Tema 9

INTRODUCCIÓN A LA INFERENCIA NO PARAMÉTRICA

Hasta ahora se suponía conocida la distribución de la variable bajo estudio salvo algún parámetro y, por tanto, se aplicaban conceptos de inferencia paramétrica. No obstante, en la práctica, no se suele conocer la forma funcional de la distribución y, en dicho caso, no se puede hacer inferencia paramétrica.

Por ello es necesario poder obtener conclusiones sobre la distribución de la variable aleatoria a partir de las observaciones sin conocer la forma funcional, lo cual es uno de los problemas que se estudian dentro de la inferencia no paramétrica. Aunque no se conozca la forma funcional si se suele tener información de tipo general sobre la variable: si es discreta o continua, simetría, curtosis, etc.

Aunque existen procesos de estimación no paramétricos, en este tema se van a estudiar solamente cuestiones de contrastes de hipótesis.

9.1. Problemas de bondad de ajuste

El problema de bondad de ajuste trata de decidir, en base a la información que proporciona una m.a.s. de una variable aleatoria, si se puede admitir que la distribución de la variable es una concreta (ejemplo: $\mathcal{N}(0,1), \exp(3)$) o bien si pertenece aun cierto tipo de distribuciones (ejemplo: normal, exponencial). Es decir, es un problema de bondad de ajuste de los datos observados a una distribución especificada.

En particular se va a contrastar si una muestra proviene de una población con una función de distribución específica, F_0, frente a que dicha función de distribución sea diferente ($F(x) \neq F_0(x)$ para algún $x \in \mathbb{R}$).

Las dos soluciones más frecuentes para resolver este contraste son:

- Test χ^2 de Pearson (1900). Es el primero que surge históricamente y se puede aplicar a variables de tipo discreto, continuo y cualitativo. Este test lo único que tiene en cuenta es una clasificación de las observaciones muestrales en distintas categorías.
- Test de Kolmogorov-Smirnov. Este test se basa en el teorema de Glivenko-Cantelli.

9.1.1. Test χ^2 de Pearson

Sea $(X_1, \ldots, X_n)$ una m.a.s. de una variable aleatoria X que se distribuye según una función de distribución F que es completamente desconocida.

A) Hipótesis nula simple: $\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$

Para resolver este problema se parte el recorrido de la función teórica correspondiente a F_0 en k subconjuntos $A_1, \ldots, A_k$ de probabilidad no nula y se consideran las siguientes probabilidades $p_i^0 = P_{F_0}[X \in A_i] > 0$, $i = 1, \ldots, k$. Sea N_i el número de observaciones muestrales en cada A_i, $i = 1, \ldots, k$. Entonces el estadístico para el contraste es

$$\chi^2(N_1, \ldots, N_k) = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \left(= -n + \sum_{i=1}^{k} \frac{N_i^2}{np_i^0} \right)$$

donde np_i^0 es el número de observaciones muestrales que cabría esperar en A_i si H_0 es cierta.

Este estadístico es una medida de la discrepancia entre las observaciones reales en cada clase y el número que debería de haber si H_0 fuera cierta.

Pearson demostró que bajo H_0 este estadístico tiene distribución asintótica,

$$\chi^2(X_1, \ldots, X_n) \rightsquigarrow_{n\to\infty} \chi^2(k-1).$$

El **Test asintótico para** $H_0 : F = F_0$, de tamaño α, es:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \chi^2(N_1, \ldots, N_k) \geq \chi^2_{k-1;\alpha} \\ 0 & \chi^2(N_1, \ldots, N_k) < \chi^2_{k-1;\alpha} \end{cases}$$

con

$$p - valor = P_{H_0}[\chi^2(N_1, \ldots, N_k) \geq \chi^2_{exp}] \approx_{n\to\infty} P[\chi^2(k-1) \geq \chi^2_{exp}]$$

siendo χ^2_{exp} el valor de estadístico en la muestra observada. Entonces, se rechaza H_0 si el $p - valor$ es menor o igual que α.

Notas:

- Este test, como ya se ha especificado, es un test asintótico, por lo tanto, habrá que especificar cómo tiene que ser n para poder usar el test. Usualmente se considera como restricción para la aplicación del test que $np_i^0 \geq 5$, $i = 1, \ldots, k$.
- Si se plantean como hipótesis nula $H_0 : F = F_0$ ó $H_0' : F = F_0'$, y tanto F_0 como F_0' asignan la misma probabilidad a todos los A_i, $i = 1, \ldots, k$, este test no distingue entre F_0 y F_0'. Para solucionar este problema se considerarán al menos 5 clases. En tal caso $\exists i : p_i^0 \leq 1/5 \Rightarrow n \geq 25$. (Para hacer particiones "buenas" se pueden usar los percentiles).
- Este test es aplicable a cualquier tipo de variable cuyos valores puedan clasificarse en un número finito de categorías pero es más apropiado para variables cualitativas ya que ellas son propiamente categóricas.

B) Hipótesis nula compuesta:

En muchos casos, la hipótesis nula H_0 no especifica una única distribución F_0, sino una familia de distribuciones posibles (p.e., una normal con parámetros desconocidos, etc), dependientes de uno o varios parámetros. En dicho caso no se puede aplicar directamente el test χ^2. Será necesario tener una estimación previa de los parámetros. Por tanto:

i) Primero se estiman los parámetros de la familia especificada en H_0, usualmente por máxima verosimilitud.

ii) Después se aplica el test con los parámetros ya estimados.

La distribución del estadístico del contraste depende de cómo se hayan obtenido las estimaciones. En particular, si los parámetros se estiman partiendo de observaciones independientes de las que se van a usar para el problema de contraste, se usa χ^2 igual que antes. Sin embargo, si se usan para el contraste los mismos datos que para la estimación, los $\widehat{p}_i$ dependen de las observaciones, y la distribución del estadístico bajo H_0 varía:

$$\widehat{\chi}^2(N_1, \ldots, N_k) = \sum_{i=1}^{k} \frac{(N_i - n\widehat{p}_i)^2}{n\widehat{p}_i} \longrightarrow \chi^2(k - q - 1)$$

donde q es el número de parámetros estimados.

Este test tiene una serie de inconvenientes:

- Es un test asintótico y, por tanto, aproximado.

- No trata los datos individualmente, sino por categorías. Por tanto, no usa toda la información contenida en la muestra. Por ello no es un buen test para variables aleatorias continuas.

El siguiente test es un test exacto, no asintótico, válido para variables aleatorias continuas y trata todos los datos de forma individual.

Ejemplo: Se recoge una muestra aleatoria simple de 30 tornillos producidos por cierta máquina y se mide su longitud, obteniéndose:

10.39	10.66	10.12	10.32	10.25	10.52	10.83	10.72	10.28	10.35
10.46	10.54	10.23	10.18	10.62	10.49	10.61	10.64	10.29	10.78
10.81	10.34	10.75	10.41	10.53	10.31	10.47	10.43	10.57	10.74

Contrastar si estos datos avalan que la distribución de la longitud de los tornillos es normal.

9.1.2. Test de Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov se basa en el teorema de Glivenko-Cantelli que, como ya se estudió, proporciona la convergencia casi segura uniformemente de la función de distribución muestral o empírica ($F^*_{X_1,\ldots,X_n}$) a la función de distribución de la variable aleatoria (F).

Para resolver el problema planteado se considera $(X_1, \ldots, X_n)$ una m.a.s. de una variable aleatoria X continua que se distribuye según una función de distribución F que es completamente desconocida. El contraste a resolver es

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

Para resolver este problema se usa el estadístico de Kolmogorov-Smirnov

$$D(X_1, \ldots, X_n) = \sup_{x \in \mathbb{R}} |F^*_{X_1,\ldots,X_n}(x) - F_0(x)|$$

el cual proporciona una medida de la discrepancia entre $F^*_{X_1,\ldots,X_n}$ y F_0. Por tanto, teniendo en cuenta que la distribución muestral converge uniformemente a la distribución teórica, se rechazará la hipótesis nula, si el valor de $D(X_1, \ldots, X_n)$ es grande. Es decir, el test de Kolmogorov-Smirnov sería:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & D(X_1, \ldots, X_n) \geq d_\alpha \\ 0 & D(X_1, \ldots, X_n) < d_\alpha \end{cases}$$

donde d_α verifica:

$$P_{H_0}\left(D(X_1, \ldots, X_n) \geq d_\alpha\right) = \alpha$$

y

$$p - valor = P_{H_0}[D(X_1, \ldots, X_n) \geq D_{exp}]$$

siendo D_{exp} el valor de estadístico en la muestra observada.

Teorema: Si F_0 es continua:

(a) La distribución de $D(X_1, \ldots, X_n)$ es independientes de F_0.

(b) $D(X_1, \ldots, X_n) \rightsquigarrow_{H_0} Z$ de Kolmogorov.

(c) Si las n observaciones son distintas, entonces

$$D_{exp} = \text{máx}\{\max_{x_i}[F^*_{X_1\ldots,X_n}(x_i) - F_0(x_i)], \max_{x_i}[F_0(x_i) - F^*_{X_1,\ldots,X_n}(x_{i^-})]\}$$

Notas:

1. Existe otra expresión de $D(X_1, \ldots, X_n)$ para observaciones iguales, aunque esto es poco probable por ser la distribución continua. Si ocurriera, por redondeos, se eliminan del estudio los elementos iguales para asegurarse que las observaciones sean distintas.

2. Si la distribución con la que se quiere comparar no está totalmente determinada, al igual que en el test χ^2, se pueden estimar los parámetros de la distribución, lo cual varía la distribución de $D(X_1, \ldots, X_n)$. Hay modificaciones del test de Kolmogorov-Smirnov, como el test de Lilliefors, para estos casos. Sin embargo, otra opción aceptable es usar el test de la χ^2 para dichos casos.

Ejemplo: Se supone que el tiempo de reacción a un determinado compuesto se distribuye según una $\mathcal{N}(10.5; 0.15^2)$. Contrastar si los siguientes datos, obtenidos en un muestreo aleatorio simple de 10 individuos a los que se ha administrado el compuesto, proporcionan evidencia para rechazar esta hipótesis:

10.39 10.66 10.12 10.32 10.25 10.52 10.83 10.72 10.28 10.35.

9.2. Problema de localización

Se van a usar tests de localización para resolver problemas de contrastes de hipótesis relativos a medidas de posición (mediana o cuantiles en general). En concreto, los posibles contrastes sobre la mediana son:

$$\begin{cases} H_0 : M_X = m \\ H_1 : M_X \neq m \end{cases} \qquad \begin{cases} H_0 : M_X = m \\ H_1 : M_X > m \end{cases} \qquad \begin{cases} H_0 : M_X = m \\ H_1 : M_X < m \end{cases}$$

y se van a utilizar dos tests para resolverlos:

- Test de los signos de Fisher: se aplica sólo a variables aleatorias de tipo continuo y se generaliza fácilmente a contraste de hipótesis sobre cuantiles.
- Test de rangos signados de Wilcoxon: usa más información muestral que el de Fisher pero exige, además de continuidad, simetría en la distribución.

9.2.1. Test de los signos de Fisher

Sea $X_1, \ldots, X_n$ m.a.s. de $X \sim F$ continua (desconocida).

Idea intuitiva: Es de esperar que, si H_0 es cierta, aproximadamente la mitad de los valores muestrales queda por encima de m y la otra mitad por debajo (concuerda con la idea de convergencia de cuantiles muestrales a cuantiles poblacionales), también con el hecho de que $F(M_X) = 1/2$.

Se puede definir el estadístico del test de los signos:

$$\begin{aligned} T(X_1, \ldots, X_n) &= \text{número de observaciones muestrales mayores que } m \\ &= \text{n}^{\text{o}} \text{ de signos positivos en } (X_i - m) \rightsquigarrow_{H_0} B(n, 1/2) \end{aligned}$$

Las regiones crítica del test, según el contraste plateado son:

- Para $H_1 : M_X > m$, se rechaza H_0 para

$$T_{exp}(= T(x_1, \ldots, x_n)) \geq k : \; P_{H_0}[T(X_1, \ldots, X_n) \geq k] \leq \alpha.$$

  Otra opción es calcular el $p - valor = P_{H_0}[T(X_1, \ldots, X_n) \geq T_{exp}]$.

- Para $H_1 : M_X < m$, se rechaza H_0 para

$$T_{exp} \leq k : \; P_{H_0}[T(X_1, \ldots, X_n) \leq k] \leq \alpha.$$

  Otra opción es calcular el $p - valor = P_{H_0}[T(X_1, \ldots, X_n) \leq T_{exp}]$.

- Para $H_1 : M_X \neq m$, se rechaza H_0 para

$$T_{exp} \leq k \text{ ó } T_{exp} \geq n - k : \; P_{H_0}[T(X_1, \ldots, X_n) \leq k] \leq \alpha/2.$$

  Otra opción es calcular el

$$p - valor = \begin{cases} 2P_{H_0}[T(X_1, \ldots, X_n) \leq T_{exp}] & \text{si } T_{exp} \leq n/2 \\ 2P_{H_0}[T(X_1, \ldots, X_n) \geq T_{exp}] & \text{si } T_{exp} \geq n/2 \end{cases}$$

Notas:

1. Este test se puede aleatorizar.

2. Si algún valor de la muestra coincide con m, $(X_i - m = 0)$, dicho dato se elimina y se reajusta el tamaño de la muestra, n.

3. Para n grande $(n \geq 20)$ se puede emplear la aproximación normal a la distribución binomial para determinar los puntos críticos (k), pero sería un test asintótico:

$$B(n,p) \approx N(np, npq) \Rightarrow \frac{2T(X_1, \ldots, X_n) - n}{\sqrt{n}} \sim N(0,1)$$

4. El test se puede generaliza a tests sobre cuantiles de cualquier orden.

Ejemplo: Una empresa que tradicionalmente comenzaba su actividad diaria a las 9 h. ha cambiado su horario para abrir a las 8 h. y se pregunta si ello ha afectado significativamente al retraso de sus empleados. Es aceptable pensar que la forma de la distribución de los retrasos no ha variado con el cambio de horario, pero se teme que se haya desplazado hacia la derecha, lo cual supondrá un incremento del tiempo perdido. Se sabe, además, que la mediana de los retrasos de los empleados era inicialmente de 5 minutos. Con el cambio de horario se selecciona a 12 empleados y se observa, en determinados días, los siguientes retrasos (en minutos):

2.5, 1.2, 7, 1.8, 8.3, 6.8, 5.2, 3.4, 4.7, 6.2, 9.1, 5.2

A partir de estos datos, contrastar la hipótesis de que la distribución de los retrasos no ha variado con el cambio de horario.

9.2.2. Test de los rangos signados de Wilcoxon

Este test sólo se puede aplicar en el caso en que se conoce que la distribución, además de continua, es simétrica. Wilcoxon propuso un test para contrastar $H_0 : M_X = m$, que además de tener en cuenta la diferencia, tiene en cuenta la magnitud de la misma, por lo que es un test muy potente. El problema de este test es que necesita que la distribución sea simétrica y que los datos sean exactos.

Sea $X_1, \ldots, X_n$ una m.a.s. de X, v.a. con distribución continua y simétrica (alrededor de la mediana) y sea $D_i = X_i - m$, $i = 1, \ldots, n$. Si algún D_i es 0, se elimina ese dato y se reajusta el número de datos n.

El método propuesto por Wilcoxon consiste en ordenar de forma creciente los valores absolutos de estas diferencias $(|D_i|)$ y anotar el rango o lugar que ocupan $(r(|D_i|))$, de ahí

el nombre del test. Si hubiera empates, es decir si hubiera datos repetidos, se le asigna a cada uno el promedio de los rangos.

Basándose en esta idea, se define el estadístico del test de Wilcoxon como:

$$T^+(X_1,\ldots,X_n) = \text{suma de los rangos de los } D_i \text{ positivos} = \sum_{i=1}^{n} r(|D_i|)I\{D_i > 0\}$$

La distribución de $T^+(X_1,\ldots,X_n)$ bajo H_0 es simétrica en torno a la media $\dfrac{n(n+1)}{4}$ y viene dada por:

$$P[T^+(X_1,\ldots,X_n) = t] = P\left[T^+(X_1,\ldots,X_n) = \frac{n(n+1)}{2} - t\right]$$

Si $n \leq 15$, la distribución bajo H_0 está tabulada para ambas colas. Si $n > 15$, se puede aproximar asintóticamente la distribución de $T^+(X_1,\ldots,X_n)$, bajo H_0, por

$$T^+(X_1,\ldots,X_n) \approx \mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right).$$

Las regiones críticas del test según el contraste planteado son:

- Si $H_1 : M_X > m$ es cierta, cabe esperar que haya más diferencias mayores que 0 que menores que 0 y que la magnitud de las que son mayores que 0 sea mayor que la magnitud de las que son menores que 0. Por tanto se rechaza H_0 si $T^+(X_1,\ldots,X_n)$ es grande: Para un nivel de significación α se rechaza H_0 si

  $$T^+_{exp}(= T^+(x_1,\ldots,x_n)) \geq k$$

  con $P[T^+(X_1,\ldots,X_n) \geq k] \leq \alpha$.

  En general se calcula el $p - valor = P_{H_0}[T^+(X_1,\ldots,X_n) \geq T^+_{exp}]$.

- Si $H_1 : M_X < m$ es cierta, cabe esperar lo contrario que antes. Por tanto se rechaza H_0 si $T^+(X_1,\ldots,X_n)$ es pequeño: Para un nivel de significación α se rechaza H_0 si

  $$T^+_{exp} \leq k$$

  con $P[T^+(X_1,\ldots,X_n) \leq k] \leq \alpha$.

  En general se calcula el $p - valor = P_{H_0}[T^+(X_1,\ldots,X_n) \leq T^+_{exp}]$.

- Si $H_1 : M_X \neq m$ es cierta, se rechaza H_0 para valores pequeños o grandes de $T^+(X_1,\ldots,X_n)$: Para un nivel de significación α se rechaza H_0 si

  $$T^+_{exp} \leq k \text{ o } T^+_{exp} \geq \frac{n(n+1)}{2} - k$$

con $P[T^+(X_1,\ldots,X_n) \leq k] \leq \alpha/2$.

En general se calcula el $p-valor$:

$$p-valor = \begin{cases} 2P_{H_0}[T^+(X_1,\ldots,X_n) \geq T^+_{exp}] & \text{si } T^+_{exp} \geq \dfrac{n(n+1)}{4} \\ 2P_{H_0}[T^+(X_1,\ldots,X_n) \leq T^+_{exp}] & \text{si } T^+_{exp} \leq \dfrac{n(n+1)}{4} \end{cases}$$

Ejemplo: A partir de los datos del ejemplo anterior, y suponiendo que la distribución de los retrasos es simétrica, contrastar la hipótesis de que ésta no varía con el cambio de horario.

Nota: Los tests de localización se usan también para contrastar la hipótesis de homogeneidad de las distribuciones correspondientes a dos muestras apareadas o relacionadas cuando se tiene constancia de que las distribuciones tienen la misma forma funcional pero una está desplazada respecto de la otra. En dicho caso se toma la variable diferencia y $H_0 : M_{X-Y} = 0$. Para aplicar cada test deberá comprobarse primeramente que se está bajo las condiciones necesarias.

9.3. Problema de independencia: test χ^2

El problema de independencia relativo a dos muestras trata, como su nombre indica, de ver si dos variables, referidas a una misma población son independientes o no. Sean X e Y dos características poblacionales distintas. Se va a contrastar:

$$\begin{cases} H_0 : X \text{ e } Y \text{ son independientes} \\ H_1 : X \text{ e } Y \text{ no son independientes} \end{cases}$$

Para resolver este contraste se va a utilizar el test χ^2 de independencia.

Test χ^2 de independencia

Sean X e Y dos variables cualitativas, teniendo X las categorías $A_1,\ldots,A_m$, e Y las $B_1,\ldots,B_k$. Para resolver el problema de contraste H_0 : X e Y son independientes, se toma una m.a.s. de individuos y se clasifican los individuos según las categorías de X e Y. Sea N_{ij} el número de individuos de la muestra que presentan las categorías A_i y B_j, $\forall i = 1,\ldots,m$, $\forall j = 1,\ldots,k$.

Con dichos datos muestrales se construye la tabla de contingencia muestral (reparto de la muestra por categorías bidimensionales, totales, marginales y total global).

$X\backslash Y$	B_1	B_2	$\dots$	B_k	Totales
A_1	N_{11}	N_{12}	$\dots$	N_{1k}	$N_{1\cdot}$
A_2	N_{21}	N_{22}	$\dots$	N_{2k}	$N_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
A_m	N_{m1}	N_{m2}	$\dots$	N_{mk}	$N_{m\cdot}$
Totales	$N_{\cdot 1}$	$N_{\cdot 2}$	$\dots$	$N_{\cdot k}$	n

donde $N_{i\cdot} = \sum_{j=1}^{k} N_{ij}$ y $N_{\cdot j} = \sum_{i=1}^{m} N_{ij}$.

Sean $P_{ij} = P[X \in A_i, Y \in B_j]$, $P_{i\cdot} = P[X \in A_i]$ y $P_{\cdot j} = P[Y \in B_j]$, $\forall i = 1, \dots, m$, $\forall j = 1, \dots, k$. El estadístico del contraste es:

$$\sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

siendo O_{ij} la frecuencia observada y E_{ij} la frecuencia esperada.

Una forma de plantear la hipótesis nula, basada en las frecuencias, es $H_0 : P_{ij} = P_{i\cdot}P_{\cdot j}$, ya que la independencia de dos variables se caracteriza porque su frecuencia conjunta sea igual al producto de sus frecuencias marginales. Por lo tanto, bajo H_0 la frecuencia esperada $E_{ij} = nP_{ij}$ quedaría $E_{ij} = nP_{i\cdot}P_{\cdot j}$.

Por otro lado, ya que las frecuencias teóricas P_{ij}, $P_{i\cdot}$ y $P_{\cdot j}$, en general, no son conocidas, se trabajará con sus estimaciones máximo verosímiles: $\widehat{P}_{ij} = N_{ij}/n$, $\widehat{P}_{i\cdot} = N_{i\cdot}/n$ y $\widehat{P}_{\cdot j} = N_{\cdot j}/n$. Finalmente, la frecuencia observada $O_{ij} = N_{ij}$. Teniendo en cuenta todo esto se llega a que el estadístico del contraste es:

$$\chi^2(N_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(N_{ij} - \frac{N_{i\cdot}N_{\cdot j}}{n})^2}{\frac{N_{i\cdot}N_{\cdot j}}{n}}$$

que bajo H_0 tiene la distribución asintótica $\chi^2_{(m-1)(k-1)}$.

Al nivel de significación α se rechaza H_0 si

$$\chi^2_{exp} \geq \chi^2_{(m-1)(k-1);\alpha},$$

siendo $\chi^2_{exp} = \chi^2(x_1, \dots, x_n)$ y $P_{H_0}[\chi^2(N_{ij}) \geq \chi^2_{(m-1)(k-1);\alpha}] = \alpha$.

El p-valor es $P_{H_0}[\chi^2(N_{ij}) \geq \chi^2_{exp}]$.

Notas

- Los requisitos mínimos para poder usar la distribución asintótica indicada son:
  - Las frecuencias esperadas deben ser mayores o iguales que 2.

$$\frac{n_{i\cdot}n_{\cdot j}}{n} \geq 2$$

- Hay que asegurarse que al menos el 80 % de las frecuencias esperadas sea mayores o iguales a 5.

Si no es así, se debe aumentar el tamaño de la muestra.

- Para aplicar el test de χ^2 de independencia a dos v.a. cualesquiera, no cualitativas, se agrupan los valores de cada variable en un número finito de categorías que respeten las condiciones necesarias para poder aplicar el test.

Ejemplo: Para estudiar si el grupo sanguíneo de los individuos tiene relación con la predisposición a la diabetes, se han seleccionado al azar 400 sujetos a los que se ha determinado el grupo sanguíneo y el nivel de glucosa en sangre en idénticas condiciones experimentales. Clasificando la segunda medida en tres niveles, los resultados han sido:

Grupo\ Nivel	Bajo	Medio	Alto
O	137	86	35
A	42	23	11
B	19	17	7
AB	14	7	2

Contrastar, al nivel de significación 0.05, si ambas variables son independientes.

9.4. Problema de homogeneidad: test χ^2

El problema de homogeneidad consiste en estudiar si una serie de poblaciones se comportan de la misma forma frente a una determinada característica. Para ello se toman m.a.s. de cada población, se mide la característica de interés en ellas y se trata de contrastar si todas las muestras proceden de variables con la misma distribución teórica. Se va a contrastar:

$$\begin{cases} H_0 : F_1 = \ldots = F_m \\ H_1 : \text{Alguna distribución es distinta} \end{cases}$$

Para resolver dicho contraste se va a utilizar el test χ^2 de homogeneidad.

Test χ^2 de homogeneidad

Este test se debe aplicar, en un principio, a características de tipo cualitativo. Se van a suponer m poblaciones, m muestras aleatorias simples, de tamaños $n_1, \ldots, n_m$ y que, en todos los casos, las variables pueden tomar valores en k categorías $A_1, \ldots, A_k$. Sean N_{ij} el número de observaciones de la muestra $i-$ésima que presenta la modalidad A_j, $i = 1, \ldots, m$, $j = 1, \ldots, k$, $N_{\cdot j} = \sum_{i=1}^{m} N_{ij}$, $n_i = \sum_{j=1}^{k} N_{ij}$ y $n = \sum_{i=1}^{m} n_i$.

Con dichos datos muestrales se construye la tabla de contingencia muestral

Muestras\Categorías	A_1	A_2	$\dots$	A_k	
1	N_{11}	N_{12}	$\dots$	N_{1k}	n_1
2	N_{21}	N_{22}	$\dots$	N_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
m	N_{m1}	N_{m2}	$\dots$	N_{mk}	n_m
	$N_{\cdot 1}$	$N_{\cdot 2}$	$\dots$	$N_{\cdot k}$	n

Si se denota por P_{ij} a la probabilidad de que un individuo de la muestra $i-$ésima presente la modalidad A_j, la hipótesis nula del contraste se puede escribir como: $H_0 : P_{1j} = P_{2j} = \ldots = P_{mj} (= P_{\cdot j})$, $j = 1, \ldots, k$. Desde el punto de vista paramétrico, este contraste se puede ver como contrastar la igualdad de los parámetros p de m multinomiales de dimensión $k-1$, que se puede resolver mediante el test de la razón de verosimilitud. Como se está estudiando el caso no paramétrico, se va a resolver el contraste con el test χ^2, pero ambos test son asintóticamente equivalentes.

Al igual que antes, el estadístico del contraste es:

$$\sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

siendo O_{ij} la frecuencia observada y E_{ij} la frecuencia esperada.

En este caso, bajo H_0 la frecuencia esperada es $E_{ij} = n_i P_{\cdot j}$, pero como no son conocidas, se trabajará con sus estimaciones máximo verosímiles: $\widehat{P}_{\cdot j} = N_{\cdot j}/n$. Finalmente, al igual que antes, la frecuencia observada es $O_{ij} = N_{ij}$, con lo que el estadístico del contraste queda:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(N_{ij} - \frac{n_i N_{\cdot j}}{n})^2}{\frac{n_i N_{\cdot j}}{n}}$$

que, bajo H_0 tiene la distribución asintótica $\chi^2_{(m-1)(k-1)}$.

Al nivel de significación α se rechaza H_0 si

$$\chi^2_{exp} \geq \chi^2_{(m-1)(k-1);\alpha},$$

siendo $\chi^2_{exp} = \chi^2(x_1, \ldots, x_n)$ y $P_{H_0}[\chi^2(N_{ij}) \geq \chi^2_{(m-1)(k-1);\alpha}] = \alpha$.
El p-valor es $P_{H_0}[\chi^2(N_{ij}) \geq \chi^2_{exp}]$.

Notas:

- Los requisitos mínimos para poder usar la distribución asintótica indicada son:
  - Los tamaños muestrales en cada población deben ser como mínimo de 20.

$$n_i \geq 20, \ \forall i = 1, \ldots, m.$$

- Las frecuencias esperadas deben ser mayores o iguales que 2.

$$\frac{n_i n_{\cdot j}}{n} \geq 2$$

- Hay que asegurarse que no más del 20 % de las frecuencias esperadas sea menores a 5.

Si no es así, se debe aumentar el tamaño de las muestras.

- A pesar de que existe una gran analogía con el test de independencia, el problema que resuelve este otro test es totalmente distinto.
- Para aplicar el test χ^2 de homogeneidad a m variables aleatorias cualesquiera, es decir, si se tienen $X_1, \ldots, X_m$ variables aleatorias cualesquiera y se desea contrastar $H_0 : F_1 = \ldots = F_m$, se particiona el rango de valores comunes a todas las variables en k subconjuntos o modalidades (A_j) de probabilidad no nula bajo todas las distribuciones. Se considera P_{ij} =probabilidad de que la variable $X_i \in A_j$, $\forall i = 1, \ldots, m$, $\forall j = 1, \ldots, k$ y se toma una m.a.s. de cada variable. Como las muestras son independientes se puede aplicar el test χ^2 de homogeneidad a ellas para resolver el contraste planteado.
- Para variables no cualitativas hay tests mucho mejores, ya que el test χ^2 no utiliza los datos, sino la pertenencia a algunos intervalos.

Ejemplo: Contrastar, a partir de los resultados de la siguiente tabla, si los distintos grupos sanguíneos se presentan con la misma frecuencia en tres grupos étnicos diferentes:

Raza\ Grupo	O	A	B	AB
1	32	11	7	2
2	47	13	17	9
3	23	7	9	6