ANOVA

Introducción, conceptos fundamentales y "como funciona por dentro"

Universitat d'Alacant
Universidad de Alicante
Dept. of Marine Science and Applied Biology
Jose Jacobo Zubcoff

ANOVA

¿Qué es y para que sirve?

Compara la distribución de una variable continua normal en dos o más poblaciones

(niveles o categorías)

Pruebas de contraste para dos o más grupos independientes (ANOVA entre sujetos): un factor completamente aleatorizado.

ANOVA

Ejemplo de problema a resolver:

Se quiere estudiar la abundancia en tres zonas pesqueras, una de las cuales tiene cierto nivel de protección (pesquera).
Se espera que comparando los valores de esta zona con la de otras dos con parecidas características el hecho resultará evidente.

Hipótesis:

H_0 : No existen diferencias entre las *k* zonas

H_1 : Hipótesis nula no cierta (al menos alguna zona es diferente al resto)

ANOVA

H_0: No existen diferencias entre los *k* niveles

H_1: La hipótesis nula no es cierta

Hipótesis nula: (todas las medias poblacionales de los "k" grupos son iguales)
H0: μ1=μ2=μ3...=μa=μ

Hipótesis alternativa: (al menos una media poblacional difiere)
H1: No es cierto H0

- Parte de un conjunto de observaciones muestrales
- *K* niveles o categorías

ANOVA

Supongamos un universo de notas de 9 alumnos de 3 grupos distintos

Grupo 1	Grupo 2	Grupo 3
5	5	5
5	5	5
5	5	5

No hay diferencia ENTRE grupos

Ni DENTRO de los grupos

$$x_{i,j} = \mu$$

ANOVA

Supongamos que aplicamos un método de enseñanza (*factor*) que afecta:

Grupo 1	Grupo 2	Grupo 3
5+1=6	5+2=7	5+0=5
5+1=6	5+2=7	5+0=5
5+1=6	5+2=7	5+0=5

$$X_{i,j} = \mu + \alpha_i$$

Donde $\boldsymbol{\alpha_i} = \{1,2,0\}$ *efecto del factor*

El factor influye en establecer diferencias ENTRE grupos
Pero NO DENTRO

ANOVA

- Por razones ALEATORIAS algunos alumnos rinden mas que otros

Grupo 1	Grupo 2	Grupo 3
5+1-1 = 5	5+2+2 = 9	5+0+3 = 8
5+1-2 = 4	5+2+0 = 7	5+0+4 = 9
5+1+0 = 6	5+2+1 = 8	5+0+0 = 5

$$X_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

Donde $\varepsilon_{i,j}$= {-1,-2,0,2,0,1,3,4,0} *efecto aleatoriedad*

La ALEATORIEDAD influye en la variabilidad DENTRO de los grupos

ANOVA

Grupo 1	Grupo 2	Grupo 3	
5+1-1 = 5	5+2+2 = 9	5+0+3 = 8	
5+1-2 = 4	5+2+0 = 7	5+0+4 = 9	
5+1+0 = 6	5+2+1 = 8	5+0+0 = 5	
$X_{1.} = 5$	$X_{2.} = 8$	$X_{3.} = 7.33$	$X_{..} = 6.78$

Calculamos las medias por grupo y la media global

ANOVA

Grupo 1	Grupo 2	Grupo 3	
5	9	8	
4	7	9	
6	8	5	
$X_{1.} = 5$	$X_{2.} = 8$	$X_{3.} = 7.33$	$X_{..} = 6.78$

Para calcular el efecto **aleatorio**: medimos las diferencias **DENTRO**

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

ANOVA

Grupo 1	Grupo 2	Grupo 3	
5	9	8	
4	7	9	
6	8	5	
$X_{1.} = 5$	$X_{2.} = 8$	$X_{3.} = 7.33$	$X_{..} = 6.78$

Para calcular el efecto del **factor**: medimos las diferencias **ENTRE**

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

ANOVA

Tenemos dos tipos de variabilidad:

- ENTRE grupos (debida al factor)
- DENTRO grupos (debida a la aleatoriedad)

Para poder afirmar que el factor produce efectos:

La variabilidad ENTRE grupos debe ser significativamente grande respecto a la DENTRO grupos

ANOVA

Generalizando …

	1	2	Niveles del factor	k
1	$X_{1,1}$	$X_{2,1}$	…	$X_{k,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{i,j}$	$X_{k,2}$
j	$X_{1,j}$	$X_{2,j}$	…	$X_{k,j}$
n	$X_{1,n1}$	$X_{2,n2}$	…	$X_{k,nk}$

$i = 1,2,3,...,k$ $\quad$ $j = 1,2,3,..., n_k$ (no balanceado)

Media al nivel i del factor = $(1/n_i) \sum_{j=1} X_{i,j}$

Media general = $(1/N) \sum \sum X_{i,j}$ Siendo $N = \sum n_i$

ANOVA

$$X_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

Asumiendo las hipótesis previas:

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_k$

O bien si consideramos $Xi,j = \mu + \alpha i$

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

Se quiere comprobar la NO INFLUENCIA del factor α

- Todas las muestras proceden de la misma población

ANOVA

$$\left.\begin{array}{l} H_0: \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1: \textit{Al menos una igualdad no es cierta} \end{array}\right\}$$

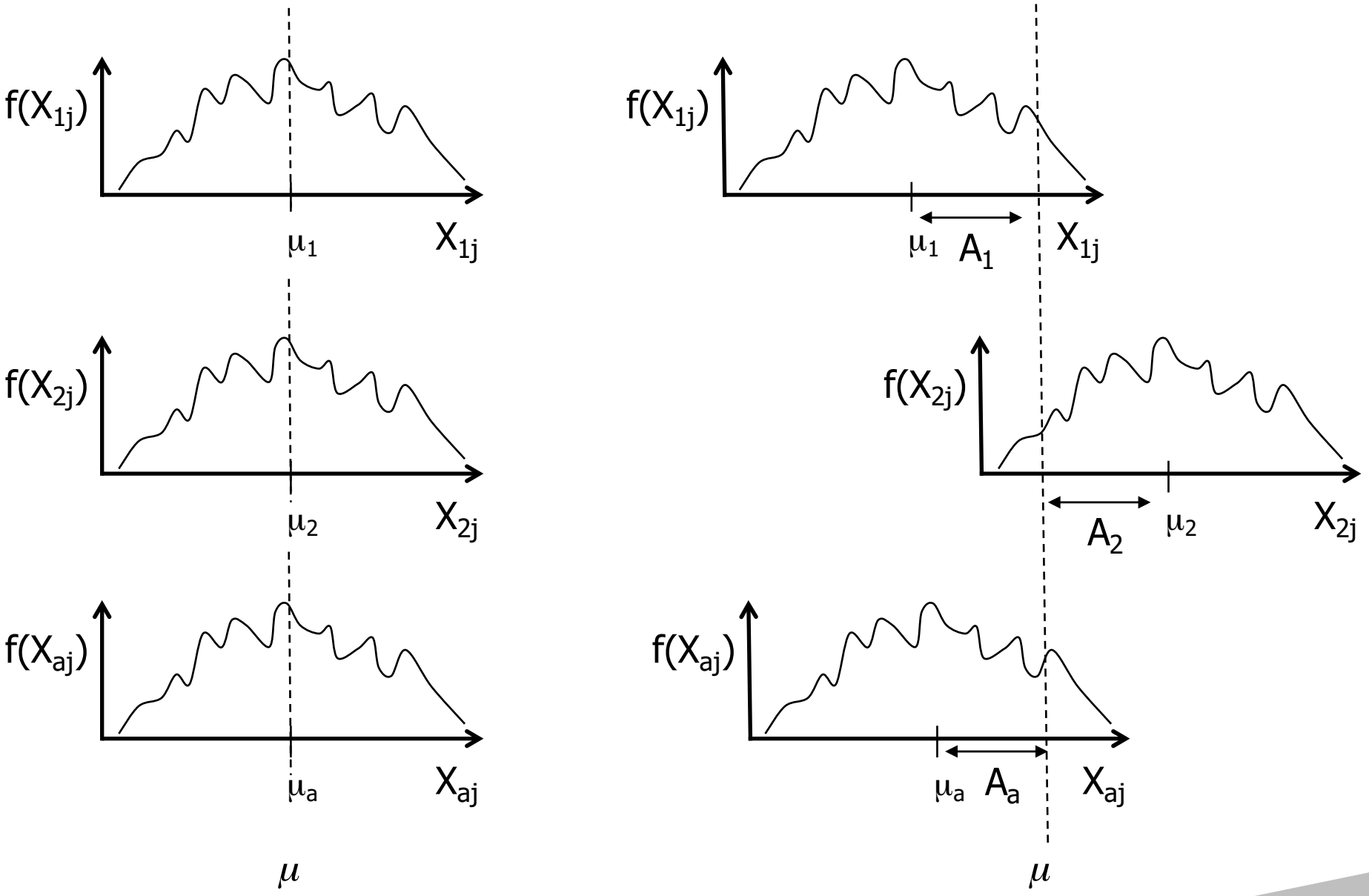
- Según la Hipótesis fijada

modelo probabilístico

NO se rechaza H_0 sí y solo sí:

$$F = \frac{\frac{Q_E}{k-1}}{\frac{Q_D}{n-k}} \leq F_{k-1,n-k,\alpha}$$

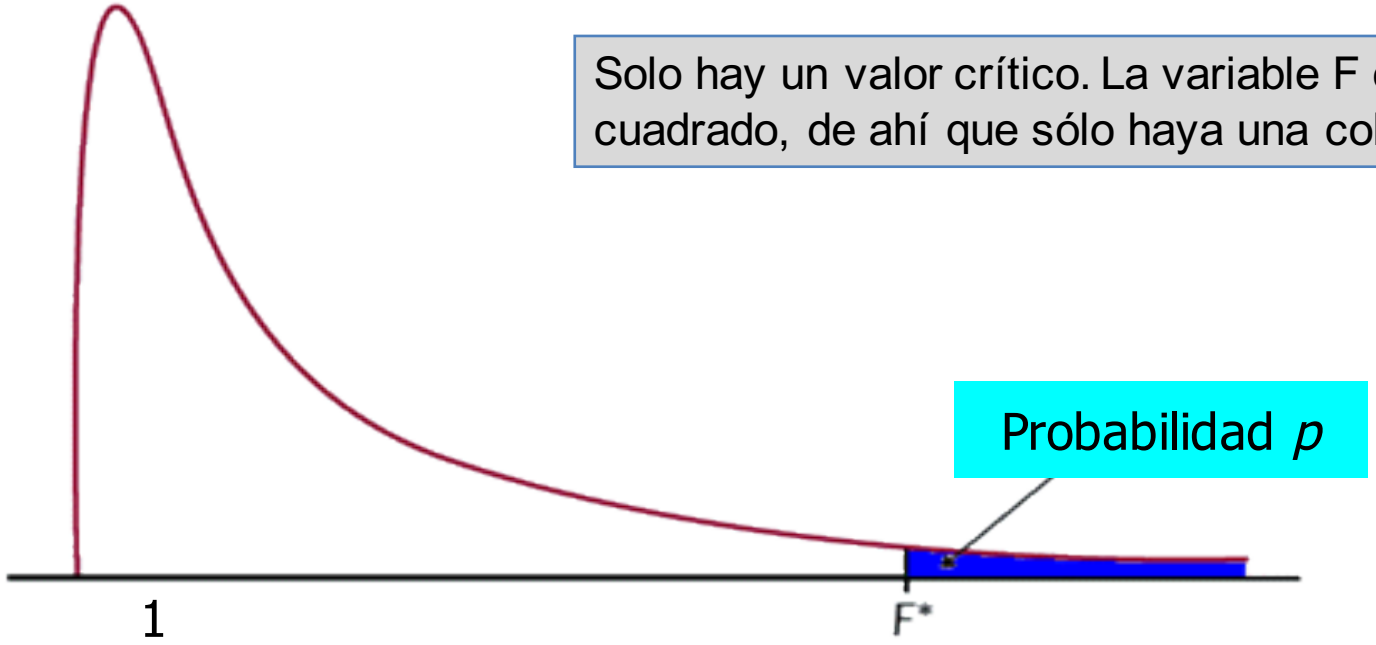
Análisis de la varianza de un factor:



Análisis de la varianza de un factor:

Construcción del estadístico de contraste:

MC_{ENTRE} / MC_{DENTRO} sigue una distribución F



Si el $F_{calc} > F_{crit}$ para (a-1) y a(n-1) g.l., se <u>rechaza</u> H_0:

al menos una de las medias X_i es significativamente diferente de las demás

ANOVA

Fuentes de variación	Sumas de cuadrados	G.L.	Cuadrados Medios	F	Signif.
ENTRE	$Q_E = \sum_{i=1}^{k} n_i \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2$	k–1	$Q_E / k-1 = S_E^2$	$F = \dfrac{Q_E / k-1}{Q_D / n-k}$	P-valor
DENTRO	$Q_D = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i.}\right)^2 = \sum_{i=1}^{k} (n-1) S_i^2$	n–k	$Q_D / n-k = S_D^2$		
TOTAL	$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{..}\right)$	n–1	$Q / n-1 = S^2$		

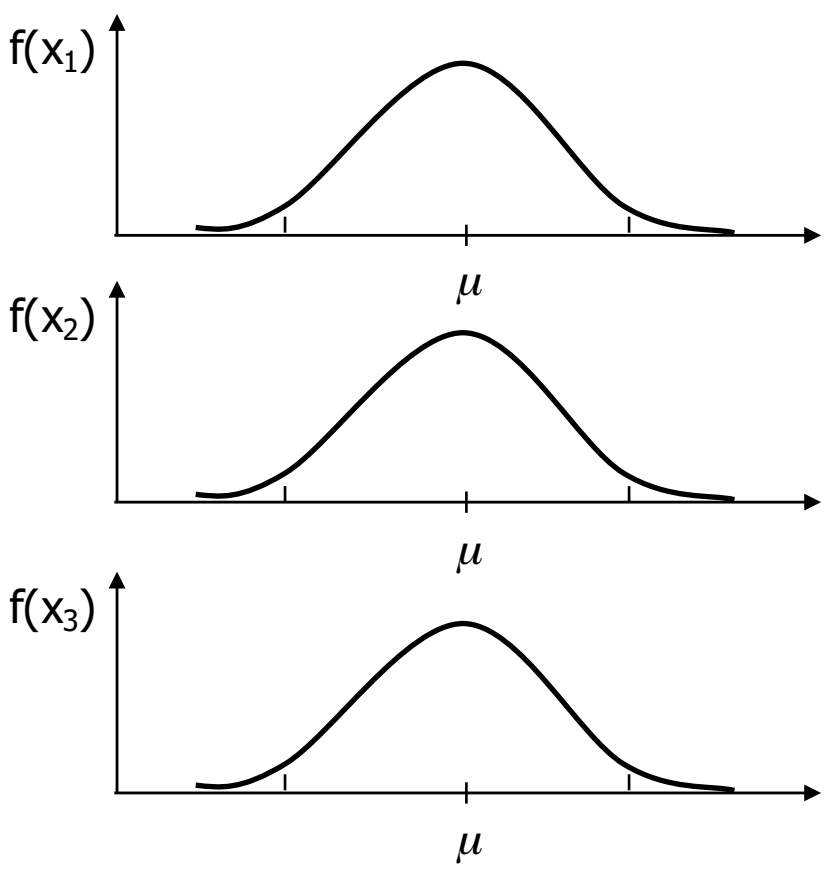
ANOVA

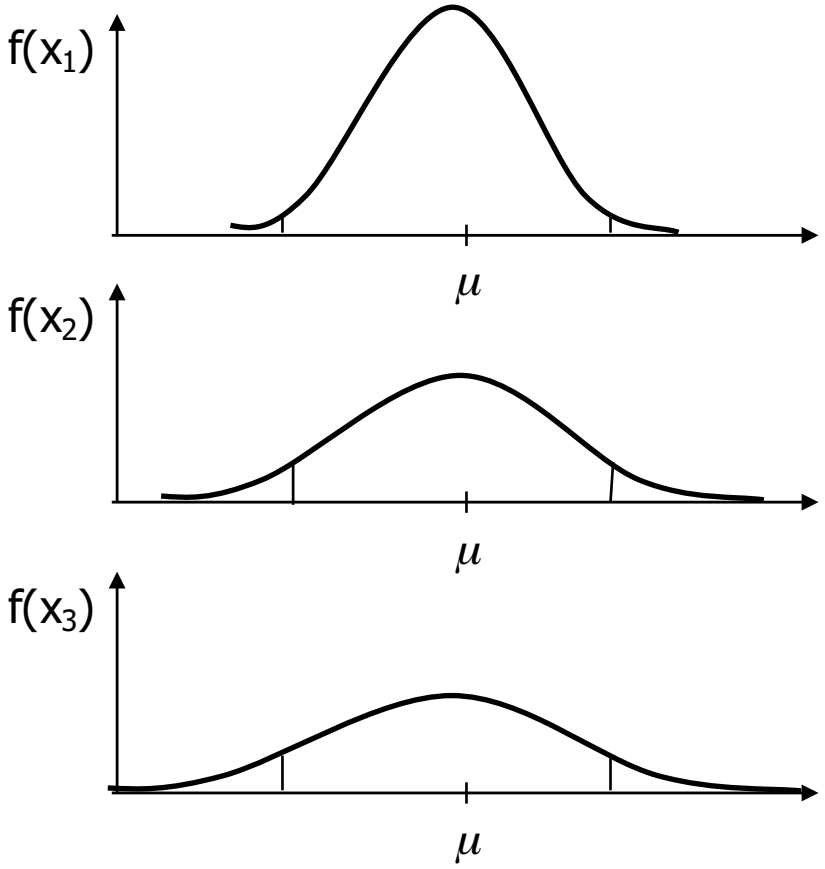
Hipótesis necesarias para realizar un ANOVA

a) Normalidad de la respuesta en cada nivel
b) Homogeneidad de las varianzas
c) Independencia de los valores obtenidos

Asunciones del análisis de la varianza:

Homogeneidad de varianzas:





Cuando varianzas ≠, se incrementa el error Tipo I

Asunciones del análisis de la varianza:

Homogeneidad de varianzas:

Test de heterogeneidad de varianzas:

Bartlett -> high alfa, sensible a NO normalidad
Levene -> es una ANOVA para VAR. Asume var iguales!!
Scheffe -> insensible a NO normalidad, pero no lo recomnieda …
Hartley ->problema cuando 1 var es pequeña….
… etc.

Test de Cochran

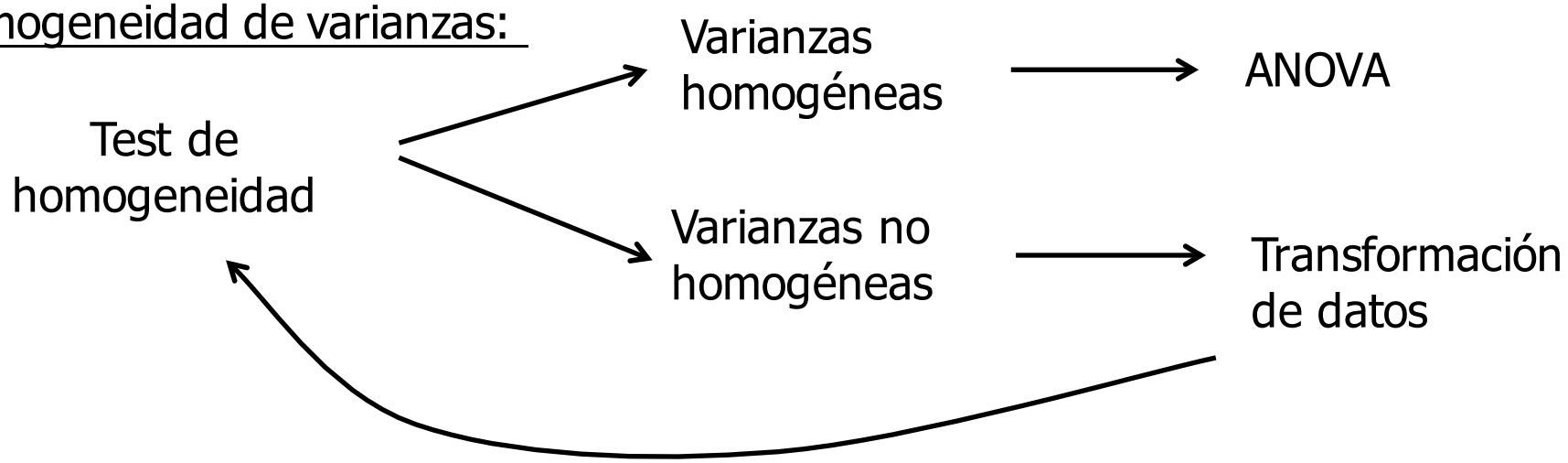
$$C = \frac{\text{mayor } s_i^2}{\sum s_i^2}$$

La distribución C para a (tratamientos) y (n-1) g.l. ha sido tabulada

Si $C_{obs} < C_c$, aceptamos la H_0 de homogeneidad de varianzas

Asunciones del análisis de la varianza:

Homogeneidad de varianzas:



Conteos (o datos que siguen una dist. de Poisson) ➔ $\sqrt{(X + 1)}$

Ratios, tasas, concentraciones, etc. ➔ log (X) o log (X + 1)

Porcentajes y proporciones ➔ $sen^{-1} \sqrt{X}$ (= arcsen X)

Asunciones del análisis de la varianza:

Transformación de datos

- Útil para eliminar heterogeneidad de la varianza.
- Sólo es efectivo si la media tiene una relación constante con la varianza.
- La transformación debe ser monotónica.
- Deben mantenerse las medias en el mismo orden.
- La transformación debe utilizarse únicamente para evitar el problema de heterogeneidad de la varianza.
- Transformaciones sistemáticas son perjudiciales.

Asunciones del análisis de la varianza:

Transformación de datos

Raíz cuadrada

- Poblaciones que siguen una **distribución de Poisson**: medias y varianzas son iguales
- Frecuencias o recuentos por unidad de tiempo o superficie.
- Principalmente con abundancias muy pequeñas.

$$\sqrt{x + 1}$$

Asunciones del análisis de la varianza:

Transformación de datos

Logarítmo

- Muestreos con valores muy altos: medias mayores y varianza mucho mayores.
- Datos distribuidos log-normal
- Medidas de tasas, concentraciones, relaciones,...
- Ej: Relación entre el número de presas comida por depredador, cantidad de clorofila por peso algal,...
- Independiente del tipo de logaritmo usado.
- Sumar una constante (1) para aplicar logaritmos por los valores que son 0.
- Problema en valores muy pequeños: solo cuando son mayores de 10.

$$\log (X+1)$$

Asunciones del análisis de la varianza:

Transformación de datos

Arcoseno

- **Porcentajes y proporciones**
- Distribución Binomial.
- Ej. Porcentaje de cobertura de *Posidonia*

$$\text{sen}^{-1}\sqrt{x}$$

Asunciones del análisis de la varianza:

Transformación de datos

logit transformation

- Transforma los valores de **porcentajes** desde - ∞ hasta + ∞

- La transformación de arcoseno limita los valores desde 0 hasta π / 2 radianes (0 hasta 90 º) .

$$(1/2) \log ((p / 1\text{-}p))$$

Asunciones del análisis de la varianza:

Si la transformación de datos no es posible

- Situaciones biológicas que presentan varianzas heterogéneas: gran agrupación de organismos.
- Cuando son experimentos bien replicados: el análisis de la varianza es suficientemente robusto.
- Experimentos grandes y balanceados.
- La validez del test y probabilidades asociadas con la distribución de la F ratio no se ven muy afectadas.

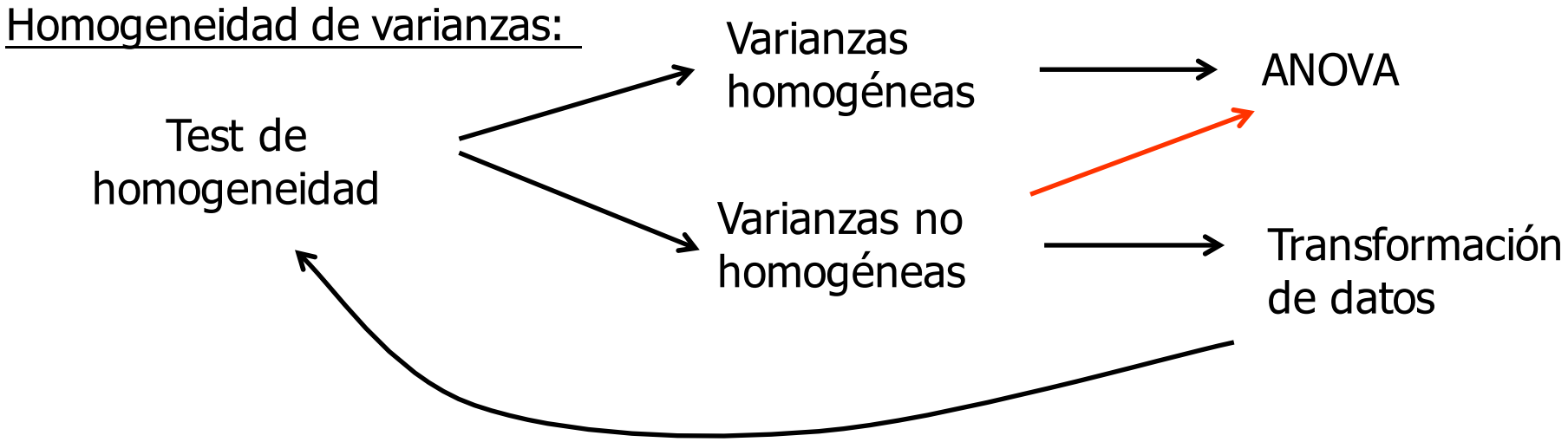
Asunciones del análisis de la varianza:

Si la transformación de datos no soluciona la Heterogeneidad

- Si se acepta la Ho no existe problema.
- Si se rechaza debe ser un α menor (0.01): así se evita error tipo I.

Utilizar un test **no paramétrico**
no soluciona el problema de
heterogeneidad de
varianzas
(debemos intentar **explicar** dicha heterogeneidad)

Asunciones del análisis de la varianza:



Si se acepta H_0 ($p > 0,05$), no hay problema

Si se rechaza H_0, considerar $\alpha_c = 0,01$

Utilizar un test no paramétrico (p.ej. Kruskal-Wallis) no soluciona el problema

Asunciones del análisis de la varianza:

Normalidad de los datos

El análisis de la varianza es suficientemente robusto a las desviaciones de la normalidad (Glass et al. 1972, Harwell et al. 1992, Lix et al. 1996)*, sobre todo cuando:

- hay un gran número de tratamientos y / o réplicas;
- los datos están equilibrados.

Las transformaciones a menudo corrigen el apuntamiento, pero cuidado cuando existe homogeneidad de varianzas.

(*) References

Glass, G.V., P.D. Peckham, and J.R. Sanders. 1972. Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. Rev. Educ. Res. 42: 237-288.

Harwell, M.R., E.N. Rubinstein, W.S. Hayes, and C.C. Olds. 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. J. Educ. Stat. 17: 315-339.

Lix, L.M., J.C. Keselman, and H.J. Keselman. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Rev. Educ. Res. 66: 579-619.

Schmider, Emanuel; Ziegler, Matthias; Danay, Erik; Beyer, Luzi; Bühner, Markus. 2010. *Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption*. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, Vol 6(4), 2010, 147-151. doi: 10.1027/1614-2241/a000016

Resumen del problema de la FALTA de independencia de datos

No independencia DENTRO de los tratamientos	No independencia ENTRE los tratamientos
Correlación positiva	
σ^2_e dentro de las muestras es subestimado F ratio excesivo Incremento del error Tipo I Diferencias sin importancia son detectadas	σ^2_e entre de las muestras es subestimado F ratio muy pequeno Incremento del error Tipo II Diferencias reales no son detectadas
Correlación negativa	
σ^2_e dentro de las muestras es sobreestimado F ratio demasiado pequeno Incremento del error Tipo II Diferencias reales no son detectadas	σ^2_e entre de las muestras es sobreestimado F ratio excesivo Incremento del error Tipo I Diferencias sin importancia son detectadas

Asunciones del análisis de la varianza:

Normalidad de los datos

El análisis de la varianza es suficientemente robusto a las desviaciones de la normalidad (Glass et al. 1972, Harwell et al. 1992, Lix et al. 1996)*, sobre todo cuando:

- hay un gran número de tratamientos y / o réplicas;
- los datos están equilibrados.

Las transformaciones a menudo corrigen el apuntamiento, pero cuidado cuando existe homogeneidad de varianzas.

(*) References

Glass, G.V., P.D. Peckham, and J.R. Sanders. 1972. Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. Rev. Educ. Res. 42: 237-288.

Harwell, M.R., E.N. Rubinstein, W.S. Hayes, and C.C. Olds. 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. J. Educ. Stat. 17: 315-339.

Lix, L.M., J.C. Keselman, and H.J. Keselman. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Rev. Educ. Res. 66: 579-619.

Schmider, Emanuel; Ziegler, Matthias; Danay, Erik; Beyer, Luzi; Bühner, Markus. 2010. *Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption*. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, Vol 6(4), 2010, 147-151. doi: 10.1027/1614-2241/a000016

Análisis *a posteriori*

Dos tipos de definición de H_A

- *a priori* (antes de realizar el experimento)
- *a posteriori* (no propongo alternativas hasta haber realizado el experimento)

Ejemplo de comparación *a priori*:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: $\mu_3 > \mu_1$; $\mu_3 > \mu_2$; $\mu_3 > \mu_4$

Los tests *a priori* son más potentes, pero están sometidos a mayor riesgo de error

Ej: Dunn Sidak

Tests *a posteriori*

- Comparan todos los posibles pares de medias entre sí, de tal modo que definen la alternativa a la H_0
- Únicamente podrá definirse una H_A sin ambigüedad en el caso de que se los distintos tratamientos se reúnan en grupos tales que:
    1. no haya diferencias entre las medias dentro de un grupo, y
    2. cada media en un grupo difiere de todas las medias del otro grupo

- Uno de los tests más utilizados es el de Student-Newman-Keuls (SNK)
- Otros tests utilizables:

Scheffe
Tukey
LSD
Bonferroni
… etc.

Tests *a posteriori*

Test SNK:

Ejemplo: 5 hábitats (a = 5)
7 réplicas por hábitat (n = 7)

Hábitat	**a**	**b**	**c**	**d**	**e**
Media	6,4	7,1	3,8	2,6	4,1
S^2	2,18	0,67	1,77	1,35	2,06

Fuente de var.	**SC**	**gl**	**MC**	**F**	**P**
Entre	99,26	4	24,82	15,38	***
Dentro	48,18	30	1,60		
Total	147,44	34			

Tests *a posteriori*

$ET = \sqrt{(MC_{DENTRO} / n)} = \sqrt{(1,60 / 7)} = 0,48$

Test SNK:

$Q_{ij} = X_i - X_j / ET$ está tabulado para H_0 verdadera

Rango	**1**	**2**	**3**	**4**	**5**		**test SNK**		
Medias	2,6	3,8	4,1	6,4	7,1	g	Q	D=QxET	
Comparaciones	5-1 4,5						5	4,10	1,97
	4-1 3,8	5-2 3,3					4	3,84	1,84
	3-1 1,5	4-2 2,6	5-3 3,0				3	3,49	1,68
	2-1 1,2	3-2 0,3	4-3 2,3	5-4 0,7			2	2,89	1,39

$\alpha = 0,05$
gl = 30

Tests *a posteriori*

$ET = \sqrt{(MC_{DENTRO} / n)} = \sqrt{(1,60 / 7)} = 0,48$

Test SNK:

$Q_{ij} = X_i - X_j$ / ET está tabulado para H_0 verdadera

Rango	**1**	**2**	**3**	**4**	**5**		**test SNK**		
Medias	2,6	3,8	4,1	6,4	7,1	g	Q	D=QxET	
Comparaciones	$^{5-1}$**4,5***						5	4,10	1,97
	$^{4-1}$**3,8***	$^{5-2}$**3,3***					4	3,84	1,84
	$^{3-1}$1,5	$^{4-2}$**2,6***	$^{5-3}$**3,0***				3	3,49	1,68
	$^{2-1}$1,2	$^{3-2}$0,3	$^{4-3}$**2,3***	$^{5-4}$0,7			2	2,89	1,39

Si $(X_i - X_j) > D$, la diferencia es significativa (*)

$\alpha = 0,05$
gl = 30

5 > 1 5 > 2 5 > 3 5 = 4
4 > 1 4 > 2 4 > 3
3 = 1 3 = 2
2 = 1

$\underline{5 = 4} > \underline{3 = 2 = 1}$

Hay 2 Subconjuntos Homogéneos

Tests *a posteriori*

Presentación de resultados

Test SNK:

Un resultado tal que:

5 4 3 2 1

¿Hay Subconjuntos Homogéneos?

… no resulta lógico (no hay una H_A identificable) !!

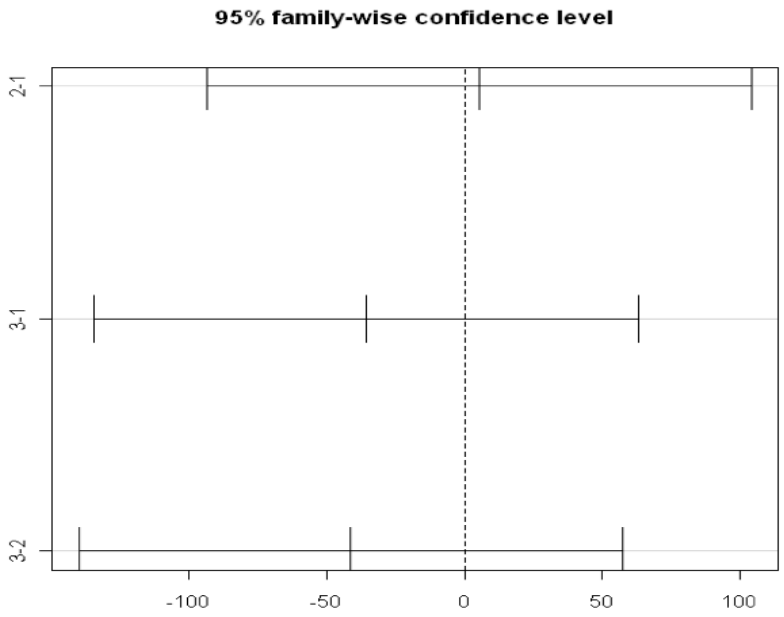
Tests *a posteriori*

Presentación de resultados

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = datos$ABUND ~ factor(datos$ZONA))

$`factor(datos$ZONA)`
           diff         lwr        upr     p adj
2-1    5.462963  -93.32526 104.25118 0.9906082
3-1 -35.685185 -134.47340  63.10303 0.6696206
3-2 -41.148148 -139.93637  57.64007 0.5871203
```



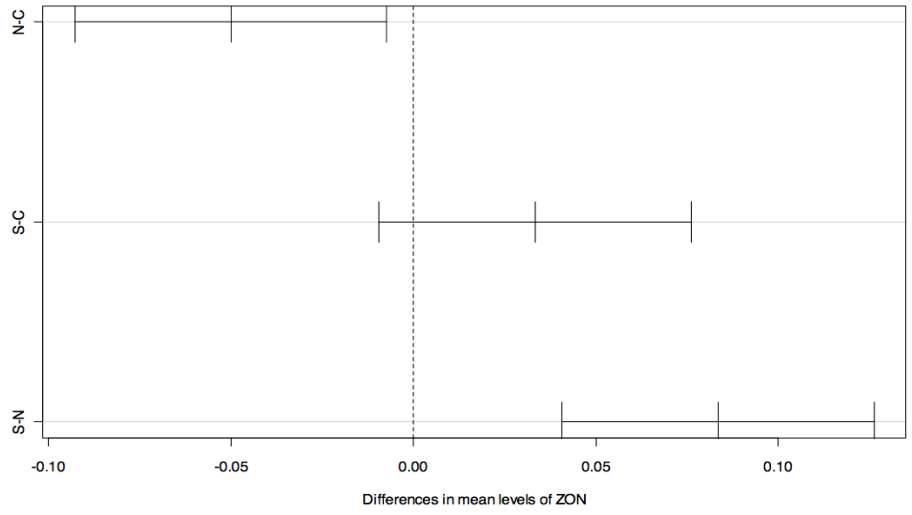
Presentación de resultados

FACTOR	Df	Sum Sq	Mean Sq	F (respecto del Residual)	F (correcto)	P-valor
Zona	2	0.12705	0.063526	11.5056	8.0627	$4.614 \cdot 10^{-5}$
Sitio	9	0.07091	0.007879	1.4269	1.3167	0.1929
Localidad	24	0.14361	0.005984	1.0837	1.0837	0.3832
Residuals	72	0.39754	0.005521			

Tabla 3. Resultados del Test de ANOVA.

			Intervalo de Confianza al 95%	
Factor	**Diferencias entre medias**	**Sig.**	**Límite inferior**	**Límite**
N-C	-0.05000828	0.0153165	-0.091921467	-0.008095102
S-C	0.03346132	0.1430502	-0.008451861	0.075374504
S-N	0.08346961	0.0000281	0.041556424	0.125382788

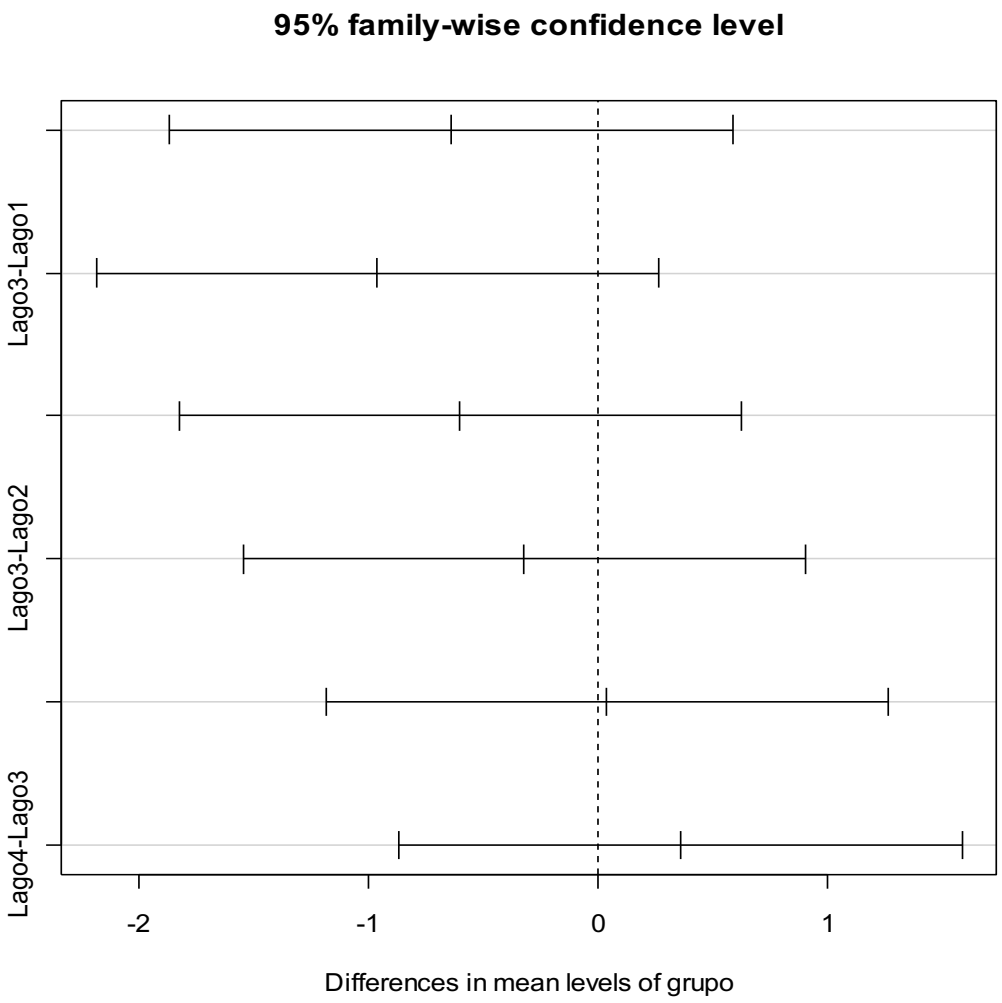
Tabla 4. Resultados del Test de Tukey.



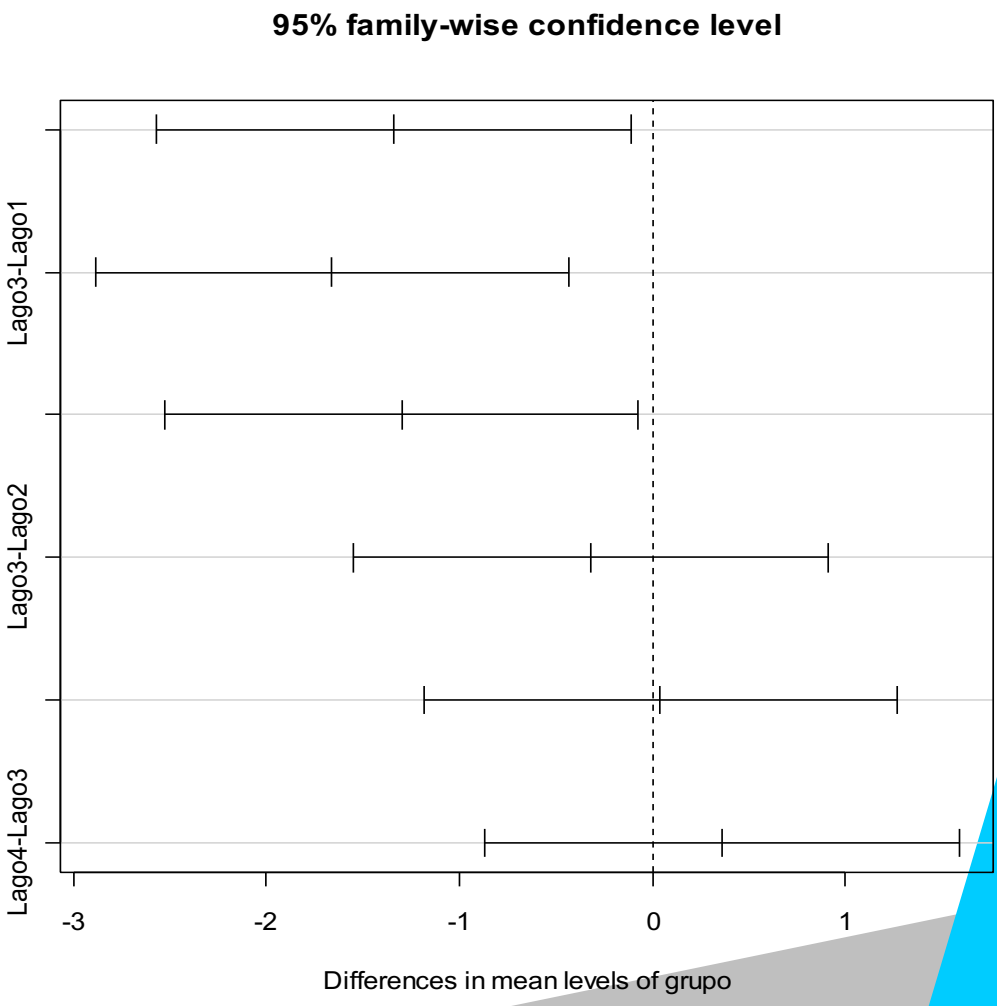
Presentación de resultados

Test a posteriori

H_0 no se rechaza
(Test no significativo)

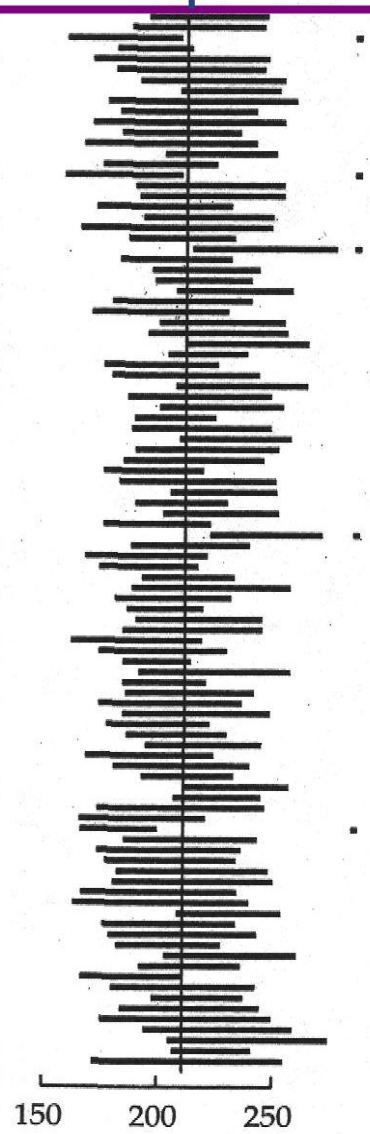


Se rechaza H_0
(Test significativo)



Presentación de resultados

Test a posteriori



Presentación de resultados

Test a posteriori

```
TukeyHSD(aov(PADINA~Zona*Profundidad))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = PADINA ~ Zona * Profundidad)

$Zona
        diff       lwr      upr     p adj
N-C -2.314815 -12.03046 7.400825 0.8395177
S-C -4.259259 -13.97490 5.456381 0.5544780
S-N -1.944444 -11.66008 7.771196 0.8838268

$Profundidad
        diff       lwr       upr p adj
2-1 -17.22222 -23.84418 -10.60026 8e-07

$`Zona:Profundidad`
             diff       lwr       upr     p adj
N:1-C:1  -6.666667 -23.42123 10.087898 0.8603058
S:1-C:1  -5.740741 -22.49530 11.013823 0.9209766
C:2-C:1 -21.111111 -37.86568 -4.356547 0.0049711
N:2-C:1 -19.074074 -35.82864 -2.319510 0.0156529
S:2-C:1 -23.888889 -40.64345 -7.134325 0.0008761
S:1-N:1   0.925926 -15.82864 17.680490 0.9999854
C:2-N:1 -14.444444 -31.19901  2.310120 0.1341259
N:2-N:1 -12.407407 -29.16197  4.347157 0.2740355
S:2-N:1 -17.222222 -33.97679 -0.467658 0.0400863
```