *Alto*	*Medio*	*Bajo*	**Total**
Pequeño	$\frac{66 \times 30}{100} = 19{,}8$	$\frac{21 \times 30}{100} = 6{,}3$	$\frac{13 \times 30}{100} = 3{,}9$	**30**
Grande	$\frac{66 \times 70}{100} = 46{,}2$	$\frac{21 \times 70}{100} = 14{,}7$	$\frac{13 \times 70}{100} = 9{,}1$	**70**
Total	**66**	**21**	**13**	**100**

Tablas de contingencia a dos criterios de clasificación: Marginales Libres (cont.)

En este caso, los grados de libertad del estadístico χ^2 están dados por el producto (*filas* – *1*) por (*columnas* – *1*), suponiendo H_0 cierta; es decir: $(2-1)\times(3-1)=2$

Supongamos que seguimos trabajando con un nivel de significación (α) del 1%. La hipótesis a ser planteada es la siguiente:

$\mathbf{H_0}$: el criterio de calidad de los establecimientos es independiente del tamaño

$\mathbf{H_1}$: el criterio de calidad de los establecimientos depende del tamaño

$$\boldsymbol{\chi^2_{H_0}} = \frac{(16-19{,}8)^2}{19{,}8} + \frac{(9-6{,}3)^2}{6{,}3} + \frac{(5-3{,}9)^2}{3{,}9} + \frac{(50-46{,}2)^2}{46{,}2} + \frac{(12-14{,}7)^2}{14{,}7} + \frac{(8-9{,}1)^2}{9{,}1} =$$

$$= 0{,}73 + 1{,}16 + 0{,}31 + 0{,}31 + 0{,}5 + 0{,}13 = \mathbf{3,14}$$

TABLAS DE CONTINGENCIA A DOS CRITERIOS DE CLASIFICACIÓN: MARGINALES LIBRES (CONT.)

TABLA 3

gl \ α	0,995	0,99	0,975	0,95	0,90	0,75	0,50	0,25	0,10	0,05	0,025	0,01	0,005
1	0,000039	0,00016	0,00098	0,0039	0,02	0,10	0,46	1,32	2,71	3,84	5,02	6,63	7,88
2	0,010	0,020	0,051	0,103	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,989	1,239	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,30
8	1,344	1,647	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,10	22,00
9	1,735	2,090	2,70	3,33	4,17	5,90	8,34	11,39					
10	2,160	2,560	3,25	3,94	4,87								

$\chi^2_{H_0} = 3{,}14 \Rightarrow$ con 2 gl, $p_v \cong 0{,}25 > \alpha = 0{,}01 \Rightarrow$ no se rechaza H_0, es decir que el criterio de calidad de los establecimientos es independiente del tamaño, con una significación del 1%.

InfoStat

Tablas de contingencia a dos criterios de clasificación: Marginales Fijos

Suponga que se muestrean, siguiendo el ejemplo anterior, 50 establecimientos pequeños y 50 grandes. Este esquema de muestreo difiere del caso anterior ya que ahora existe un factor de condicionamiento: el tamaño. Antes se tomaba una muestra de 100 establecimientos sin tener en cuenta ninguna de sus características, generando una tabla con marginales libres.

Ahora el muestreo para cada tamaño de establecimiento genera una tabla con **marginales fijos** para las filas, como se muestra a continuación:

	Criterio de calidad			
Tamaño	*Alto*	*Medio*	*Bajo*	**Total**
Pequeño	16	29	5	**50**
Grande	35	10	5	**50**
Total	**51**	**39**	**10**	**100**

Obsérvese que las filas resumen las distribuciones condicionales muestrales del criterio de calidad de los establecimientos, para cada tamaño.

Tablas de contingencia a dos criterios de clasificación: Marginales Fijos (cont.)

El interés es el mismo que en caso anterior, esto es establecer si *la calidad está o no asociada al tamaño*. Reconociendo la generación de la tabla, es decir, cómo es recogida esa información, la hipótesis que se puede verificar es que "*las proporciones de calidad son las mismas para cualquier tamaño*". La prueba para contrastar esta hipótesis se conoce como **prueba χ^2 para la homogeneidad de proporciones**.

La hipótesis nula establece, para este caso, que las distribuciones condicionales de la variable utilizada como criterio columna respecto de aquella utilizada como criterio fila (en este caso, la variable con marginales fijos) son iguales.

$\mathbf{H_0}$: $\pi_{i1} = \cdots = \pi_{ic}$ $(i = 1, \ldots, f)$ o bien $\mathbf{H_0}$: $\pi_{1j} = \cdots = \pi_{fj}$ $(j = 1, \ldots, c)$

vs. $\mathbf{H_1}$: alguna igualdad no se cumple

Tablas de contingencia a dos criterios de clasificación: Marginales Fijos (cont.)

Para la construcción de la variable pivotal se estiman las frecuencias esperadas cuando la H_0 cierta:

$$\chi^2_{H_0} = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} \sim \chi^2_{(f-1)(c-1)}$$

f = nº de filas
c = nº de columnas
fo = frecuencia observada
fe = frecuencia esperada

Frecuencias observadas

	Criterio de calidad			
Tamaño	*Alto*	*Medio*	*Bajo*	**Total**
Pequeño	16	29	5	**50**
Grande	35	10	5	**50**
Total	**51**	**39**	**10**	**100**

Frecuencias esperadas

	Criterio de calidad			
Tamaño	*Alto*	*Medio*	*Bajo*	**Total**
Pequeño	25,5	19,5	5	**50**
Grande	25,5	19,5	5	**50**
Total	**51**	**39**	**10**	**100**

Tablas de contingencia a dos criterios de clasificación: Marginales Fijos (cont.)

En este caso, los grados de libertad del estadístico χ^2 están dados por el producto (*filas* – *1*) por (*columnas* – *1*), suponiendo H_0 cierta; es decir: $(2-1)\times(3-1)=2$

Supongamos que seguimos trabajando con un nivel de significación (α) del 1%. La hipótesis a ser planteada es la siguiente:

H_0: las proporciones del criterio de calidad son las mismos para cualquier tamaño

H_1: las proporciones del criterio de calidad difieren acorde al tamaño

$$\boldsymbol{\chi^2_{H_0}} = \frac{(16-25{,}5)^2}{25{,}5} + \frac{(29-19{,}5)^2}{19{,}5} + \frac{(5-5)^2}{5} + \frac{(35-25{,}5)^2}{25{,}5} + \frac{(10-19{,}5)^2}{19{,}5} + \frac{(5-5)^2}{5} =$$

$$= 3{,}54 + 4{,}63 + 0 + 3{,}54 + 4{,}63 + 0 = \mathbf{16{,}34}$$

TABLAS DE CONTINGENCIA A DOS CRITERIOS DE CLASIFICACIÓN: MARGINALES FIJOS (CONT.)

TABLA 3

gl \ α	0,995	0,99	0,975	0,95	0,90	0,75	0,50	0,25	0,10	0,05	0,025	0,01	0,005
1	0,000039	0,00016	0,00098	0,0039	0,02	0,10	0,46	1,32	2,71	3,84	5,02	6,63	7,88
2	0,010	0,020	0,051	0,103	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,989	1,239	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,30
8	1,344	1,647	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,10	22,00
9	1,735	2,090	2,70	3,33	4,17	5,90	8,34	11,39					
	2,160	2,560	3,25	3,94	4,87								

$\chi^2_{H_0} = 16{,}34 \Rightarrow$ con 2gl, $p_v \cong 0{,}005 < \alpha = 0{,}01 \Rightarrow$ se rechaza H_0, es decir que las proporciones del criterio de calidad difieren acorde al tamaño, al nivel del 1%.

InfoStat

Mediana de Mood

Sabemos, de acuerdo a lo visto, que cuando el CV de una muestra supera el 20%, la media aritmética pierde la representatividad del grupo, al verse afectada –tal vez– por la presencia de valores extremos (outliers) y, en estos casos, utilizábamos la mediana por ser más robusta.

Ahora bien, si queremos comparar dos o más grupos, y (al menos) uno de ellos tiene un CV > 20%, no podemos utilizar una prueba de diferencia de medias o un ANOVA (tema que no se trata en el presente curso).

Luego se calcula la mediana combinada de los valores de todos los grupos, que la simbolizaremos Ψ (Psi). Los datos se disponen en una tabla de contingencia (de $2 \times k$, siendo k el número de grupos a comparar) teniendo en cuenta si pertenecen al grupo $\leq \Psi$ o bien $> \Psi$.

A partir de ello, se trabaja de la misma manera que una prueba Ji-cuadrado de homogeneidad (marginales fijos). Siendo el planteo de hipótesis, el siguiente:

Mediana de Mood (cont.)

$H_0: \Theta_1 = \Theta_2 = \cdots = \Theta_n$ vs. H_1: alguna igualdad no se cumple

Dónde Θ_i (i = 1, 2, …, n) representan las medianas poblacionales.

$$\chi^2_{H_0} = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}} \sim \chi^2_{(f-1)(c-1)}$$

f = número de filas
c = número de columnas
fo = frecuencia observada
fe = frecuencia esperada

Consideración importante

Cuando, y siempre que se utilice la distribución χ^2, más del 20% de las **fe** < 5, se debe utilizar la corrección de Yates o corrección por continuidad y consiste en restarle 0,5 al valor absoluto de la diferencia entre la frecuencia observada y la frecuencia esperada, antes de elevarlo al cuadrado. Es decir, el sumando de la variable pivotal o estadístico de prueba se transforma en:

$$\frac{(|fo - fe| - 0{,}5)^2}{fe}$$

Referencias Bibliográficas

- Bologna, E. (2011). *Estadística para Psicología y Educación*. Córdoba: Brujas.
- Glass, V. & Stanley, J.C. (1996). *Métodos Estadísticos aplicados a las Ciencias Sociales*. México: Prentice–Hall Hispanoamericana, S.A.
- Gorgas García, J., Cardiel López, N. & Zamorano Calvo, J. (2009). *Estadística Básica para Estudiantes de Ciencias*. Madrid: Departamento de Astrofísica y Ciencias de la Atmósfera. Facultad de Ciencias Físicas. Universidad Complutense de Madrid.
- Hernández Sampieri R., Fernández–Collado C. & Baptista Lucio P. (2010). *Metodología de la investigación* (6ª ed.). México: McGraw-Hill Interamericana.
- Penna, F.O., Esteva, G.C., Cobos, O.H. & Ulagnero, C.A. (2018). *Fórmulas y Tablas III (para cursos de Estadística básica)* (2ª ed.). San Luis: Nueva Editorial Universitaria.
- Triola M. (2018). *Estadística* (12ª ed.). México: Pearson Educación.