Distribuciones χ^2, F y t no centrales

Def: Sean $X_1, \cdots, X_\nu$ va. independientes tales que $X_i \sim N(\xi_i, 1)$. Luego:

$$U = \sum_{i=1}^{\nu} X_i^2 \sim \chi'^2_{\nu,\delta}$$

donde el parámetro de no centralidad es $\delta = \left(\sum_{i=1}^{\nu} \xi_i^2 \right)^{1/2}$.

Se puede ver que si $Y_i \sim N(0,1)$ independientes entonces:

$$U = (Y_1 + \delta)^2 + \sum_{i=2}^{\nu} Y_i^2$$
$$U = (Y_1 + \delta)^2 + \chi^2_{\nu-1}$$

Propiedades:

$$E(\chi'^2_{\nu,\delta}) = \nu + \delta$$

$$Var(\chi'^2_{\nu,\delta}) = 2\nu + 4\delta^2$$

Suma de χ^2 no centrales independientes:

Si $U_1 \sim \chi'^2_{\nu_1,\delta_1}$ independiente de $U_2 \sim \chi'^2_{\nu_2,\delta_2}$, entonces

$$U_1 + U_2 \sim \chi'^2_{\nu_1+\nu_2,(\delta_1^2+\delta_2^2)^{1/2}}$$

Distribución F no central:

Def: Si $U_1 \sim \chi'^2_{\nu_1,\delta_1}$ independiente de $U_2 \sim \chi^2_{\nu_2}$, tenemos que

$$\frac{U_1/\nu_1}{U_2/\nu_2} \sim F'_{\nu_1,\nu_2,\delta_1}$$

es decir, F no central de ν_1 y ν_2 grados de libertad y parámetro de no cen-

tralidad δ_1.

Distribución t no central:

Def: Sean $X \sim N(\delta, 1)$ independiente de $U \sim \chi^2_\nu$, tenemos que

$$\frac{X}{\sqrt{U/\nu}} \sim t'_{\nu,\delta}$$

es decir, t no central con ν y parámetro de no centralidad δ.

Observación: Notemos que $t'_{\nu,\delta} = F'_{1,\nu,\delta}$

Potencia del test de F

Consideremos la base ortonormal de IR^n:

$$\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q, \boldsymbol{\alpha}_{q+1}, \ldots, \boldsymbol{\alpha}_r, \boldsymbol{\alpha}_{r+1}, \ldots, \boldsymbol{\alpha}_n$$

donde

$$\mathcal{V}_{r-q} : \{\boldsymbol{\alpha}_{q+1}, \ldots, \boldsymbol{\alpha}_r\}$$

$$\mathcal{V}_r : \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q, \boldsymbol{\alpha}_{q+1}, \ldots, \boldsymbol{\alpha}_r\}$$

Por lo tanto,

$$\mathbf{y} \in \Re^n \Longrightarrow \mathbf{y} = \sum_{j=1}^{n} z_j \boldsymbol{\alpha}_j \Longrightarrow \boldsymbol{\alpha}_i' \mathbf{y} = z_i$$

y si definimos a $\mathbf{T}$ como la matriz que tiene filas $\boldsymbol{\alpha}_i'$, entonces

$$\mathbf{z} = \mathbf{T}\mathbf{y}$$

Observemos que bajo el modelo Ω

$$z_i \text{ independientes y } z_i \sim N(\xi_i, \sigma^2)$$

donde

$$\xi_{r+1} = \xi_{r+2} = \cdots = \xi_n = 0$$

Bajo el modelo restringido ω, tenemos que

$$\xi_1 = \xi_2 = \cdots = \xi_q = 0$$

Usamos el estadístico F:

$$\frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2}{qs^2} = \frac{\sum_{i=1}^{q} z_i^2}{qs^2}$$

donde

$$s^2 = \frac{\sum_{i=r+1}^{n} z_i^2}{n-r}$$

Ya probamos que $\{z_1, \cdots, z_q\}$ y $\{z_{r+1}, \cdots, z_n\}$ son independientes y como $E(z_i) = 0$ si $i \geq r+1 \Longrightarrow$

$$\sum_{i=r+1}^{n} \left(\frac{z_i}{\sigma}\right)^2 = \frac{(n-r)s^2}{\sigma^2} \sim \chi^2_{n-r}$$

Sin embargo, si H_0 es cierta

$$\sum_{i=1}^{q} \left(\frac{z_i}{\sigma}\right)^2 \sim \chi^2_q$$

de lo contrario

$$\frac{z_i}{\sigma} \sim N(\frac{\xi_i}{\sigma}, 1)$$

$$\Downarrow$$

$$\sum_{i=1}^{q} \left(\frac{z_i}{\sigma}\right)^2 \sim \chi'^2_{q,\delta} \quad \text{con } \delta^2 = \sum_{i=1}^{q} \left(\frac{\xi_i}{\sigma}\right)^2$$

Por lo tanto, si H_0 no es cierta

$$F = \frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2}{qs^2} \sim F'_{q,n-r,\delta}$$

y la potencia del test será:

$$P(F'_{q,n-r,\delta} \geq F_{q,n-r,\alpha})$$

donde

$$\delta = \sqrt{\sum_{i=1}^{q} \left(\frac{\xi_i}{\sigma}\right)^2}$$

¿Cómo se calcula δ en términos de los parámetros originales?

$$\mathbf{z} = \mathbf{T}\mathbf{y} \Longrightarrow z_i = \boldsymbol{\alpha}_i'\mathbf{y} = \sum_{j=1}^{n} \alpha_{ij} y_j \Longrightarrow \xi_i = E(z_i) = \boldsymbol{\alpha}_i'\eta$$

en consecuencia

$$\xi_i = E(z_i) = \sum_{j=1}^{n} \boldsymbol{\alpha}_{ij}\eta_j$$

Tenemos las siguientes igualdades:

$$\begin{aligned} \|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 &= \sum_{i=1}^{q} z_i^2 \\ \sigma^2\delta^2 &= \sum_{i=1}^{q} \xi_i^2 \end{aligned}$$

y reemplazando a las z_i's obtenemos

$$\begin{aligned} \|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 &= \sum_{i=1}^{q} \left(\sum_{j=1}^{n} \alpha_{ij} y_j \right)^2 \\ \sigma^2\delta^2 &= \sum_{i=1}^{q} \left(\sum_{j=1}^{n} \alpha_{ij} \eta_j \right)^2 \end{aligned}$$

Con lo cual obtenemos la **Regla 1**: Bajo el modelo Ω

obtenemos $\sigma^2\delta^2$ reemplazando en la suma de cuadrados $\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}_\omega}\|^2$ cada Y_i por su valor esperado.

Cuadrados Medios

En el denomirador del estadístico F tenemos: $s^2 = \frac{\|\mathbf{y} - \widehat{\boldsymbol{\eta}}\|^2}{n-r}$ y su esperanza es σ^2.

En el numerador del estadístico F tenemos:

$$\frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2}{q} = \frac{\sum_{i=1}^{q} z_i^2}{q}$$

luego

$$E\left(\frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2}{q}\right) = E\left(\frac{\sum_{i=1}^{q} z_i^2}{q}\right)$$
$$= \frac{1}{q}\sum_{i=1}^{q} E(z_i^2)$$

$$= \frac{1}{q}\sum_{i=1}^{q}(\sigma^2 + \xi_i^2)$$
$$= \sigma^2 + q^{-1}\sigma^2\delta^2$$

Podemos calcular $\sigma^2\delta^2$ con la **Regla 1**. Observemos que en realidad aquí no es necesaria la normalidad, sólo alcanza con tener el modelo

$$\Omega' : E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad \Sigma = \sigma^2\mathbf{I}$$

¿Cómo quedaría en el caso de regresión lineal?

$$\Omega : Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ independientes}$$

Consideremos

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Bajo $\omega = \Omega \cap H_0$ tenemos que $Y_i = \beta_0 + \epsilon_i$, entonces el estimador de mínimos cuadrados será $\widehat{\beta}_0 = \overline{Y}$.

Para calcular la potencia necesitamos:

$$\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 = \sum_{i=1}^{n} (\widehat{\beta_0} + \widehat{\beta_1} x_i - \overline{Y})^2$$

Usando la Regla 1, reemplazamos por los valores esperados bajo Ω:

$$\begin{aligned}
\sigma^2\delta^2 &= \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i - \frac{\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)}{n} \right)^2 \\
&= \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \overline{x})^2 \\
&= \sum_{i=1}^{n} {\beta_1}^2 \ (x_i - \overline{x})^2 \\
&= {\beta_1}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2
\end{aligned}$$

por lo tanto

$$\delta^2 = \frac{{\beta_1}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2}{\sigma^2}$$

Análisis de la Varianza de 1 Factor (ANOVA 1)

En el Análisis de la Varianza de 1 Factor nos interesa comparar las medias de k poblaciones . Supongamos que tenemos k poblaciones y llamamos $\beta_1, \cdots, \beta_k$ a sus medias y que además cada población se distribuye según una normal y todas tienen la misma varianza σ^2.

Es decir, observamos

$$\begin{array}{lll} y_{11}, y_{12}, \ldots, y_{1j} \quad \ldots, y_{1n_1} & \sim & N(\beta_1, \sigma^2) \\ y_{21}, y_{22}, \ldots, y_{2j} \quad \ldots, y_{2n_2} & \sim & N(\beta_2, \sigma^2) \\ & \cdots & \\ y_{k1}, y_{k2}, \ldots, y_{kj} \quad \ldots, y_{kn_k} & \sim & N(\beta_k, \sigma^2) \end{array}$$

donde y_{ij} es la j-ésima observación de la i-ésima población, todas independientes. En total se tienen $n = \Sigma_{i=1}^{k} n_i$ observaciones.

Podemos escribir este modelo como.

$$\begin{aligned} y_{ij} &= \beta_i + \epsilon_{ij} \quad i =, \cdots, k \quad j = 1, \cdots, n_i \\ \epsilon_{ij} &\sim N(0, \sigma^2) \quad \text{independientes} \end{aligned}$$

Deseamos testear:

$$H_0 : \beta_1 = \cdots = \beta_k \quad H_1 : \text{ existen } i \neq j : \beta_i \neq \beta_j$$

Podríamos escribir esto en forma matricial definiendo:

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \dots \\ \dots \\ \dots \\ y_{2n_2} \\ \dots \\ y_{k1} \\ y_{k2} \\ \dots \\ y_{kn_k} \end{pmatrix} ; \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ . & . & . & . & . \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 1 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} ; \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \dots \\ \dots \\ \beta_k \end{pmatrix}$$

donde $rg(X) = k$. En consecuencia en este modelo todas las funciones de la forma $\mathbf{c}'\boldsymbol{\beta}$ son estimables.

Ejemplo (ANOVA 1) En la siguiente tabla se muestran los porcentajes de contenido de ácidos grasos no saturados activos (PAPFUA) presentes en 6 margarinas dietéticas:

IMPERIAL	PARKAY	BLUE BONNET	CHIFFON	MAZOLA	FLEISCHMANN'S
14.1	12.8	13.5	13.2	16.8	18.1
13.6	12.5	13.4	12.7	17.2	17.1
14.4	13.4	14.1	12.6	16.4	18.3
14.3	13	14.3	13.9	17.3	18.4
	12.3			18	

Las preguntas que se plantean los investigadores son:

a) Se desea saber si hay diferencias en los contenidos medios de PAPFUA de las 6 margarinas consideradas.

b) La margarinas Mazola y Fleischmann's son de tipo cereal, mientras que las otras son de tipo soja. Interesa obtener un intervalo de confianza para $\frac{\beta_1+\beta_2+\beta_3+\beta_4}{4} - \frac{\beta_5+\beta_6}{2}$.

En este caso $k = 6$, $n_1 = n_3 = n_4 = n_6 = 4$ y $n_2 = n_5 = 5$, por lo tanto $n = 26$

Volvamos al caso general

Buscamos minimizar:

$$\mathcal{S}(\boldsymbol{\beta}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \beta_i)^2$$

luego

$$\frac{\partial \mathcal{S}(\boldsymbol{\beta})}{\partial \beta_r} = -2 \sum_{j=1}^{n_r} (y_{rj} - \beta_r) = 0 \quad r = 1, \cdots, k$$

Por lo tanto, para cada $r = 1, \ldots, k$

$$\widehat{\beta}_r = \frac{\sum_{j=1}^{n_r} y_{rj}}{n_r} = \overline{Y}_{r.}$$

Por otro lado, minimizar bajo $\omega = \Omega \cap H_0$ es buscar el mínimo de

$$\mathcal{S}^*(\beta) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \beta)^2$$

luego

$$\frac{\partial \mathcal{S}^*(\beta)}{\partial \beta} = -2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \beta) = 0$$

y en consecuencia

$$\widehat{\beta} = \frac{\Sigma_{i=1}^{k} \Sigma_{j=1}^{n_i} y_{ij}}{n} = \overline{Y}_{..} \qquad (= \frac{\Sigma_{i=1}^{k} n_i \overline{y}_{i.}}{n})$$

Para calcular el estadístico F necesitamos:

$$\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{Y}_{i.} - \overline{Y}_{..})^2 = \sum_{i=1}^{k} n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2$$

$$\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2 = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

Suma de Cuadrados Entre Grupos$=\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2$: es una medida pesada dispersión de las k poblaciones respecto de la media general.

Suma de Cuadrados Dentro de los Grupos$=\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2$: es una medida combinada

de la dispersión dentro de cada muestra.

La hipótesis nula a testear

$$H_0 : \beta_1 = \cdots = \beta_k$$

se puede escribir

$$H_0 : \beta_2 - \beta_1 = \cdots = \beta_k - \beta_1 = 0$$

que es de la forma

$$\mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ . & . & . & & \\ . & . & . & & \\ . & . & . & & \\ -1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ . \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ . \\ . \\ . \\ . \\ 0 \end{pmatrix}$$

donde $rg(\mathbf{C}) = k - 1$, luego $q = k - 1$ y por lo tanto, el estadístico del test será:

$$F = \frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2/(k-1)}{\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}_\omega\|^2/(n-k)}$$

y rechazaremos H_0 si

$$F > F_{k-1,n-k,\alpha}$$

Con todo esto podemos armar la Tabla de Análisis de la Varianza de 1 Factor que es la salida típica de muchos programas que se utilizan para calcular este test (ver Cuadro 2).

SC		g.l.	M.S.	E(M.S.)	F
Entre	$\sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$k-1$	$(1) = \frac{\sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2}{k-1}$	$\sigma^2 + (k-1)^{-1}\sum_{i=1}^{k} n_i(\beta i - \overline{\beta}_{..})^2$	(1)/(2)
Dentro	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$	$n-k$	$(2) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2}{n-p}$	σ^2	
Tot. Cor.	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2$	$n-1$			

Cuadro 2: Tabla de ANOVA

Bajo Ω, F tiene una distribución $\mathcal{F}$ no central con parámetro de no centralidad

dado por la **Regla 1**:

$$\sigma^2\delta^2 = \sum_{i=1}^{k} n_i(\beta_i - \overline{\beta}_{..})^2$$

donde $\overline{\beta}_{..} = \Sigma_{i=1}^{k} \frac{n_i}{n}\beta_i$

Si la hipótesis de igualdad de medias es rechazada, seguramente nos deseare-mos identificar aquellas β_i que difieren entre sí, estaremos interesados en las diferencias $\beta_i - \beta_j$.

Otras veces, como en el ejemplo, podrían interesarnos algunas combinaciones particulares, tales como

$$\beta_1 - \frac{\beta_2 + \beta_3}{2} \quad \text{o} \quad \frac{1}{2}(\beta_1 + \beta_2) - \frac{1}{3}(\beta_3 + \beta_4 + \beta_5)$$

Estas son combinaciones lineales de los parámetros de la forma:

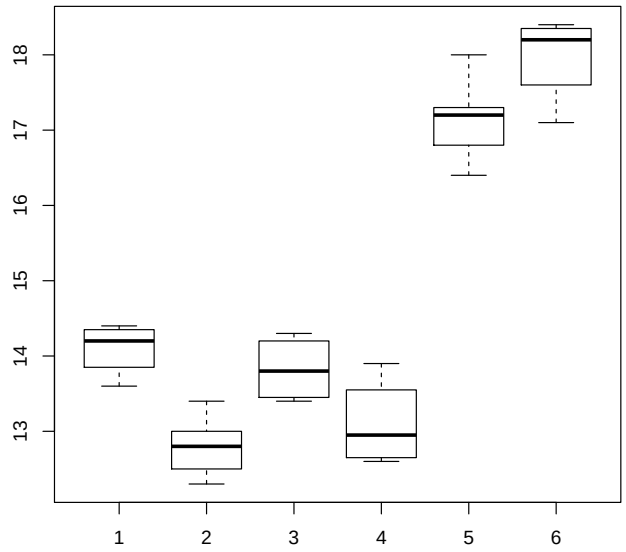
$$\sum_{i=1}^{p} c_i\beta_i \quad \text{con} \quad \sum_{i=1}^{p} c_i = 0$$

Estas combinaciones reciben el nombre de contrastes. Podríamos utilizar cualquiera

de los métodos vistos, si estuviéramos interesados en muchos contrastes el métodos de Scheffé podría ser el más adecuado. Para algunos casos particulares veremos el método introducido por Tukey.

Por ahora volvamos al ejemplo:

```
margarinas<-read.table("C:\\Users\\Ana\\ModeloLineal\\doctex\\margarinas.txt",header=T)
margarinas
   PAPFUA TIPO
1    14.1    1
2    13.6    1
3    14.4    1
4    14.3    1
5    12.8    2
6    12.5    2
7    13.4    2
.
.
25   18.3    6
26   18.4    6
attach(margarinas)
tipo.f<- factor(TIPO)
plot(tipo.f,PAPFUA)
```




```
salida<- aov(PAPFUA~tipo.f)
anova(salida)

Analysis of Variance Table

Response: PAPFUA
```

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
tipo.f     5 104.992 20.9984  79.736 1.642e-12 ***
Residuals 20   5.267  0.2634
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Como el p-valor es pequeñísimo el test de F rechaza la hipótesis de igualdad de medias.

Tests simultáneos para diferencias de medias

Bonferroni: $\alpha = 0{,}05 \quad \frac{\alpha}{2\binom{6}{2}} = 0{,}05/30 = 0{,}002$. Cada intervalo es de la forma:

$$\overline{y}_{i.} - \overline{y}_{j.} \pm t_{20,0,002}\, s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

donde $t_{20,0,002} = 3{,}331$

6	5	1	3	4	2
17,975	17,140	14,100	13,825	13,100	12,800
			- - - - - - - - - - - - - - - - -		

Hay tres grupos de medias que no son significativamente diferentes.

Scheffé: $\alpha = 0{,}05$

Vamos a probar que en contxto del modelo $y_{ij} = \beta_i + \epsilon_{ij}$, $\beta_j - \beta_1 .$, $j = 2, \ldots, k$ es una base de dimensión $k-1$ que genera el subespacio de todos los contrastes y por lo tanto

la probabilidad de que todos los contrastes satisfagan simultáneamente las desigualdades

$$\widehat{\psi} \pm \sqrt{(k-1)F_{k-1,n-k,\alpha}} s \sqrt{\sum_{i=1}^{k} c_i^2/n_i}$$

es $1-\alpha$

Cada intervalo es de la forma:

$$\overline{y}_{i.} - \overline{y}_{j.} \pm \sqrt{(k-1)F_{k-1,n-k,0,05}}\; s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$\overline{y}_{i.} - \overline{y}_{j.} \pm \sqrt{5F_{5,20,0,05}}\, s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

donde $F_{5,20,0,05} = 2{,}71$

6	5	1	3	4	2
17,975	17,140	14,100	13,825	13,100	12,800

La conclusión es la misma.

Ejercicio Adicional de la Práctica 3: programar estos dos tipos de intervalos.

Intervalo de Confianza para el contraste buscado en b)

Bonferroni: $\alpha = 0{,}05$

El intervalo es de la forma general:

$$\widehat{\psi} \pm t_{n-r,0,05/2} \sqrt{\widehat{var}(\widehat{\psi})}$$

y en este caso

$$\widehat{\psi} \pm t_{n-k,0,05/2} \sqrt{\widehat{var}(\widehat{\psi})}$$

donde $t_{20,0,05} = 2{,}086$

Tenemos que

$$\begin{aligned} \widehat{\psi} &= \frac{\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4}{4} - \frac{\hat{\beta}_5 + \hat{\beta}_6}{2} \\ &= \frac{\overline{y}_{1.} + \overline{y}_{2.} + \overline{y}_{3.} + \overline{y}_{4.}}{4} - \frac{\overline{y}_{5.} + \overline{y}_{6.}}{2} \\ &= -4{,}1015 \end{aligned}$$

Además:

$$\widehat{var}(\widehat{\psi}) = s^2 \left(\frac{1}{16} \left(\frac{1}{4} + \frac{1}{5} + \frac{1}{4} + \frac{1}{4} \right) + \frac{1}{4} \left(\frac{1}{5} + \frac{1}{4} \right) \right) = 0{,}0473$$

El inetervalo resultante es

$$(-4{,}1015 - 2{,}086 * 0{,}0217, -4{,}1015 + 2{,}086 * 0{,}0217) = (-4{,}199972, -4{,}002528)$$

Otra parametrización

Otra manera de escribir el modelo sería

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

donde:

μ: es el efecto general

α_i: es el efecto del tratamiento i

En ese caso tendríamos

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \cdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \cdots \\ y_{2n_2} \\ . \\ . \\ y_{k1} \\ y_{k2} \\ \cdots \\ y_{kn_k} \end{pmatrix}; \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ . & . & . & . \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ . & . & . & . \\ 1 & 0 & 1 & 0 \\ . & . & . & . \\ . & . & . & . \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ . & . & . & . \\ 1 & 0 & 0 & 1 \end{pmatrix}; \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_k \end{pmatrix}$$

¿Son todas las funciones estimables en este modelo?

Es claro que la matriz de diseño $\mathbf{X}$ tiene $rg(\mathbf{X}) = k < p = k + 1$ y por lo tanto no toda función paramétrica es estimable.

Ya vimos en el caso $k = 3$ que, por ejemplo, α_1 no es estimable.

De acuerdo con el Teorema que probamos muchas clases atrás deberíamos incluir una restricción adicional. Para lograr la identificabilidad de los parámetros son frecuentes:

$$\alpha_k = 0 \quad \text{o} \quad \sum_{i=1}^{k} \alpha_i = 0 \quad \text{etc.}$$

Es muy usada la restricción $\Sigma_{i=1}^{k} \alpha_i = 0$, que es natural ya que:

$$\begin{aligned} \eta_{ij} &= E(y_{ij}) = \mu + \alpha_i = \mu + \overline{\alpha} + \alpha_i - \overline{\alpha} \\ &= \tilde{\mu} + \tilde{\alpha}_i \end{aligned}$$

donde $\sum_{i=1}^{k} \tilde{\alpha}_i = 0$

Notemos que usando esta restricción tenemos que:

$$\eta_{ij} = E(y_{ij}) = \mu + \alpha_i \Longrightarrow \eta_{.j} = k\mu$$

$$\Longrightarrow \mu = \overline{\eta}_{.j}$$

por lo tanto

$$\alpha_i = \eta_{ij} - \overline{\eta}_{.j}$$

$\hat{\mu}$ y $\hat{\alpha}_i$ están unívocamente determinados por los $\hat{\eta}_{ij}$:

$$\hat{\mu} = \overline{\hat{\eta}}_{.j} \quad \hat{\alpha}_i = \hat{\eta}_{ij} - \overline{\hat{\eta}}_{.j}$$

Si quisiéramos plantear las ecuaciones normales para estimar los parámetros podríamos plantear:

$$\begin{aligned}
\frac{\partial S}{\partial \mu} &= -2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i) = 0 \\
\frac{\partial S}{\partial \alpha_i} &= -2 \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i) = 0 \\
\sum_{i=1}^{k} \alpha_i &= 0 \quad \leftarrow \text{restricción adicional}
\end{aligned}$$

Por lo tanto:

$$\begin{aligned} n\mu + \sum_{i=1}^{k} n_i \alpha_i &= \sum_{i=1}^{k} n_i \overline{y}_{i.} \\ \mu + \alpha_i &= \overline{y}_{i.} \\ \sum_{i=1}^{k} \alpha_i = 0 & \end{aligned}$$

Notemos que la primera ecuación es dependiente de las k siguientes. Luego:

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{k} \frac{\overline{y}_{i.}}{k} \\ \hat{\alpha}_i &= \overline{y}_{i.} - \sum_{j=1}^{k} \frac{\overline{y}_{j.}}{k} \end{aligned}$$

que están unívocamente determinados por los y_{ij}

Volviendo al tema de comparaciones múltiples

Método de Tukey

Mientras el método S de Scheffé utiliza la distribución F, este método usa *la distribución del rango studientizado* $q_{l,\nu}$, que presentaremos a continuación.

El método T sirve para realizar contrastes simultáneos que involucran I parámetros $(\theta_1, \ldots, \theta_I)$ con la restricción de que sus estimadores $\hat{\theta}_i$ tengan igual varianza. De allí, que en principio en el contexto de ANOVA 1 Factor asumiremos que $n_i = m \; \forall i = 1, ..., k$

Deduciremos el método para el caso en que $\hat{\theta}_i$ son independientes y los contrastes de interés de la forma $\theta_i - \theta_j$.

Def.: Distribución del rango studientizado $q_{I,\nu}$: Sean $x_1, x_2, \ldots, x_I$ v.a. independientes tales que $x_i \sim N(0,1)$, $R = \text{máx}_{1 \leq i \leq I} x_i - \text{mín}_{1 \leq i \leq I} x_i$ y $U \sim \chi^2_\nu$ independiente de las $x_i's$. Entonces:

$$\frac{\text{máx}_{1 \leq i \leq I} x_i - \text{mín}_{1 \leq i \leq I} x_i}{\sqrt{\frac{U}{\nu}}} = \frac{R}{\sqrt{\frac{U}{\nu}}} \sim q_{I,\nu}$$

Teorema de Tukey

Sean $\hat{\theta}_i$ v.a. independientes $1 \leq i \leq I$, tales que $\hat{\theta}_i \sim N(\theta_i, a^2\sigma^2)$, con $a > 0$ constante conocida y s^2 un estimador de σ^2, independiente de $\hat{\theta}_i \; \forall i$, y tal que

$\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu$. Entonces

La probabilidad de que todas los $\frac{1}{2}I(I-1)$ diferencias $\theta_i - \theta_j$ satisfagan simultáneamente

$$\hat{\theta}_i - \hat{\theta}_j - Ts \leq \theta_i - \theta_j \leq \hat{\theta}_i - \hat{\theta}_j + Ts$$

donde $T = aq_{I,\nu,\alpha}$ es $1-\alpha$.

Ejemplo: Supongamos que queremos comparar las medias de 4 tratamientos: T_1, T_2, T_3 y T_4 y nos interesan los contrastes:

$$\alpha_i - \alpha_j$$

que es equivalente a comparar $\beta_i - \beta_j$.

Sabemos que $\hat{\beta}_i = \overline{y}_{i.}$ y que $\overline{y}_{1.}, \ldots, \overline{y}_{4.}$ son independientes. Además $\overline{y}_{i.} \sim N(\beta_i, \frac{\sigma^2}{n_i})$. Para poder usar Tukey, entonces $n_i = m \; \forall i$.

Por lo tanto:

$$P(\cap_{i,j} \overline{y}_{i.} - \overline{y}_{j.} - q_{4,4m-4,\alpha} s\sqrt{\frac{1}{m}} \leq \beta_i - \beta_j \leq \overline{y}_{i.} - \overline{y}_{j.} - q_{4,4m-4,\alpha} s\sqrt{\frac{1}{m}})$$

Extensiones del Método de Tukey

1. Teorema de Tukey

Bajo las condicones del Teorema anterior la probabilidad de que todos los contrastes de la forma $\psi = \Sigma_{i=1}^{I} c_i\theta_i, \Sigma_{i=1}^{I} c_i = 0$ satisfagan simultáneamente

$$\hat{\psi} - Ts \sum_{i=1}^{I} |c_i|/2 \leq \psi \leq \hat{\psi} - Ts \sum_{i=1}^{I} |c_i|/2$$

donde $T = aq_{I,\nu,\alpha}$ y $\hat{\psi} = \Sigma_{i=1}^{I} c_i\hat{\theta}_i$, es $1 - \alpha$.

2. Método de Tukey–Kramer Para el caso de muestras de diferente tamaño hay diferentes propuestas para extender el método de Tukey. El método T–K aplicado al problema de ANOVA 1 Factor para n_i observaciones para cada nivel i, $i = 1, \cdots, k$, propone los intervalos

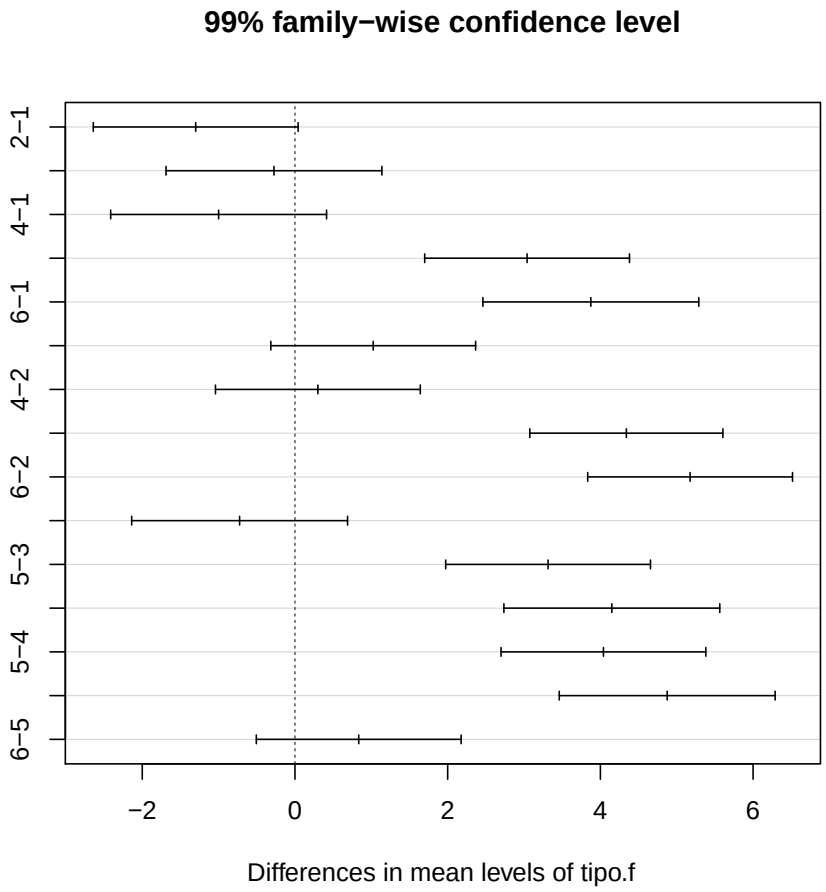
$$\overline{y}_{i.} - \overline{y}_{j.} - q_{k,n-k,\alpha} s \sqrt{\frac{1}{2}(\frac{1}{n_i} + \frac{1}{n_j})} \leq \beta_i - \beta_j \leq \overline{y}_{i.} - \overline{y}_{j.} - q_{k,n-k,\alpha} s \sqrt{\frac{1}{2}(\frac{1}{n_i} + \frac{1}{n_j}))}$$

Volvamos a nuestro ejemplo de las margarinas

```
salida<- aov(PAPFUA~tipo.f)

anova(salida)

FLUOR.tuk<-TukeyHSD(salida,"tipo.f",ordered=FALSE,conf.level=0.99)
plot(FLUOR.tuk)
```


99% family−wise confidence level
2−1
4−1
6−1
4−2
6−2
5−3
5−4
6−5
−2
0
2
4
6
Differences in mean levels of tipo.f

Comparación de los métodos para ANOVA 1 Factor

Notemos que ls tres tipos de intervalos son de la misma forma y que para comparar sus longitudes basta considerar:

$$r_{T,S} = \frac{long.Tukey}{long.Scheffe} = \frac{q_{k,\nu,\alpha} \sum_{i=1}^{k} |c_i|/2}{\sqrt{(k-1)F_{k-1,\nu,\alpha} \sum_{i=1}^{k} c_i^2}}$$

$$r_{B,S} = \frac{long.Bonferroni}{long.Scheffe} = \frac{t_{\nu,\alpha/(k(k-1))}}{\sqrt{(k-1)F_{k-1,\nu,,\alpha}}}$$

$$r_{T,B} = \frac{long.Tukey}{long.Bonferroni} = \frac{r_{T,S}}{r_{B,S}}$$

En la siguiente tabla extraída de Stapleton (1995) mostramos los cocientes para contrastes de la forma $\beta_i - \beta_j$ para $\alpha = 0{,}05$, $k = 3, 5, 7, 10$, $\nu = 10, \infty$.

Table 5.4.1 Ratios of Lengths Among Tukey, Bonferroni, and Scheffé SCI's

	$\nu = 10$				$\nu = \infty$			
	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 3$	$k = 5$	$k = 7$	$k = 10$
$r_{T,S}$	0.958	0.884	0.824	0.765	0.956	0.886	0.831	0.768
$r_{B,S}$	1.002	0.960	0.919	0.875	0.976	0.913	0.854	0.793
$r_{T,B}$	0.956	0.918	0.897	0.874	0.980	0.970	0.973	0.968