POBLACIÓN Y MUESTRA

La muestra aleatoria

Una población en estadística es el conjunto de todas las observaciones en las que estamos interesados. Se llama **tamaño de la población** al número de individuos que la componen, siendo cada posible observación un individuo; así pues, las poblaciones pueden ser finitas e infinitas.

Cada observación en una población es un valor de una variable aleatoria **X** con una función de probabilidad o densidad determinada **f(x)** Normalmente, se denomina a las poblaciones con el nombre de la distribución de la variable; es decir, hablaremos de poblaciones normales, binomiales, etc.

Para estudiar una población existen dos posibilidades. Una de ellas consiste en estudiar todos sus elementos y sacar conclusiones; la otra consiste en estudiar sólo una parte de ellos, una muestra, elegidos de tal forma que nos digan algo sobre la totalidad de las observaciones de la población. El mejor método ser el primero, cuando es posible, lo cual sólo ocurre en las poblaciones finitas y razonablemente pequeñas; en el caso de poblaciones muy grandes o infinitas será muy difícil o imposible realizar un estudio total. En este caso necesitaremos tomar una muestra y nos surgirá el problema de cómo hacer para que la muestra nos diga algo sobre el conjunto de la población.

La condición más obvia que se le puede pedir a una muestra es que sea representativa de la población. Está claro que si no conocemos la población no podemos saber si la muestra es representativa o no. La única forma de tener cierta garantía de que esto ocurra es tomar nuestra muestra de forma que cada individuo de la población y cada subgrupo posible de la población tengan igual probabilidad de ser elegidos. A este tipo de muestras se les llama muestras aleatorias o muestras al azar.

Una **muestra aleatoria de tamaño n** es un conjunto de n individuos tomado de tal manera que cada subconjunto de tamaño n de la población tenga la misma probabilidad de ser elegido como muestra; es decir, si la población tiene tamaño N, cada una de las combinaciones posibles de n elementos debe ser equiprobable.

Población de aliens

Muestra de aliens

Los sistemas de muestreo se basan normalmente en la asignación de un número a cada uno de los individuos de la población y la posterior obtención de una muestra de n números aleatorios que se obtendrá por sorteo utilizando bolas numeradas, ordenadores, etc

Otra variante del muestreo es cuando se divide la población en n grupos, que no correspondan con ninguna clasificación relacionada con el problema en estudio, que se ordenan. Por sorteo se elige un elemento del primer grupo y a continuación los elementos correspondientes de los demás grupos. Este tipo de muestra se denomina muestra al azar sistemático.

Si la población está subdividida en grupos podemos tomar otro tipo de muestra en la que cada grupo de la población está representado por un porcentaje de individuos igual al porcentaje de individuos de la población integrados en ese grupo. Este tipo se llama muestra al azar estratificado.

Parámetros y estadísticos

Parámetros poblacionales

Se llama parámetros poblacionales a cantidades que se obtienen a partir de las observaciones de la variable y sus probabilidades y que determinan perfectamente la distribución de esta, así como las características de la población, por ejemplo: La media, μ, la varianza σ^2, la proporción de determinados sucesos, P.

Los Parámetros poblacionales son números reales, constantes y únicos.

Parámetros muestrales

Los Parámetros muestrales son resúmenes de la información de la muestra que nos "determinan" la estructura de la muestra.

Los Parámetros muestrales no son constantes sino variables aleatorias pues sus valores dependen de la estructura de la muestra que no es siempre la misma como consecuencia del muestreo aleatorio. A estas variables se les suele llamar estadísticos.

Los estadísticos se transforman en dos tipos: estadísticos de centralidad y estadísticos de dispersión.

Estadísticos de centralidad:

Son medidas de la tendencia central de la variable. los más conocidos son:

1) La media aritmética

Es el valor esperado de las observaciones de la muestra calculado como si la muestra fuera una variable completa, es decir, multiplicando observaciones por frecuencias y sumando.

Si x_1, x_2,.., x_n representan una muestra de tamaño n de la población, la media aritmética se calcula como:

$$\overline{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

La media aritmética es la medida de la tendencia central que posee menor varianza. Engloba en ella toda la información de la muestra; esto, con ser una ventaja, supone una cierta desventaja pues los valores muy extremos, en muestras pequeñas afectan mucho a la media.

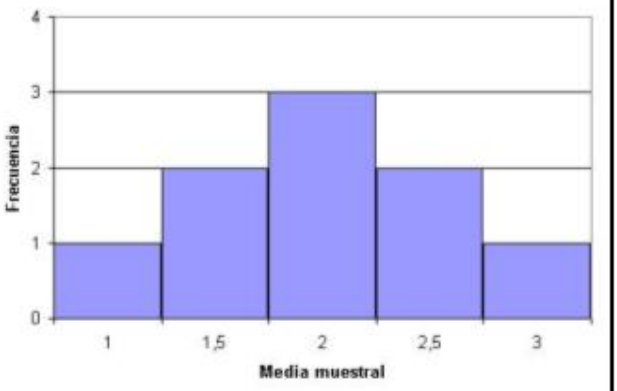
La media de la media aritmética es igual a la de las observaciones (μ) y su varianza es igual a la de las observaciones partida por n. En poblaciones normales, la distribución de la media es normal,

$$\overline{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Si la población no es normal, pero la muestra es grande (n ≥ 30), por el teorema central del límite la distribución de la media será asintóticamente normal.

muestras de tamaño 2		Media muestral
X_1	X_2	
1	1	1
1	2	1,5
1	3	2
2	1	1,5
2	2	2
2	3	2,5
3	1	2
3	2	2,5
3	3	3

Población {1 , 2 , 3}



2) La mediana

En una variable se define como el punto para el cual la función de distribución alcance el valor 0.5; en una muestra la mediana es el valor central.

Para calcularla se ordenan las observaciones de menor a mayor. Si n es impar, la mediana es la observación central

$$\tilde{x} = x_{\frac{n+1}{2}}$$

Si n es par, la mediana se define como la media de las dos observaciones centrales

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

En resumen, podríamos decir que la mediana es el valor que es mayor o igual que el 50% de las observaciones de la muestra y menor o igual que el otro 50%.

No tiene por qué ser igual a una de las observaciones de la muestra.

Es más fácil de calcular que la media aritmética y apenas se afecta por observaciones extremas; sin embargo tiene mayor varianza que X y sólo toma en cuenta la información de los valores centrales de la muestra.

3) La moda

Es el valor más frecuente.

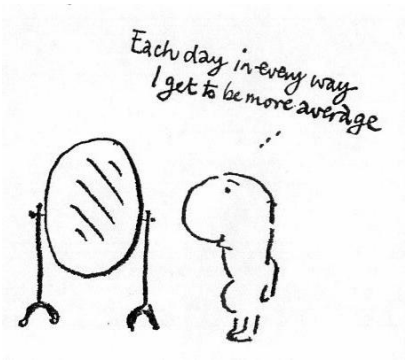
Su cálculo es el más simple de los tres correspondientes a estadísticos de centralidad pero la moda es el estadístico de mayor varianza.

La moda puede no existir y cuando existe no es necesariamente única. No tiene sentido en muestras pequeñas en las que la aparición de coincidencias en los valores es con gran frecuencia más producto del azar que de otra cosa.

La media es el estadístico de centralidad más usado cuando uno espera que la población tenga una distribución más o menos simétrica, sin estar clasificada en grupos claramente diferenciados.

En el caso de distribuciones muy asimétricas, con una cola muy larga, la mediana es, normalmente, el valor de elección dado que la media suele estar desplazada respecto al núcleo principal de observaciones de la variable. En estos casos, la mediana es el valor que mejor expresa el punto donde se acumulan mayoritariamente las observaciones de la variable.

En el caso de poblaciones o muestras subdivididas en grupos claramente definidos la media y la mediana carecen, normalmente, de sentido y los valores que más claramente reflejan el comportamiento de las observaciones de la variable son las modas.



Otros estadísticos de centralidad son los cuantiles.

Los cuantiles o percentiles

Un percentil X, P_X, es un valor de la distribución muestral o poblacional de la variable que es mayor o igual que el X% de las observaciones de la variable $P(Y \leq P_X) = X\%$.

Existe un tipo especial de cuantiles llamados **cuartiles**.

Los cuartiles son tres valores que dividen la distribución en cuatro partes equivalentes porcentualmente.

- El primer cuartil es el valor que es mayor o igual que el 25% de las observaciones de la muestra y menor o igual que el 75%.
- El segundo cuartil es la mediana.
- El tercer cuartil es mayor o igual que el 75% de las observaciones de la muestra y menor o igual que el 25%.

Estadísticos de dispersión

Los estadísticos de dispersión son parámetros muestrales que expresan la dispersión de los valores de la variable respecto al punto central, es decir, su posición relativa. Los más importantes son:

El rango

Es la diferencia entre las dos observaciones extremas, la máxima menos la mínima. Expresa cuantas unidades de diferencia podemos esperar, como máximo, entre dos valores de la variable.

El rango estima el campo de variación de la variable.

Se afecta mucho por observaciones extremas y utiliza únicamente una pequeña parte de la información.

La varianza

Es la desviación cuadrática media de las observaciones a la media muestral.

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Su concepto es análogo al de la varianza poblacional. No obstante esta expresión de cálculo de la varianza muestral no se utiliza mucho pues sus valores tienden a ser menores que el de la auténtica varianza de la variable (debido a que la propia media muestral tiene una varianza que vale un enésimo de la de las observaciones) Para compensar esta deficiencia y obtener valores que no subestimen la varianza poblacional (cuando estamos interesados en ella y no en la varianza muestral) utilizaremos una expresión, esencialmente igual que la anterior salvo que el denominador está disminuido en una unidad.

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Normalmente, estaremos interesados en saber cosas acerca de la varianza poblacional y no de la varianza muestral. Por tanto, en adelante, cuando hablemos de varianza muestral, salvo indicación expresa, nos referiremos a la segunda.

Es el estadístico de dispersión más usado por las propiedades de su distribución. Si la población de la que procede la muestra es normal:

$\frac{(n-1)s^2}{\sigma^2} \approx \chi^2$ con n-1 grados de libertad.

Además, utiliza toda la información de la muestra.

Su mayor inconveniente consiste en que se expresa en unidades cuadráticas. Por ello, para muchos propósitos se utiliza otro estadístico de dispersión que la desviación típica.

Si no disponemos de una calculadora, el cálculo de la varianza puede ser complicado porque, habitualmente, los valores de las desviaciones de las observaciones a la media resultan ser números con varias cifras decimales. Por ello, se suele utilizar una ecuación que deriva directamente de la anterior:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$$

o, alternativamente, la equivalente a aquella de "la media de los cuadrados menos el cuadrado de la media".

$$s^2 = \left(\overline{X_i^2} - \overline{X_i}^2\right) \cdot \frac{n}{n-1}$$

La desviación típica

Es la raíz cuadrada positiva de la varianza y, por tanto, se expresa en las unidades de medida de la variable.

$$s = +\sqrt{s^2}$$

Su concepto es análogo al de la desviación típica poblacional.

Coeficiente de variación

Es el cociente entre la desviación típica y la media aritmética muestrales y expresa la variabilidad de la variable en tanto por uno, sin dimensiones.

$$C.V. = \frac{s}{\overline{x}} \quad \text{o bien} \quad C.V. = \frac{s}{\overline{x}} \cdot 100$$

Permite comparar muestras de variables de distinta naturaleza o muestras de la misma variable en poblaciones en las que el orden de magnitud de las observaciones sea muy diferente.

Pruebas chi-cuadrado de ajuste e independencia

Las pruebas chi-cuadrado son un grupo de contrastes de hipótesis que sirven para comprobar afirmaciones acerca de las funciones de probabilidad (o densidad) de una o dos variables aleatorias.

Estas pruebas no pertenecen propiamente a la estadística paramétrica pues no establecen suposiciones restrictivas en cuanto al tipo de variables que admiten, ni en lo que refiere a su distribución de probabilidad ni en los valores y/o el conocimiento de sus parámetros.

Se aplican en dos situaciones básicas:

a) Cuando queremos comprobar si una variable, cuya descripción parece adecuada, tiene una determinada función de probabilidad. La prueba correspondiente se llama chi-cuadrado de ajuste.

b) Cuando queremos averiguar si dos variables (o dos vías de clasificación) son independientes estadísticamente. En este caso la prueba que aplicaremos ser la chi-cuadrado de independencia o chi-cuadrado de contingencia.

Chi-cuadrado de ajuste

En una prueba de ajuste la hipótesis nula establece que una variable X tiene una cierta distribución de probabilidad con unos determinados valores de los parámetros. El tipo de distribución se determina, según los casos, en función de: La propia definición de la variable, consideraciones teóricas al margen de esta y/o evidencia aportada por datos anteriores al experimento actual.

A menudo, la propia definición del tipo de variable lleva implícitos los valores de sus parámetros o de parte de ellos; si esto no fuera así dichos parámetros se estimarán a partir de la muestra de valores de la variable que utilizaremos para realizar la prueba de ajuste.

Como en casos anteriores, empezaremos definiendo las hipótesis.

Hipótesis nula: X tiene distribución de probabilidad f(x) con parámetros y_1,..., y_p

Hipótesis alternativa: X tiene cualquier otra distribución de probabilidad.

Es importante destacar que el rechazo de la hipótesis nula no implica que sean falsos todos sus aspectos sino únicamente el conjunto de ellos; por ejemplo, podría ocurrir que el tipo de distribución fuera correcto pero que nos hubiésemos equivocado en los valores de los parámetros.

Obviamente, necesitaremos una muestra de valores de la variable X. Si la variable es discreta y tiene pocos valores posible estimaremos las probabilidades de dichos valores mediante sus frecuencias muestrales; si la variable es continua o si es una discreta con muchos o infinitos valores estimaremos probabilidades de grupos de valores (intervalos).

Metodológicamente, la prueba se basa en la comparación entre la serie de frecuencias absolutas observadas empíricamente para los valores de la variable (**Oi**) y las correspondientes frecuencias absolutas teóricas obtenidas en base a la función de probabilidad supuesta en la hipótesis nula (**Ei**).

Así pues, una vez calculadas las frecuencias absolutas de cada valor o intervalo de valores, obtendremos el número total de observaciones de la muestra (T) sumando las frecuencias observadas

$$T = \sum_i O_i$$

Para calcular las frecuencias esperadas repartiremos este número total de observaciones (T) en partes proporcionales a la probabilidad de cada suceso o grupo de sucesos. Para ello calcularemos dichas probabilidades utilizando la función de probabilidad definida en la hipótesis nula f(x), de modo que, cada valor Ei tendrá la siguiente expresión:

$$E_i = f(x_i) \cdot T$$

Por tanto, tendremos los siguientes datos para la prueba:

Valor de la variable	x_1 x_2 x_3 ... x_i ... x_k
Frecuencias observadas	O_1 O_2 O_3 ... O_i ... O_k
Frecuencias esperadas	E_1 E_2 E_3 ... E_i ... E_k

Si la hipótesis nula es cierta, las diferencias entre valores observados y esperados (que siempre existirán por tratarse de una muestra aleatoria) son atribuibles, exclusivamente, al efecto del azar. En estas condiciones, se puede calcular un parámetro que depende de ambos, cuya distribución se ajusta a una chi-cuadrado.

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \approx \chi^2$$

Si, por el contrario, la hipótesis nula fuera falsa los Ei ya no serían, realmente, los valores esperados de las frecuencias; por tanto, las diferencias entre los valores "esperados" y los observados reflejarían no sólo el efecto del azar sino también las diferencias entre los Ei y la auténtica serie de valores esperados (desconocida) Como consecuencia, las diferencias de los numeradores de la expresión anterior tienden a ser más grandes y, por estar elevadas al cuadrado, la suma de cocientes ser positiva y mayor que lo que se esperaría para los valores de una chi-cuadrado.

$$\sum_i \frac{(O_i - E_i)^2}{E_i} >> \chi^2$$

Por tanto, el parámetro anterior será el estadístico de contraste de la prueba de hipótesis y la región crítica se encontrar siempre en la cola derecha de la distribución chi-cuadrado. Evidentemente, esta prueba será siempre de una sola cola.

Estadístico de contraste $\chi^{2^*} \approx \sum_i \frac{(O_i - E_i)^2}{E_i}$

Se acepta la hipótesis nula si $\chi^{2^*} < \chi^2_{1-\alpha,\nu}$, el percentil 1 – α de la distribución chi-cuadrado con ν grados de libertad.

Cabe señalar que en las pruebas chi-cuadrado lo corriente es que pretendamos comprobar que una variable tiene una cierta distribución y, por tanto, habitualmente, nos vemos obligados a colocar nuestra propia hipótesis en la hipótesis nula. Únicamente podremos colocar nuestra hipótesis en la alternativa en el caso excepcional de que pretendamos demostrar que cierto tratamiento produce una distorsión de la distribución básica de la variable en estudio.

El número de grados de libertad de la variable chi-cuadrado se calcula de la siguiente forma:

- A priori, tendrá tantos grados de libertad como parejas frecuencia observada - frecuencia esperada.
- A esta cantidad se debe restar el número de restricciones lineales impuestas a las frecuencias observadas, es decir, el número de parámetros que es necesario calcular **directamente** a partir de los valores observados para establecer los valores esperados. Este número es, como mínimo, uno ya que siempre tendremos que calcular el número total de observaciones de la muestra.

Una condición básica para que podamos llevar a cabo una prueba chi-cuadrado es que las frecuencias de las distintas clases deben ser suficientemente altas como para garantizar que pequeñas desviaciones aleatorias en la muestra no tengan importancia decisiva sobre el valor del estadístico de contraste.

Las reglas que determinan cuando es posible o no realizar el contraste varían mucho de unos autores a otros. En un extremo de máxima rigidez se encuentran aquellos que opinan que no se puede realizar la prueba cuando alguna de las frecuencias, observadas o esperadas, sea menor que 5. En el otro extremo se encuentran quienes opinan que, para que la prueba sea viable ninguna de las frecuencias esperadas debe ser menor que 1 y no más del 25% pueden ser menores que 5; en lo que refiere a las frecuencias observadas no existirían límites. La autora de este texto simpatiza más con la segunda postura, no sólo por razones prácticas, sino porque lo razonable es que la distribución esperada esté adecuadamente definida y, por tanto, no debe incluir valores muy bajos; sin embargo, los valores extremos en la distribución observada simplemente reflejan diferencias importantes entre la distribución supuesta por la hipótesis nula y la real.

Sea cual sea el criterio que elijamos, si resultara que la prueba no es viable podríamos recurrir a

englobar los valores o clases de valores con sus vecinos más próximos y pasar así a engrosar sus frecuencias. Este procedimiento no puede llevarse hasta el absurdo pero proporciona una salida digna a situaciones complejas. En casos excepcionales se pueden englobar valores que no sean vecinos porque exista algún nexo lógico de conexión entre ellos.

Cuando sea necesario agrupar valores, los grados de libertad no se deben calcular hasta que tengamos establecidas definitivamente las parejas de frecuencias observadas y esperadas con las que calcularemos el estadístico de contraste.

Chi-cuadrado de contingencia o independencia

La prueba chi-cuadrado de contingencia sirve para comprobar la independencia de frecuencias entre dos variables aleatorias, X e Y.

Las hipótesis contrastadas en la prueba son:

Hipótesis nula: X e Y son independientes.

Hipótesis alternativa: X e Y no son independientes (No importa cual sea la relación que mantengan ni el grado de esta.

La condición de independencia, tal como fue definida en la página anterior era: X e Y son independientes si y sólo si para cualquier pareja de valores x e y la probabilidad de que X tome el valor x e Y el valor y, simultáneamente, es igual al producto de las probabilidades de que cada una tome el valor correspondiente.

$$X \text{ e } Y \text{ son independientes} \iff \forall x,y \;\; f(x,y) = f(x)\cdot f(y)$$

Por tanto, todo lo que necesitamos serán unas estimas de las funciones de probabilidad de ambas variables por separado (f(x) y f(y)) y de la función de probabilidad conjunta (f(x,y))

Empezaremos la prueba tomando una muestra de parejas de valores sobre la que contaremos la frecuencia absoluta con la que aparece cada combinación de valores (x_i,y_j) o de grupos de valores (i,j) (O_{ij}) La tabla siguiente, en la que se recogen estos datos, es en realidad nuestra estimación de la función de probabilidad conjunta multiplicada por el número total de datos (T).

X \ Y	y_1	y_2	....	y_i	....	y_J	$F_i = \sum_i O_{ij}$
x_1	O_{11}	O_{12}	....	O_{1i}	....	O_{1J}	F_1
x_2	O_{21}	O_{22}	....	O_{2i}	....	O_{2J}	F_2
....	....	....	....	....	....	....	....
x_i	O_{i1}	O_{i2}	....	O_{ii}	....	O_{iJ}	F_i
...	....	....	....	....	....	....	....
x_I	O_{I1}	O_{I2}	....	O_{Ii}	....	O_{IJ}	F_I
$C_j = \sum_j O_{ij}$	C_1	C_2	....	C_i	....	C_J	T

Para obtener las estimas de las funciones de probabilidad marginales debemos sumar por filas y por columnas los valores de las frecuencias conjuntas. Las sumas de filas (F_i) son, en cada caso, el número de

veces que hemos obtenido un valor de X (x_i) en cualquier combinación con distintos valores de Y, es decir, son nuestra estima de la función de probabilidad de X multiplicada por el número total de observaciones; análogamente, las sumas de columnas (C_j) son nuestra estima de la función de probabilidad de Y multiplicada por el número total de observaciones.

El número total de observaciones lo podemos obtener como la suma de todas las frecuencias observadas o, también, como la suma de las sumas de filas o de las sumas de columnas:

$$T = \sum_{ij} O_{ij} = \sum_{i} F_i = \sum_{j} C_j$$

Así pues, si las variables fueran independientes debería cumplirse que

$$\forall i,j \qquad \frac{O_{ij}}{T} = \frac{F_i}{T} \cdot \frac{C_j}{T} = \frac{F_i \cdot C_j}{T^2}$$

Naturalmente, nadie espera que esta condición se cumpla exactamente debido al efecto de los errores de muestreo aleatorio. Por tanto, nuestro problema consiste en distinguir entre las diferencias producidas por efecto del muestreo y diferencias que revelen falta de independencia.

Podemos convertir la ecuación anterior a frecuencias absolutas multiplicando por T:

- Si X e Y son independientes, O_{ij} debe ser igual a $\frac{F_i \cdot C_j}{T}$ y, por tanto,
- bajo la hipótesis de independencia, $\frac{F_i \cdot C_j}{T}$ es el valor esperado de O_{ij} (E_{ij})

Tal como pasaba en la prueba anterior, si las variables son independientes, es decir, si las frecuencias E_{ij} son realmente los valores esperados de las frecuencias O_{ij}, se puede calcular un parámetro que depende de ambas que tiene distribución chi-cuadrado,

$$\sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2$$

Por otra parte, si las variables no son independientes, las diferencias entre las series de frecuencias observadas y esperadas serán mayores que las atribuibles al efecto del azar y, al estar elevadas al cuadrado en el numerador de la expresión anterior, ésta tenderá a ser mayor que lo que suele ser el valor de una variable chi-cuadrado.

$$\sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} >> \chi^2$$

Por tanto, el parámetro anterior ser el estadístico de la prueba de hipótesis y la región crítica se encontrar siempre en la cola derecha de la distribución chi-cuadrado. Nuevamente, esta prueba será siempre de una sola cola.

Estadístico de contraste $\chi^{2^*} \approx \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Se acepta la hipótesis nula si $\chi^{2^*} < \chi^2_{1-\alpha,\nu}$, el percentil 1 – α de la distribución chi-cuadrado con ν grados de libertad.

Tal como ocurría en la prueba anterior lo corriente es que queramos demostrar que dos variables son independientes, es decir, que, habitualmente, nos veremos obligados a colocar nuestra hipótesis en la hipótesis nula.

El número de grados de libertad de la chi-cuadrado que sirve de contraste se calcula de la siguiente forma:

- A priori tendremos tantos grados de libertad como combinaciones de valores x_i, y_j tengamos (I J)
- A este número tendremos que restarle I debido a que, para calcular las frecuencias esperadas, necesitamos calcular las I sumas de filas en la tabla anterior. Conocidas las sumas de filas obtenemos el número total de observaciones sin perder ningún grado de libertad.
- A continuación, necesitaremos calcular, a partir de las frecuencias observadas J - 1 de las sumas de columnas; la restante podemos obtenerla restando la suma de las anteriores del total de observaciones (T).

En resumen, el número de grados de libertad de la prueba es el producto del número de filas menos uno por el número de columnas menos uno.

$$\nu = I\,J - I - (J-1) = IJ - I - J + 1 = (I-1)(J-1)$$

En cuanto a la magnitud mínima necesaria de las frecuencias observadas y esperadas, rigen las mismas normas que en el caso de la prueba de ajuste. En este caso, si nos viéramos obligados a juntar valores para sumar frecuencias, debemos unir columnas o filas completas (y contiguas). Obviamente, los grados de libertad no deben calcularse hasta que no se hayan realizado todas las agrupaciones necesarias y quede claro cual es el número de filas y columnas de la tabla definitiva.

Como hemos visto, esta prueba no hace ninguna suposición acerca del tipo de distribución de ninguna de las variables implicadas y utiliza únicamente información de la muestra, es decir, información contingente. Esta es la razón por la que, habitualmente, se le llama chi-cuadrado de contingencia.

Comparación múltiple de distintas proporciones o probabilidades

Una aplicación concreta de la chi-cuadrado de independencia es la comparación múltiple de las distintas proporciones o probabilidades de un suceso en I poblaciones diferentes.

Supongamos que tenemos I poblaciones en las cuales las observaciones se pueden clasificar como A o no-A. Llamemos P_i a la probabilidad del suceso A en cada población i y P a la frecuencia media de A en el conjunto de las poblaciones; la probabilidad del suceso no-A en cada población i ser 1 - Pi y la media de todas ellas valdrá 1 - P.

Las hipótesis de la prueba serán:

Hipótesis nula: $\forall i,\ P_i = P$

Hipótesis alternativa: $\exists i,j \mid P_i \neq P_j$

Si tomamos una muestra de tamaño n_i en cada población y contamos en cada caso el número de sucesos A aparecidos en la muestra obtendríamos la siguiente tabla:

muestra / suceso	1	2	....	l	
A	x_1	x_2	....	x_l	$\sum_i x_i$
no A	$n_1 - x_1$	$n_2 - x_2$	....	$n_l - x_l$	$\sum_i (n_i - x_i)$
	n_1	n_2	....	n_l	$\sum_i n_i$

Esta es una tabla típica a la que se puede aplicar la metodología de la prueba chi-cuadrado de independencia. Veamos como corresponden las hipótesis de una y otra prueba. Si la clasificación de las observaciones en sucesos A y no-A fuera independiente de la clasificación en muestras, la frecuencia relativa de A (y la de no-A) serían iguales en todos los casos y los valores esperados de las frecuencias absolutas se calcularían multiplicando la estima común de la frecuencia relativa global por el número de observaciones en cada muestra.

La estima global de la frecuencia de A se hallara dividiendo el número total de sucesos A por el número total de observaciones:

$$\hat{P} = \frac{\sum_i x_i}{\sum_i n_i}$$

lo cual no es otra cosa que el cociente entre la suma de la fila uno (F_1) y el total de observaciones (T)

Por tanto, el valor esperado de la frecuencia observada de A en la muestra i ($E_{A,i}$) será:

$$E_{A,i} = \hat{P} \cdot n_i = \frac{\sum_i x_i}{\sum_i n_i} \cdot n_i = \frac{F_1}{T} \cdot C_i$$

La estima global de la frecuencia de no-A se hallara dividiendo el número total de sucesos no-A por el número total de observaciones:

$$\hat{P} = \frac{\sum_i (n_i - x_i)}{\sum_i n_i}$$

lo cual no es otra cosa que el cociente entre la suma de la fila dos (F_2) y el total de observaciones (T)

Por tanto, el valor esperado de la frecuencia observada de no-A en la muestra i ($E_{no\text{-}A,i}$) será:

$$E_{no-Ai} = \left(1 - \hat{P}\right) \cdot n_i = \frac{\sum_i (n_i - x_i)}{\sum_i n_i} \cdot n_i = \frac{F_2}{T} \cdot C_i$$

Es decir, los valores esperados se calcularían, en pura lógica, tal como indica el procedimiento estándar de la prueba de contingencia. En definitiva:

Hipótesis nula: $\forall i,\ P_i = P \Leftrightarrow$ **La clasificación en sucesos es independiente de la clasificación en poblaciones.**

Hipótesis alternativa: $\exists i,j \mid P_i \neq P_j \Leftrightarrow$ **La clasificación en sucesos no es independiente de la clasificación en poblaciones.**

En resumen, la prueba de comparación múltiple de proporciones se realizar mediante una prueba de contingencia que nos dirá si las probabilidades son todas iguales o si, al menos, existe una que sea diferente de las demás.

Los grados de libertad serán siempre:

$$\nu = (2-1)(I-1) = I-1$$

Prueba de homogeneidad de muestras

Otra de las aplicaciones interesantes de la prueba chi-cuadrado de independencia consiste en la comprobación de la homogeneidad de distintas muestras de una variable.

Supongamos que hemos obtenido J muestras de tamaño n_j de una misma variable aleatoria (X) y queremos comprobar si son homogéneas, es decir, si la variable tiene la misma distribución de probabilidad en todas ellas, bien para utilizarlas conjuntamente, bien porque se trate de identificar diferencias entre las poblaciones de procedencia de las distintas muestras. Las frecuencias observadas serán las de la tabla siguiente, en la que F_i es la frecuencia absoluta total del valor x_i y T es el número total de observaciones

$$T = \sum_i x_i$$

X \ muestra	1	2	....	j	....	J	$F_i = \sum_i O_{ij}$
x_1	O_{11}	O_{12}	....	O_{1j}	....	O_{1J}	F_1
x_2	O_{21}	O_{22}	....	O_{2j}	....	O_{2J}	F_2
....	....	....	....	....	....	....	....
x_i	O_{i1}	O_{i2}	....	O_{ij}	....	O_{iJ}	F_i
...	....	....	....	....	....	....	....
x_I	O_{I1}	O_{I2}	....	O_{Ij}	....	O_{IJ}	F_I
$C_j = \sum_j O_{ij}$	n_1	n_2	....	n_j	....	n_J	T

El razonamiento en este caso es idéntico al anterior. Si las muestras son homogéneas, se puede obtener una estima conjunta de la frecuencia de cada valor x_i (F_i / T) y el valor esperado de la frecuencia absoluta de x_i en cada muestra se calcular como el producto de dicha frecuencia por el tamaño de la muestra correspondiente

$$E_{ij} = \frac{F_i}{T} \cdot n_j = \frac{F_i \cdot C_j}{T}$$

Así pues, las hipótesis de la prueba serán:

Hipótesis nula: Las muestras son homogéneas ⇔ La clasificación de las observaciones según los valores de la variable es independiente de la clasificación en muestras.

Hipótesis alternativa: Las muestras no son homogéneas. ⇔ La clasificación de las observaciones según los valores de la variable no es independiente de la clasificación en muestras.

Obviamente, la prueba se realizar según la metodología habitual.

En este caso, a la prueba chi-cuadrado de contingencia se le suele llamar chi-cuadrado de homogeneidad.