# Pruebas de Hipótesis Multiples

Cuando queremos hacer comparaciones de mas de dos poblaciones, una alternativa es comparar todos los grupos a la vez con el método de Análisis de Varianza (**ANOVA**)

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_k$$

En un ANOVA siempre se van a considerar varias fuentes de variación y si se rechaza la hipótesis nula, la pregunta a plantearse es “en qué grupo se dió la diferencia”

rohen

Si lo que se va a probar es el efecto de solo un factor sobre alguna característica de interés, el análisis de varianza apropiado a usar es el ANOVA en una dirección o de un solo factor.

Si las unidades experimentales se han elegido aleatoriamente, tenemos lo que se llama un experimento diseñado como *completamente aleatorizado*.

En un Análisis en el que se quiere suponer que la media de los diferentes grupos bajo tratamiento es la misma, $\mu_1 = \mu_2 = \cdots = \mu_k$ suponemos que las ***varianzas en cada grupo son iguales*** $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, y que las ***muestras en cada grupo vienen de una distribución normal***.

rohen

## *Fuentes de Variación.*

Variación Total (Total SS)

Dentro de los Grupos (Within groups SS)

Entre Grupos (Between groups SS)

$$TSS = BSS + WSS$$

**Variación Total (Total SS)**

Es la variación presente en todos los *n* datos y tiene asociada *n-1* grados de libertad

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left( X_{ij} - \overline{X} \right)^2$$

**Dentro de los Grupos**

El estimador de la variación total será la variación ponderada de todos los grupos, con *n-k* grados de libertad asociados a este estimador.

$$WSS = \sum_{i=1}^{k} \left[ \sum_{j=1}^{n_i} \left( X_{ij} - \overline{X}_i \right)^2 \right]$$

*WSS*/(*n-k*) se conoce como error cuadrado medio (*EMS*).

**Entre Grupos (Between groups SS):**

La cantidad de variabilidad entre los grupos es la cantidad más importante en la prueba estadística. Nos indica si realmente existe una diferencia entre los grupos

$$BSS = n_i \sum_{i=1}^{k} \left( \overline{X}_i - \overline{X} \right)^2$$

Tiene asociados *k-1* grados de libertad. (*k es el número de grupos a estudiar)*

*BSS*/(*k-1*) = *BMS* es la cantidad a comparar con *WSS*/(*n-k*)

rohen

## Probando la Hipótesis Nula

Si la hipótesis nula es verdadera, entonces tanto *BMS* como *EMS* estiman cada una a $\sigma^2$, la varianza común de los *k* grupos. Pero si las *k* medias de los grupos no son iguales, entonces *BMS* será más grande que *EMS*.

Así, la prueba para la igualdad de medias es una prueba de cociente de varianzas de una cola

$$F = \frac{BSS/(k-1)}{ESS/(n-k)} \sim F_{k-1,n-k}$$

En caso de rechazarse la hipótesis nula, para encontrar cuál es el grupo que marca la diferencia podemos usar pruebas *t* por pares. **cuidado!**

rohen

ANOVA es robusto bajo considerable heterogeneidad de la varianza, siempre y cuando los tamaños de muestra dentro de los grupos se mantengan parecidos.

Si hay duda de la validez de la hipótesis de normalidad del ANOVA o tenemos variables cualitativas, la prueba no paramétrica de Kruskal-Wallis es la prueba análoga a la ANOVA de un solo factor, usando rangos en vez de medias.

La prueba *Kruskal Wallis*

Cuando no se cumplen los supuestos del análisis de varianza múltiple, K-W puede alcanzar una potencia hasta del 95%. Se basa en el rango de las *n* observaciones que inicialmente se obtienen sin distinguir grupos.

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

*$n_i$ es el número de observaciones en el grupo i.* $n = \sum_{i=1}^{k} n_i$
*$R_i$ es la suma de los $n_i$ rangos del gupo i*

De rechazarse la hipótesis nula, se procede a encontrar el grupo o los grupos que son diferentes.

rohen

Se rechaza $H_o : \mu_1 = \mu_2 = \cdots = \mu_k$. Es muy tentador realizar pruebas por pares: $H_o : \mu_i = \mu_j, i \neq j$.

Problema: aumenta error tipo I, disminuye la potencia.

Cada prueba *t* para cada pareja de grupos se realiza a un nivel de significancia $\alpha$ y hay una probabilidad de $100(1-\alpha)\%$ de correctamente no rechazar *Ho* cuando las dos medias poblacionales son iguales.

Para un conjunto de *K* hipótesis, la probabilidad de correctamente no rechazar todas ellas es de $(1-\alpha)^K$, por lo que la de rechazar incorrectamente a todas es de $1-(1-\alpha)^K$. Mientras mayor sea el número de muestras por pares a realizar, mayor será la probabilidad de error tipo I, lo cual hace más fácil rechazar la hipótesis nula, aún cuando sea verdadera.

rohen

Probabilidad de cometer al menos un error tipo I usando pruebas $t$ para todas las $K$ posibles parejas de los $k$ grupos

| | | Nivel de significancia a usado en las pruebas $t$ | | | | |
|---|---|---|---|---|---|---|
| $k$ | $K$ | 0.1 | 0.05 | 0.01 | 0.005 | 0.0001 |
| 2 | 1 | 0.10 | 0.05 | 0.01 | 0.005 | 0.001 |
| 3 | 3 | 0.27 | 0.14 | 0.03 | 0.015 | 0.003 |
| 4 | 6 | 0.47 | 0.26 | 0.06 | 0.030 | 0.006 |
| 5 | 10 | 0.65 | 0.40 | 0.10 | 0.049 | 0.010 |
| 6 | 15 | 0.79 | 0.54 | 0.14 | 0.072 | 0.015 |
| 10 | 45 | 0.99 | 0.90 | 0.36 | 0.202 | 0.044 |
| | $\infty$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Hay $K = k\ (k-1)/2$ comparaciones por pares de $k$ medias.

Posibles ajustes más adelante …

rohen

Prueba no paramétrica para comparaciones múltiples.
**Prueba de Tukey**.

Comparaciones múltiples se pueden efectuar, usando sumas de rangos en vez de medias.

Las sumas de rangos $R_i$ se ordenan por magnitud y se prueban de las sumas más distantes a las más cercanas.

El error estándar se calcula como $SE = \sqrt{\left(\frac{n(n+1)}{12} - \frac{\sum t}{12(n-1)}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$ en caso de haber empates.

La estadística de prueba es $Q = \left(\overline{R}_j - \overline{R}_i\right)/SE$ que se contrasta con valores críticos $q$.

*(Cf. J.H.Zar.- Biostatistical Analysis)*

rohen

# Ejemplo

| Muestras Ordenadas de acuerdo a su rango medio ( $i$ ) | 1 | 2 | 4 | 3 |
|---|---|---|---|---|
| Suma de Rango ( $R\,i$ ) | 55.0 | 132.5 | 163.5 | 145.0 |
| Tamaño de Muestra | 8 | 8 | 8 | 7 |
| Rango medio ( $R\,i$ ) | 6.88 | 16.56 | 20.44 | 20.71 |

Muestra 4 y 3 son diferentes de la 1, pero en la muestra 2 hay ambigüedad

Cuando se prueba una sola hipótesis, generalmente nos interesa controlar la tasa de falsos positivos y maximizar al mismo tiempo la probabilidad de detectar un efecto positivo, cuando realmente existe, es decir, maximizar la potencia de la prueba.

Al analizar expresión de genes, por ejemplo, se contrastan varios grupos y generalmente por pares. Cuando se aplican pruebas repetidas, el *valor-p* es conceptualmente aplicado a cada prueba (controla la tasa de falsos positivos en una prueba)

rohen

La cantidad más comúnmente controlada es la tasa de error de la familia de pruebas (*FWER* - *familywise error rate*).

El ajuste que hace Bonferroni es dividir el nivel de significancia $\alpha$ preestablecido por $K$ el número de pares de pruebas a realizar.

Otros métodos están basados en la tasa de falsos descubrimientos (*FDR*) (rechazar *Ho* cuando es verdadera)

Posibles resultados de $K$ pruebas estadísticas

| | Aceptar Hipotesis Nula | Rechazar Hipótesis Nula | Total |
|---|---|---|---|
| Hipótesis Nula Cierta | $U$ | $V$ | $k_o$ |
| Hipótesis Alternativa Cierta | $T$ | $S$ | $k_1$ |
| | $W$ | $R$ | $K$ |

rohen

La definición formal de *FWER* es $P(V \geq 1)$

Benjamini y Hochberg definen la tasa de falsos descubrimientos como

$$FDR = E\left[\frac{V}{R} \middle| R > 0\right] \cdot P(R > 0)$$

Donde $V$ es el número de hipótesis rechazadas incorrectamente y $R$ el número de hipótesis rechazadas (correcta o incorrectamente). El cociente $V/R$ es considerado cero cuando $R=0$. Sin embargo no perdamos de vista que $P(R = 0)$ puede ser grande.

El proceso de Benjamini y Hochberg propone ordenar los *valores-p* resultantes de las $K$ pruebas, de modo que $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$. Si calculamos

$$\hat{k} = \arg\max_{1 \leq k \leq K} \left\{ k : p_{(K)} \leq \alpha \cdot k / K \right\}$$

entonces rechazando las hipótesis nulas correspondientes a $p_{(1)}, \cdots, p_{(\hat{k})}$ proporciona $FDR = \alpha \cdot k_0 / K \leq \alpha$

rohen

En la expresión de *FDR* el término $E\left[\frac{V}{R}\middle|R>0\right]$ es lo que se define como *positive false discovery rate* (*pFDR*) en el cual se condiciona que al menos un positivo haya ocurrido

Por ello, cuando la prueba estadística viene de una mezcla aleatoria de las distribuciones de la hipótesis nula y la alternativa, *pFDR* puede escribirse como una probabilidad *a-posteriori* (Bayes).

Bajo condiciones generales, *V/R*, *FDR* y *pFDR* convergen simultáneamente en todas las regiones de significancia a la forma de *pFDR* Bayesiana.

*pFDR* puede ser usado para definir el *valor-q* que es el análogo al *valor-p* de las pruebas.

rohen

## Interpretación Bayesiana

*pFDR* se calcula sobre una región de significancia fija. Suponga que deseamos calcular *K* pruebas idénticas de una hipótesis nula versus una alternativa basada en las estadística $T_1, T_2, \cdots, T_K$ . Para una región de significancia $\Gamma$, defimos

$$pFDR(\Gamma) = E\left[\frac{V(\Gamma)}{R(\Gamma)} \middle| R > 0\right]$$

Donde $V(\Gamma) = \#\{nulas T_i : T_i \in \Gamma\}$ y $R(\Gamma) = \#\{T_i : T_i \in \Gamma\}$

Sea $H_i = 0$ cuando la *i-ésima* hipótesis nula es verdadera y $H_i = 1$ cuando es falsa ($i = 1, \ldots, K$). Sea $\pi_0$ la probabilidad *a-priori* de que una hipótesis sea cierta y ($1 - \pi_0$) la de que sea falsa (Bernoulli).

rohen

$$pFDR = P\left(H = 0 \middle| T \in \mathbf{\Gamma}\right)$$

$$= \frac{\pi_o P\left(T \in \mathbf{\Gamma} \middle| H = 0\right)}{\pi_o P\left(T \in \mathbf{\Gamma} \middle| H = 0\right) + \left(1 - \pi_o\right) P\left(T \in \mathbf{\Gamma} \middle| H = 1\right)}$$

$$= \frac{\pi_o \left[ error\ tipo\ I\ de\ \mathbf{\Gamma} \right]}{\pi_o \left[ error\ tipo\ I\ de\ \mathbf{\Gamma} \right] + \left(1 - \pi_o\right) \left[ potencia\ de\ \mathbf{\Gamma} \right]}$$

Al aumentar el *error tipo I*, aumenta *pFDR* y al aumentar la *potencia en* $\Gamma$, disminuye *pFDR*

rohen

Al escribir *pFDR* como $P\left(H=0 \middle| T \in \Gamma\right)$ está relacionada con el *error tipo I* y la llamamos *error tipo I a-posteriori* (Bayes). Esta cantidad nos da una medida global en la que no se proporciona información específica de los valores de cada estadística, solo si cae o no en la región de rechazo Γ.

***Valor-q***

Proporciona una medida de error de la hipótesis a probar, para cada estadística observada con respecto a *pFDR*.

rohen

## Ejemplo para calcular *q-value*

$$(T_i, H_i) \text{ v.a.i.i.d. con } T_i | H_i \sim (1 - H_i) \cdot N(0,1) + H_i \cdot N(2,1)$$

Dado que observamos las variables aleatorias

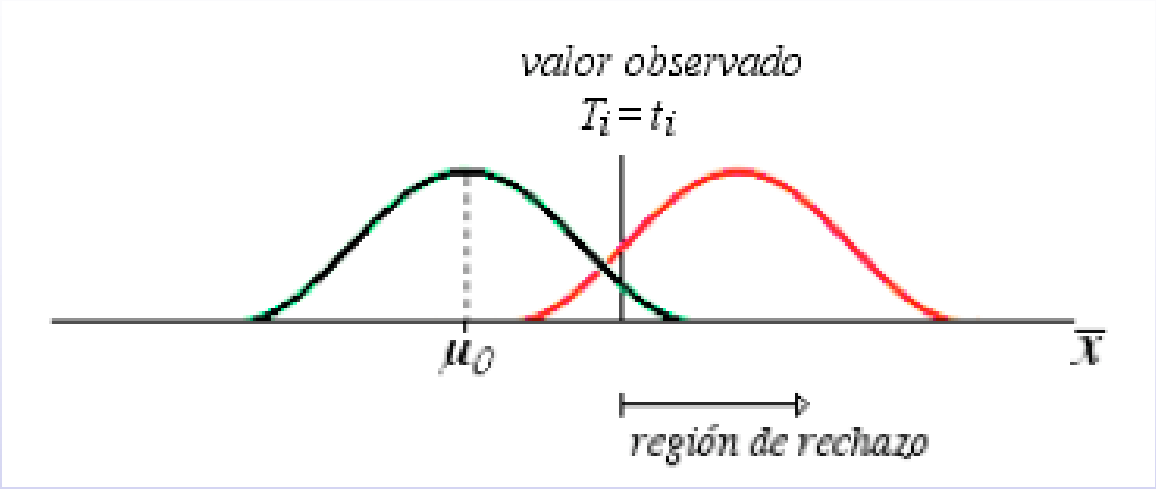
$T_1 = t_1, \cdots T_K = t_K$ el *valor* – $p$ puede ser calculado como

$$valor - p(t_i) = P(T_i \geq t_i | H = 0) = P(N(0,1) \geq t_i)$$

*Valor-p($t_i$)* da la *tasa de error tipo I* si rechazamos cualquier estadística como un valor más extremo de $t_i$

Cf. ecuación diapositiva 21

rohen




$pFDR(T \geq t_i)$ involucra el cálculo del área bajo *Ho* a la derecha de $t_i$ (*valor-p*) y el área bajo $H_1$ a la derecha de $t_i$ (potencia de la prueba).

Entonces $pFDR(T \geq t_i)$ es lo que llamamos *valor-q*($t_i$) (una sola región de significancia)

rohen

Definiendo

$$valor-q = \inf_{\Gamma_\alpha : t \in \Gamma_\alpha} pFDR(\Gamma_\alpha) = \inf_{\Gamma_\alpha : t \in \Gamma_\alpha} P\left(H = 0 \middle| T \in \Gamma_\alpha\right)$$

vemos que el *valor-q* es una versión Bayesiana del *valor-p*

*p-value a-posteriori*. (Bayes)

rohen

# Referencias


http://ftp.medprev.uma.es/libro/html.htm

http://www.hrc.es/bioest/M_docente.html

http://ocw.jhsph.edu/Topics.cfm?topic_id=33

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1199583/?tool=pubmed

http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aos/1074290335


rohen