Contraste de hipótesis simples

El objetivo es determinar el test más potente para contrastar dos hipótesis simples al nivel de significación α.

Teorema (Lema de Neyman-Pearson para contrastes determinísticos)**.** Sea X una variable aleatoria con distribución F_θ y supongamos que el espacio paramétrico se reduce a dos posibles valores $\Theta = \{\theta_0, \theta_1\}$. Sean $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$ dos hipótesis simples y $\vec{X} = (X_1, \ldots, X_n)$ una muestra aleatoria simple de X con función de densidad o probabilidad $f_\theta(x_1, \ldots, x_n)$. Sea $C_0 \subset \mathcal{X}^n$ un subconjunto medible del espacio muestral que verifica

$$\{\vec{x} \mid f_{\theta_1}(\vec{x}) > k f_{\theta_0}(\vec{x})\} \subset C_0 \subset \{\vec{x} \mid f_{\theta_1}(\vec{x}) \geq k f_{\theta_0}(\vec{x})\}$$

para algún $k > 0$ y $P_{\theta_0}(C_0) = \alpha$. Entonces C_0 es la región crítica asociada a un test de significación α y máxima potencia entre los tests[1] de nivel α.

Demostración. La demostración se realizará para el caso continuo, aunque en el caso discreto se haría de forma análoga.

La región crítica C_0 está construida de manera que, si la muestra $\vec{x}$ verifica $f_{\theta_1}(\vec{x}) > k f_{\theta_0}(\vec{x})$, entonces $\vec{x} \in C_0$ y se rechaza H_0. Por el contrario si $\vec{x}$ verifica $f_{\theta_1}(\vec{x}) < k f_{\theta_0}(\vec{x})$ entonces $\vec{x} \notin C_0$ y no se rechaza H_0. En el caso $f_{\theta_1}(\vec{x}) = k f_{\theta_0}$ entonces $\vec{x}$ puede estar o no en C_0 y no sabemos si se rechaza o no H_0.

Sea C otra región crítica con $P_{\theta_0}(C) \leq \alpha$. Comprobaremos que $P_{\theta_1}(C_0) \geq P_{\theta_1}(C)$.

Partiendo cada una de las dos regiones críticas en dos conjuntos disjuntos, lo común y el resto, y aplicando propiedades de la probabilidad, se tiene:

$$\left.\begin{array}{r} C_0 = (C_0 \cap C) \cup (C_0 \cap C^c) \\ C = (C_0 \cap C) \cup (C_0^c \cap C) \end{array}\right\} P_{\theta_1}(C_0) - P_{\theta_1}(C) = P_{\theta_1}(C_0 \cap C^c) - P_{\theta_1}(C_0^c \cap C)$$

Como $C_0 \cap C^c \subset C_0$

$$P_{\theta_1}(C_0 \cap C^c) = \int_{C_0 \cap C^c} f_{\theta_1}(\vec{x}) d\vec{x} \geq \int_{C_0 \cap C^c} k f_{\theta_0}(\vec{x}) d\vec{x} = k P_{\theta_0}(C_0 \cap C^c)$$

Como $C_0^c \cap C \subset C_0^c$

$$P_{\theta_1}(C_0^c \cap C) = \int_{C_0^c \cap C} f_{\theta_1}(\vec{x}) d\vec{x} \leq \int_{C_0^c \cap C} k f_{\theta_0}(\vec{x}) d\vec{x} = k P_{\theta_0}(C_0^c \cap C)$$

[1] Entre los determinísticos. Véase el teorema siguiente para el caso de contrastes aleatorizados.

luego

$$\begin{aligned}
P_{\theta_1}(C_0) - P_{\theta_1}(C) &= P_{\theta_1}(C_0 \cap C^c) - P_{\theta_1}(C_0^c \cap C) \\
&\geq k[P_{\theta_0}(C_0 \cap C^c) - P_{\theta_0}(C_0^c \cap C)] \\
&\geq k[P_{\theta_0}(C_0 \cap C^c) + P_{\theta_0}(C_0 \cap C) - P_{\theta_0}(C_0 \cap C) - P_{\theta_0}(C_0^c \cap C)] \\
&= k[P_{\theta_0}(C_0) - P_{\theta_0}(C)] \geq 0
\end{aligned}$$

ya que $k > 0$ y $P_{\theta_0}(C) \leq \alpha = P_{\theta_0}(C_0)$
En consecuencia, $P_{\theta_1}(C_0) \geq P_{\theta_1}(C)$.

Otra manera de verlo sería considerando que, si se define

$$\varphi = \varphi(\vec{x}) = \begin{cases} 1 & f_1 = f_{\theta_1}(\vec{x}) > k \cdot f_{\theta_0}(\vec{x}) = k \cdot f_0 \\ \varphi(\vec{x}) & f_1 = k \cdot f_0 \\ 0 & f_1 < k \cdot f_0 \end{cases}$$

y siendo ψ otro contraste a nivel de significación $\alpha = E_{\theta_0}(\varphi)$, es decir, $E_{\theta_0}(\psi) \leq \alpha$, entonces

$$\int (\varphi - \psi)\,(f_1 - k\,f_0) = \int_{X(\Omega)^n} [\varphi(\vec{x}) - \psi(\vec{x})]\,[f_1(\vec{x}) - k\,f_0(\vec{x})]\,d\vec{x} \geq 0$$

pues cuando $f_1 - k\,f_0 > 0$ entonces $\varphi = 1 \geq \psi \Longrightarrow \varphi - \psi \geq 0$ y cuando $f_1 - k\,f_0 < 0$ entonces $\varphi = 0 \Longrightarrow \varphi - \psi \leq 0$, luego

$$\begin{aligned}
0 \leq \int (\varphi - \psi)\,(f_1 - k\,f_0) &= \int \varphi\,f_1 - \int \psi\,f_1 - k\left(\int \varphi\,f_0 - \int \psi\,f_0\right) \\
&= E_{\theta_1}\varphi - E_{\theta_1}\psi - k\,(E_{\theta_0}\varphi - E_{\theta_0}\psi) \\
&= \text{potencia}_\varphi - \text{potencia}_\psi - \underbrace{k}_{>0}\underbrace{(\alpha - \text{tamaño}_\psi)}_{\geq 0} \\
&\leq \text{potencia}_\varphi - \text{potencia}_\psi \\
&\Rightarrow \text{potencia}_\varphi \geq \text{potencia}_\psi
\end{aligned}$$

$\square$

Una versión más general del resultado anterior utilizando contrastes aleatorizados es la siguiente:

Teorema (Lema de Neyman-Pearson para contrastes aleatorizados)**.** Sean $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$ dos hipótesis simples y $(X_1, \ldots, X_n)$ una muestra aleatoria simple. Entonces se verifica que:

- Cualquier test de la forma

$$\varphi(\vec{x}) = \begin{cases} 1 & \text{si} \quad f_{\theta_1}(\vec{x}) > k f_{\theta_0}(\vec{x}) \\ \gamma(\vec{x}) & \text{si} \quad f_{\theta_1}(\vec{x}) = k f_{\theta_0}(\vec{x}) \\ 0 & \text{si} \quad f_{\theta_1}(\vec{x}) < k f_{\theta_0}(\vec{x}) \end{cases} \tag{1}$$

  con $k > 0$ es el más potente entre los de su tamaño.

- Para cada $\alpha \in (0,1)$ existe un test con $\gamma(\vec{x}) = \gamma = \mathrm{c}^{\mathrm{te}}$ de la forma (1) con tamaño α.
- La función crítica de un test de máxima potencia al nivel α para contrastar H_0 frente a H_1 es de la forma descrita en (1) salvo a lo sumo en un conjunto del espacio muestral que tiene probabilidad cero bajo H_0 y H_1.

Demostración. El primer punto se sigue directamente de la expresión usada al final de la demostración anterior,

$$\int (\varphi - \psi)\,(f_1 - k\,f_0) \geq 0$$

Para el segundo punto, como

$$\alpha = E_{\theta_0}\varphi = \Pr[f_1 > k\,f_0] + \gamma\ \Pr[f_1 = k\,f_0]$$

basta tomar

$$\gamma = \frac{\alpha - \Pr[f_1 > k\,f_0]}{\Pr[f_1 = k\,f_0]}$$

Para el tercer punto, para un test ψ de máxima potencia a nivel de significación α,

$$\begin{aligned}
&\text{potencia}_\varphi = \text{potencia}_\psi \quad \bigwedge \quad \alpha = \text{tamaño}_\varphi \geq \text{tamaño}_\psi \\
&\models\!\!\!\Longrightarrow \quad \int \varphi\,f_1 = \int \psi\,f_1 \quad \bigwedge \quad \int \varphi\,f_0 \geq \int \psi\,f_0 \\
&\models\!\!\!\Longrightarrow \quad 0 \leq \int (\varphi - \psi)\,(f_1 - k\,f_0) \leq 0 \\
&\models\!\!\!\Longrightarrow \quad \int (\varphi - \psi)\,(f_1 - k\,f_0) = 0
\end{aligned}$$

y, como el integrando es no negativo,

$$(\varphi - \psi)\,(f_1 - k\,f_0) = 0 \ \overset{\Pr}{\forall}\ \vec{x}$$

luego, casi seguro,

$$\begin{aligned}
f_1 \neq k\,f_0 &\quad\Longrightarrow\quad \varphi = \psi \\
f_1 = k\,f_0 &\quad\Longrightarrow\quad \psi \text{ puede tomar valores arbitrarios}
\end{aligned}$$

luego ψ es de la forma (1). □

Ejemplo (caso gausiano)**.** Sea X una variable aleatoria con distribución $N(\mu, \sigma)$ con σ conocido, de la que se extrae una m.a.s. de tamaño n para contrastar las hipótesis

$$\begin{aligned}
H_0 &\equiv \mu = \mu_0 \\
H_1 &\equiv \mu = \mu_1
\end{aligned} \qquad \mu_1 < \mu_0$$

Aplica el criterio de Neyman-Pearson para hacer el contraste de hipótesis a nivel de significación α.

Resolución.

$$L(x_1,\ldots,x_n,\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2}\frac{\sum(x_i-\mu)^2}{\sigma^2}\right]$$

$$\begin{aligned}
\frac{L(x_1,\ldots,x_n,\mu_1,\sigma)}{L(x_1,\ldots,x_n,\mu_0,\sigma)} &= \frac{\exp\left[-\frac{1}{2}\frac{\sum(x_i-\mu_1)^2}{\sigma^2}\right]}{\exp\left[-\frac{1}{2}\frac{\sum(x_i-\mu_0)^2}{\sigma^2}\right]} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 + n\mu_1^2 - 2\mu_1\sum x_i - \sum x_i^2 - n\mu_0^2 + 2\mu_0\sum x_i\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu_1^2-\mu_0^2) + 2(\mu_0-\mu_1)\sum x_i\right]\right\}
\end{aligned}$$

$$\begin{aligned}
\text{R.C.} &= \left\{\exp\left(-\frac{1}{2\sigma^2}\left[n(\mu_1^2-\mu_0^2)+2(\mu_0-\mu_1)\sum x_i\right]\right) > k\right\} \\
&= \left\{-\frac{1}{2\sigma^2}\left[n(\mu_1^2-\mu_0^2)+2(\mu_0-\mu_1)\sum x_i\right] > \ln k\right\} \\
&= \left\{\frac{1}{2\sigma^2}\left[n(-\mu_1^2+\mu_0^2)+2(-\mu_0+\mu_1)\sum x_i\right] > \ln k\right\} \\
&= \left\{n(-\mu_1^2+\mu_0^2)+2(-\mu_0+\mu_1)\sum x_i > 2\sigma^2\ln k\right\} \\
&= \left\{2(\mu_1-\mu_0)\sum x_i > 2\sigma^2\ln k + n(\mu_1^2-\mu_0^2)\right\} \\
&= \left\{\sum x_i < k^*\right\}
\end{aligned}$$

pues $\mu_1-\mu_0<0$ y con $k^* = \frac{2\sigma^2\ln k + n(\mu_1^2-\mu_0^2)}{2(\mu_1-\mu_0)}$. Entonces

$$P[\text{R.C.} \mid H_0] = \alpha \iff P\left[\sum X_i < k^* \mid H_0\right] = \alpha \iff P[N(n\mu_0,\sqrt{n}\sigma) < k^*] = \alpha$$

$\square$

Veamos un ejemplo con una muestra pequeña:

```
> alfa <- 0.05
> n <- 9
> mu0 <- 6
> mu1 <- 5
> sigma <- 1
> (k. <- qnorm (alfa, n*mu0, sqrt(n)*sigma)) # k*
```

```
[1] 49.06544
```

$$\text{R.C} = \left\{\sum X_i < 49\text{'}06544\right\}$$

```
> (X <- rnorm (n, mu1, sigma)) # bajo H1
```

```
[1] 5.114042 4.928529 5.283143 5.169557 6.520205 3.896682 5.716259 5.168419
[9] 4.743231

> sum(X)

[1] 46.54007

> pnorm (sum(X), n*mu0, sqrt(n)*sigma) # P-valor

[1] 0.006447709
```

Nótese que esta región crítica vale para cualquier μ_1 que sea menor que μ_0.

Ejemplo (caso Poisson)**.** Sea $X \equiv P(\lambda)$ una variable aleatoria de la que se extrare una m.a.s. de tamaño n, $(x_1, \ldots, x_n)$, para contrastar la hipótesis $H_0 : \lambda = \lambda_0$ frente a $H_0 : \lambda = \lambda_1$ con $\lambda_0 < \lambda_1$.

Utiliza el lema de Neyman-Pearson para contrastar esas hipótesis.

Resolución.

$$L(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^{n} p(x_i, \lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod (x_i!)}$$

La región crítica para el contraste se elige de la siguiente forma:

$$\text{R.C.} = \left\{ \frac{L(x_1, \ldots, x_n, \lambda_1)}{L(x_1, \ldots, x_n, \lambda_0)} > k \right\} = \left\{ \frac{e^{-n\lambda_1} \lambda_1^{\sum x_i}}{e^{-n\lambda_0} \lambda_0^{\sum x_i}} > k \right\}$$

$$= \left\{ e^{-n(\lambda_1 - \lambda_0)} \left(\frac{\lambda_1}{\lambda_0} \right)^{\sum x_i} > k \right\} = \left\{ \left(\frac{\lambda_1}{\lambda_0} \right)^{\sum x_i} > k\, e^{n(\lambda_1 - \lambda_0)} \right\}$$

Aplicando logaritmos neperianos se obtiene:

$$\text{R.C.} = \left\{ \ln \left(\frac{\lambda_1}{\lambda_0} \right) \sum x_i > \ln \left(k\, e^{n(\lambda_1 - \lambda_0)} \right) \right\} =$$

$$= \left\{ \sum x_i > k^* \right\}$$

teniendo en cuenta que $\frac{\lambda_1}{\lambda_0} > 1$ y tomando $k^* = \frac{\ln(k) + \ln(n(\lambda_1 - \lambda_0))}{\ln(\lambda_1) - \ln(\lambda_0)}$.

Para un valor de k^* y nivel de significacion $\alpha = P(\sum x_i > k^* \mid \lambda_0)$ el test con región crítica $\{\sum x_i > k^*\}$ es el más potente para cualquier valor de λ_1 mayor que λ_0. □

Por ejemplo, dada la muestra aleatoria simple $\vec{x} = (7, 4, 4, 6, 4, 5, 8, 9, 4, 7)$ procedente de una distribución de Poisson, contrastemos las hipótesis $H_0 : \lambda = 5$ y $H_1 : \lambda = 6$ al nivel de significación $\alpha = 0\text{'}04239093$.

La región crítica óptima es de la forma:

$$\text{R.C.} = \left(\sum_{i=1}^{10} x_i > k^* \right) \qquad \text{con } P \left(\sum_{i=1}^{10} x_i > k^* \,\middle|\, \lambda = \lambda_0 \right) = \alpha$$

y sólo es necesario buscar el valor de k^*.

$$P\left(\sum_{i=1}^{10} x_i > k^* \,\middle|\, \lambda_0\right) = P\left(\text{Poisson}(10\lambda_0) > k^*\right) = 0\text{'}04239093$$

que se obtiene con $k^* = 62$:

```
> x <- c(7, 4, 4, 6, 4, 5, 8, 9, 4, 7)
> 1 - ppois(62, 50)

[1] 0.04239093

> sum(x)

[1] 58

> mean(x)

[1] 5.8
```

El p-valor del contraste es:

```
> 1 - ppois(sum(x)-1, 50)

[1] 0.1448594
```

Si se desea realizar el contraste anterior con un tamaño del test igual a 0'05, el problema está en que la distribución de Poisson es discreta y no toma todos los valores del intervalo $(0, 1)$:

$$P\left[\sum_{i=1}^{10} x_i > 62 \,\middle|\, \lambda_0\right] + P\left[\sum_{i=1}^{10} x_i = 62 \,\middle|\, \lambda_0\right] = 0\text{'}04239093 + 0\text{'}01328984 = 0\text{'}05568078$$

es decir, con la R.C. $= \left\{\sum_{i=1}^{10} x_i > 62\right\}$ no se alcanza el valor de α y con R.C. $= \left\{\sum_{i=1}^{10} x_i \geq 62\right\}$ se sobrepasba.

Para resolver esta situación se puede recurrir a un contraste aleatorizado, que se define de la siguiente forma:

$$\varphi(\vec{x}) = \begin{cases} 1 & \text{si} \quad f_{\theta_1}(\vec{x}) > k f_{\theta_0}(\vec{x}) \\ \gamma & \text{si} \quad f_{\theta_1}(\vec{x}) = k f_{\theta_0}(\vec{x}) \\ 0 & \text{si} \quad f_{\theta_1}(\vec{x}) < k f_{\theta_0}(\vec{x}) \end{cases}$$

de forma que

$$E[\varphi(\vec{x})] = P\left[\sum_{i=1}^{10} x_i > k^* \,\middle|\, \lambda_0\right] + \gamma\, P\left[\sum_{i=1}^{10} x_i = k^* \,\middle|\, \lambda_0\right] = \alpha$$

Para alcanzar el tamaño α, basta con que γ verifique:

$$\gamma = \frac{\alpha - P\left[\sum_{i=1}^{10} x_i > 62 \,\middle|\, \lambda_0\right]}{P\left[\sum_{i=1}^{10} x_i = 62 \,\middle|\, \lambda_0\right]} = 0\text{'}5725475$$

Por lo tanto el test aleatorizado con tamaño 0'05 sería

$$\varphi(\vec{x}) = \begin{cases} 1 & \text{si } \sum_{i=1}^n x_i > 62 \\ 0\text{'}5725 & \text{si } \sum_{i=1}^n x_i = 62) \\ 0 & \text{si } \sum_{i=1}^n x_i < 62) \end{cases}$$

El test funciona de la siguiente manera:

- Cuando la suma de los valores de la muestra es mayor que 62 se rechaza H_0.
- En el caso de que la suma es menor que 62, no se rechaza H_0.
- Si la suma es 62, se realiza un sorteo con probabilidad 0'5725 de rechazo y 0'4275 de no rechazo.

Ejemplo. Sea $X \equiv B(p)$ de la que se extrae una muestra $(X_1, \ldots, X_n)$ para resolver el contraste $H_0 : p = p_0$ frente a $H_1 : p = p_1$ con $p_0 < p_1$.

¿Cómo se resolvería usando el lema?

Resolución. La verosimilitud sería

$$L(\vec{x}, p) = \prod_{i=1}^n P[X_i = x_i] = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

luego la región crítica tendría la forma:

$$
\begin{array}{rcl}
\text{R.C.} & = & \left\{ \dfrac{L(\vec{x}, p_1)}{L(\vec{x}, p_0)} > k \right\} \\
& = & \left\{ \dfrac{p_1^{\sum x_i}(1-p_1)^{n-\sum x_i}}{p_0^{\sum x_i}(1-p_0)^{n-\sum x_i}} > k \right\} \\
& = & \left\{ \dfrac{p_1^{\sum x_i}(1-p_1)^{n}(1-p_1)^{-\sum x_i}}{p_0^{\sum x_i}(1-p_0)^{n}(1-p_0)^{-\sum x_i}} > k \right\} \\
\left\langle k^* = k\dfrac{(1-p_0)^n}{(1-p_1)^n} \right\rangle & = & \left\{ \dfrac{p_1^{\sum x_i}(1-p_1)^{-\sum x_i}}{p_0^{\sum x_i}(1-p_0)^{-\sum x_i}} > k^* \right\} \\
& = & \left\{ \dfrac{p_1^{\sum x_i}(1-p_0)^{\sum x_i}}{p_0^{\sum x_i}(1-p_1)^{\sum x_i}} > k^* \right\} \\
& = & \left\{ \left(\dfrac{p_1(1-p_0)}{p_0(1-p_1)} \right)^{\sum x_i} > k^* \right\} \\
\langle \ln \text{ es una función creciente} \rangle & = & \left\{ \sum x_i \ln \left(\dfrac{p_1(1-p_0)}{p_0(1-p_1)} \right) > \ln k^* \right\} \\
& = & \left\{ \sum x_i \left(\ln \dfrac{p_1}{p_0} + \ln \dfrac{1-p_0}{1-p_1} \right) > \ln k^* \right\} \\
\left\langle p_0 < p_1 \implies \left\{ \begin{array}{l} \ln \frac{p_1}{p_0} > 0 \\ \ln \frac{1-p_0}{1-p_1} > 0 \end{array} \right. \right\rangle & = & \left\{ \sum x_i > \dfrac{\ln k^*}{\ln \frac{p_1}{p_0} + \ln \frac{1-p_0}{1-p_1}} \right\} \\
\left\langle k^{**} = \dfrac{\ln k^*}{\ln \frac{p_1}{p_0} + \ln \frac{1-p_0}{1-p_1}} \right\rangle & = & \left\{ \sum x_i > k^{**} \right\}
\end{array}
$$

$\square$

Por ejemplo, considérese $p_0 = 0\text{'}1$, $p_1 = 0\text{'}2$ y $n = 160$. En este caso,

$$\sum_{i=1}^{n} X_i \overset{H_0}{\equiv} B(160, 0\text{'}1)$$

Si se pretende establecer un nivel de significación cercano a 0'05, puede considerarse

$$\alpha = 0\text{'}0487185096736285 = P\left[\sum X_i > 22\right]$$

y la región crítica sería

$$\text{R.C.} = \left\{\sum X_i > 22\right\} = \left\{\hat{p} = \bar{X} > \frac{22}{160} = 0\text{'}1375\right\}$$

Como en los casos anteriores, téngase en cuenta que esta región crítica vale para cualquier valor p_1 mayor que p_0, independientemente de su valor concreto.