

DISEÑO DE EXPERIMENTOS EN BIOCIENCIAS (II)



ANÁLISIS ESTADÍSTICO Y SU DIAGNÓSTICO CON R

Antonio Monleón-Getino (coordinador y editor)

Section of Statistics. Group of Research BIOST³

<http://www.fbg.ub.edu/investigadors/grups-recerca-ub/estadistica-clinica-biodiversitat/>

Department of Genetics, Microbiology and Statistics
University of Barcelona

(14-09-2020)

(A mis hijas y a Pilar)

ISBN 978-0-244-15950-4



9 780244 159504

Sergi Gallardo Masip (2017): "dibujo daliniano con mensaje oculto"



Este libro es la segunda parte y se basa en parte en el libro “Diseño y planificación de estudios científicos: Calidad de datos (data management) y principios de diseño experimental)” (Monleón-Getino, 2017). Este manual que se presenta constituye un tratado introductorio al diseño de experimentos con ejemplos, propuestas y alternativas, basada en 15 años de docencia del diseño y análisis de experimentos, aportando la experiencia, ejemplos analizados por el autor y los apuntes de clase. Así como la recopilación de ejemplos resueltos propios, de otros autores y de material libre encontrado en internet. Se presentan ejemplos de diseño de experimentos y su resolución aplicados a las biociencias, la medicina y la bioinformática.

Lulu Press Inc. 21/02/2019 Creative commons license

www.lulu.com



(second version, “non corrected”)

Disclaimer:

This work is general work whose objective is to offer the lay public an updated vision on the data science. Part of the material that is exposed in it has been compiled from a revised bibliography (which has been cited) and from material found on the internet (which has been tried to cite in whole or in part) or in teaching material from other universities. If any material has not been properly quoted please excuse the inconvenience caused and please inform the author to do so in the required format.

Contenido

| | |
|---|-----|
| Contenido | 5 |
| Agradecimientos..... | 8 |
| Prólogo..... | 8 |
| 1-Fundamentos estadísticos del diseño de experimentos en las biociencias..... | 10 |
| 2-Diseño y análisis práctico de experimentos con R..... | 216 |
| 2.1-El método científico..... | 217 |
| 2.2-¿Por qué realizar experimentos?..... | 217 |
| 2.3-Los experimentos nos ayudan a contestar preguntas..... | 218 |
| 2.4-Estadística frecuentista y bayesiana..... | 219 |
| 2.6-Multiplicidad de pruebas estadísticas, riesgo alfa y potencia estadística ¿Cómo funciona?..... | 221 |
| 2.6.1- Un poco de teoría..... | 221 |
| 2.6.2- Estudios multiobjetivo (Y), significación o potencia estadística. Un ejemplo práctico. ¿Cómo establecer la regla de decisión?..... | 240 |
| 2.7. Potencia estadística y cálculo del tamaño muestral en diseño de experimentos | 243 |
| 2.7.1. ¿Cómo se calcula el tamaño muestral de un estudio? Ideas generales..... | 243 |
| 2.7.2.Cálculo del tamaño muestral en pruebas T-test para comparar medias de dos poblaciones..... | 248 |
| 2.7.3 Ejemplo 1 cálculo del tamaño muestral. ¿Es eficaz un programa de pérdida de peso? | 250 |
| 2.7.4. Ejemplo 2 cálculo del tamaño muestral en un diseño experimental: destreza manual..... | 253 |
| 2.7.5. Cálculo del tamaño muestral en el caso de la comparación de dos proporciones | 256 |
| 2.7.6.Cálculo del tamaño muestral en diseños ANOVA (“One way”) mediante el uso de la función <code>pwr.anova.test()</code> | 258 |
| 2.7.7. Cálculo del tamaño muestral en un diseño experimental ANOVA de varios factores mediante simulación..... | 260 |
| 2.8 Introducción al análisis de la varianza (ANOVA) unifactorial “One-way” | 263 |
| 2.8.1 Concepto y suposiciones..... | 263 |
| 2.8.2 Ejemplo: Ensayo clínico de un fármaco..... | 265 |

| | |
|--|-----|
| 2.8.3 Mejora de la producción de calabazas con diferentes fertilizantes. Cálculo manual de un ANOVA | 293 |
| 2.8.4 Diversidad biológica y contaminación de zinc en las Montañas Rocosas (USA) | 301 |
| 2.8.5 Mejora de la resistencia de los tejidos de algodón mediante diseño experimental | 304 |
| 2.8.6 Ejercicio de simulación de una distribución F en una prueba ANOVA..... | 320 |
| 2.8.7. Un factor aleatorio, ejemplo del análisis del efecto lote en el rendimiento de colorante (batch)..... | 324 |
| 2.8.8. Ejemplo del uso de ANOVA no paramétrico..... | 334 |
| 2.9.ANOVA de 2 factores: introducción y asunciones | 347 |
| 2.9.1. Presión ecológica sobre <i>Patella vulgata</i> (lapa de roca)..... | 347 |
| 2.9.2. Ejemplo 2: Influencia de la dieta y del género en la pérdida de peso | 357 |
| 2.9.3. Cebos para polillas..... | 363 |
| 2.9.5. Dos factores aleatorios: ejemplo de la eficacia de la penicilina como inhibidor del crecimiento de <i>Bacillus subtilis</i> | 368 |
| 2.9.6 ANOVA multifactorial robusto: test de permutaciones librería RVAideMemoire, uso en múltiples factores..... | 375 |
| 2.9.7. ANOVA multifactorial robusto: test de permutaciones función adonis() librería Vegan..... | 404 |
| 2.9.10-Anova basado en Bootstrap (Aova multifactorial con diseños no balanceados) | 406 |
| 2.10.-Diseños con bloques aleatorizados..... | 412 |
| 2.10.1. Palillos en la alimentación | 413 |
| 2.10.2. Diseño de probióticos en biotecnología..... | 415 |
| 2.10.3. Uso de cuadrados latinos en diseños de producción agrícola | 417 |
| 2.11.Diseños jerárquicos o anidados con factores fijos y aleatorios | 422 |
| 2.11.1 Proveedores de nitrógeno y lotes (diseño mixto, con factores fijos y aleatorios)..... | 423 |
| 2.11.2 Medida de la salinidad en pozos de diversas zonas agrícolas..... | 428 |
| 2.11.3: Heredabilidad de las cualidades anaerobias (3 factores anidados)..... | 438 |
| 2.12.1. Diseños de experimentos longitudinales (medidas repetidas)..... | 447 |
| 2.12.2 Ejemplo simple de medidas repetidas | 448 |
| 2.12.3 Crecimiento de los dientes en niños y niñas..... | 450 |
| 3.PROBLEMAS DE CLASE PROPUESTOS/RESUELTOS | 457 |
| 3.1 Problema: Variabilidad de los operarios en la fábrica de plásticos..... | 458 |

| | |
|--|-----|
| 3.2. Influencia del genotipo en la actividad del enzima Manosa-6-fosfat isomerasa (MPI) en los crustáceos anfípodos <i>Platorchestia platensis</i> | 471 |
| 3.3. Influencia de las citosinas en la respuesta inmunitaria de los animales de granja. | 474 |
| 3.4. Estudio in vivo sobre el efecto de las cicinas en la expansión de las células progenitoras del CD34+ | 483 |
| 3.5. Fiabilidad de la medida: muestra y técnico de laboratorio. | 493 |
| 3.6. Influencia del hexobarbital en enzimas metabolizadores y efecto del sexo de los animales | 502 |
| 3.7. Oxígeno disuelto en el agua (ODA) y la turbulencia en el agua de los ríos..... | 511 |
| 3.8. Microbiología de los quesos..... | 515 |
| 3.8. Fertilizantes orgánicos y nitrógeno..... | 520 |
| 4. Cuestiones conceptuales (tipo test)..... | 528 |
| 4.1. Primer test..... | 529 |
| 4.2 Test 2..... | 531 |
| 4.2. Segundo test..... | 533 |
| 5. Bibliografía..... | 536 |
| 5. Bibliografía..... | 537 |



Agradecimientos

Mi más sincero agradecimiento al Dr Jordi Ocaña Rebull, director de mi tesis doctoral y mi maestro, que me ha permitido utilizar el material creado por el para las clases de diseño de experimentos y análisis de datos en diferentes grados de la Facultad de Biología de la Universidad de Barcelona. El Dr Jordi Ocaña es el autor de algunas de las funciones que aquí se utilizan tales como `source("anova2cross.R")`, `source("anova2net.R")` y otras parecidas, que permiten calcular correctamente los estadísticos F de las tablas anova en el caso de factores aleatorios.

A otros profesores de la Facultad de Biología por su apoyo, a mis compañeros y a las empresas que he asesorado y que han utilizado el material del libro para mejorar, entender sus experimentos y sobre todo obtener evidencia científica. También a otros profesores como Hafid Laayouni de la UPF, Francesc Carmona y Miquel Salicrú de los que me he inspirado en diversos de los ejercicios que presento. Intento citar siempre la autoría de los mismos.

A Sergi Gallardo Masip, joven dibujante y artista prometedor, por su dibujo de la portada.

A Martín Ríos y a Santiago Ríos por la revisión del texto.

Prólogo

Este libro es la segunda parte y se basa en parte en el libro “Diseño y planificación de estudios científicos: Calidad de datos (data management) y principios de diseño experimental” (Monleón-Getino, 2017), donde se exponen los principios del diseño de experimentos y su análisis en el campo de la biología y la medicina, sin entrar en la base matemática-estadística del diseño de experimentos y sin desarrollar concretamente ningún ejemplo. Es por ello que este manual que se presenta constituye un manual concreto de diseño de experimentos con ejemplos, propuestas y con la parte teórica basada en 15 años de docencia del diseño y análisis de experimento, aportando la experiencia, ejemplos analizados por el autor y los apuntes de clase.

Se han tomado ejemplos propios, así como otros ejemplos utilizados en clase durante la docencia de la asignaturas “Diseño de experimentos y análisis de datos” por el profesor Jordi Ocaña Rebull, catedrático del Departamento de Genética, Microbiología y Estadística. Otros ejemplos han sido tomados de otras asignaturas en las que el autor imparte clases, donde se menciona su uso.

Otra buena introducción puede encontrarse en: [A First Course in Design and Analysis of Experiments](#)

Este manual, eminentemente práctico, está destinado a los estudiantes e investigadores de las biociencias: biólogos, médicos, biotecnólogos, bioinformáticos ambientólogos y todos aquellos que realicen diseño de experimentos y su análisis estadístico; con alternativas al clásico análisis ApNOVA y su diagnóstico. Les ayudará a planificar sus diseños, así como a entender cómo analizarlos en R y a entender las dificultades que pueden encontrarse, y a solucionarlas con alternativas y casos prácticos.

El principio de este manual es una introducción teórico-práctica del fundamento estadístico del diseño de experimentos, basado fundamentalmente en las pruebas de hipótesis y el análisis de la varianza (ANOVA)

Los capítulos posteriores contienen problemas reales, su análisis y diagnóstico, así como interesantes alternativas modernas. Al final se presentan diferentes problemas solucionados en clase de diseño de experimentos.

Se presentan ejemplos y escripts realizados con el lenguaje R (<https://cran.r-project.org/>)

1-Fundamentos estadísticos del diseño de experimentos en las biociencias

ANOVA INTRODUCCION Y TEORIA

Toni Monleón-Getino

Junio de 2020

Secció d'Estadística
Facultat de Biologia



UNIVERSITAT DE
BARCELONA

Nota sobre el material

Procedencia de este material:

- ▶ Basado en material docente de las clases de Diseño de experimentos y análisis de datos de la Facultat de Biologia (Universitat de Barcelona)
- ▶ Basado en material procedente del Dr Jordi Ocaña Rebull y otros profesores.
- ▶ Basado en material libre de internet.
- ▶ Ver introducción en: <https://tutor-web.net/stats/546>
- ▶ Transformado y adaptado parcialmente de:
<http://staff.pubhealth.ku.dk/tag/Teaching/BasicStats/homepage/Lecturenotes.html>

CONTENIDO

INTRODUCCION

- Causas de variación en los datos
- Incertidumbre estadística

ALGUNOS CONCEPTOS ESTADÍSTICOS

- Test de hipótesis, error tipo I y II, p-valor
- Intervalos de confianza

ANOVA: INTRODUCCION

ONE-WAY ANOVA

- Modelo lineal de un factor fijo
- Inferencia en el caso normal
- Verificación de la validez del modelo
- Comparaciones múltiples
- Modelo con un factor aleatorio

ANOVA N-WAY

- Diseño de dos factores cruzados
- Diseños con bloques factoriales
- El modelo con bloques y un factor fijo
- Diseños multifactoriales con uno o más factores aleatorios

- Diseños factoriales con uno o más factores aleatorios
- Diseños mixtos: mezcla de factores fijos y aleatorios

DISEÑOS CON FACTORES JERÁRQUICOS

- Concepto de factor jerárquico
- Tipos de diseños
- Model lineal

ANOVA BAYESIANO

- Ventaja de los métodos bayesianos y su popularidad
- Bases matemáticas de los métodos bayesianos
- Ejemplo de ANOVA bayesiano

Otros temas de ANOVA

INTRODUCCION

Toni Monleón-Getino

Junio de 2020

Los datos son variables

Un estadístico es una variable que se utiliza para recibir un valor, tal como pasa con:

4.12 %,

Junto con una explicación como por ejemplo:

"Este es el porcentaje de presencia de tal organismo en el microbiota."

Descomponiendo la varianza

La variabilidad de los datos biológicos, está constituida por una mezcla de:

- ▶ Error en la medida o en el muestreo (Measurement error, sampling scheme)
- ▶ Variabilidad aleatoria (Random variation)
- ▶ Genotipo (Genotype)
- ▶ Exposición, estilo de vida, entorno (Exposure, life style, environment)
- ▶ Tratamiento al que ha sido sometido (Treatment)

Se pueden obtener conclusiones estadísticas si se pueden explicar las causas de la variabilidad en los datos.

Estadística descriptiva (resumiendo la información)

Dependiendo de la naturaleza de las variables, pueden resumirse de diferente manera:

Variables categóricas:

- ▶ count (%)

Variables continuas:

- ▶ raw values (si n es pequeño)
- ▶ range (min, max)
- ▶ location: median (IQR=inter quartile range)
- ▶ location: means (SD)

¿Cómo cuantificar la variabilidad?

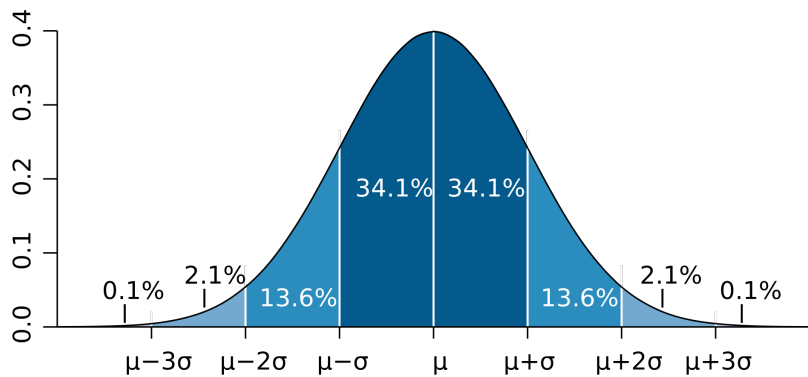
Una muestra de datos X_1, \dots, X_N tiene una **desviación estándar (sd)**; que se define como

$$SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2}; \quad \bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

SD mide la variabilidad de las medidas en una muestra.

La **varianza** de una muestra se define como SD^2 . El término 'standard deviation' (desviación estándar) se refiere a la distribución normal.

Distribución normal



¿Por qué es tan especial la distribución normal?

- ▶ Es simétrica alrededor de la media, así la media es igual a la mediana.
 - ▶ La media es el valor más probable. La media y la desviación estándar describen la distribución completa.
 - ▶ La distribución de medidas, como altura, distancia, volumen es a menudo normal.
-
- ▶ La distribución de los estadísticos, como la media, la proporción, la diferencia de medias, etc., a menudo es aproximadamente normal.
-

Cuantificación de la incertidumbre estadística

- ▶ Para la inferencia estadística y la realización de conclusiones, a través de valores p e intervalos de confianza, es crucial cuantificar la variabilidad de los estadísticos (media, proporción, diferencia de medias, razón de riesgo, etc.):
- ▶ El error estándar es la desviación estándar de la estadística.
- ▶ El error estándar es una medida de la incertidumbre estadística.

Cuantificando la incertidumbre estadística

Example: Queremos estimar el peso medio desconocido de un tipo de órgano en ratones de tipo "Control". El **error estándar** de la media se define como

$$SE = SD/\sqrt{N} \quad \text{donde } N \text{ es el tamaño de la muestra.}$$

Basado en $N = 4$ valores, 0.012, 0.0088, 0.0069, 0.009:

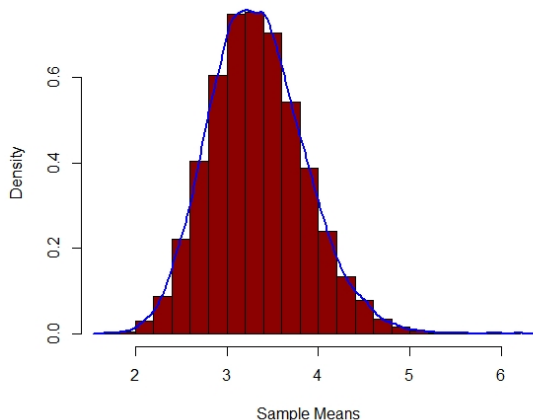
- ▶ mean: $\hat{\beta} = 0.0091$
- ▶ standard deviation: $SD = 0.002108$
- ▶ empirical variance: $var = 0.0000044$
- ▶ standard error: $SE = 0.002108/2 = 0.001054$

Cuantificando la incertidumbre estadística

```
#Simulate a distribution of the sample means.
sample_means = rep(NA, 30000)
for(i in 1:30000){
  sample_means[i] = mean(rexp(42,0.3))
}
hist(sample_means, main = "", xlab = "Sample Means"
, prob = T, col = "darkred")
lines(density(sample_means,na.rm = T), col = "blue"
, lwd = 2)
```

```
> sample_means
[1] 3.106190 3.100709 3.042631 3.128561 3.359215 3.249579
[11] 4.372723 3.313985 3.026776 3.130826 3.184767 3.752995
[21] 3.543038 3.531453 3.063798 3.493093 3.733996 3.224339
[31] 3.407127 3.060985 3.079313 1.904697 3.037635 3.013649
[41] 3.106687 4.196079 3.373259 2.483192 3.052203 3.401703
```

El error estándar es la desviación estándar de la media.



Los valores medios (hipotéticos) se distribuyen aproximadamente de manera normal, ¡incluso si los datos no se distribuyen normalmente!

Varianza vs incertidumbre estadística

Standard error y *standard deviation*, a menudo se confunden. La diferencia entre estos dos términos refleja la importante distinción entre descripción de datos e inferencia, algo que todos los investigadores deberían apreciar.

Algunas normas a tener en cuenta:

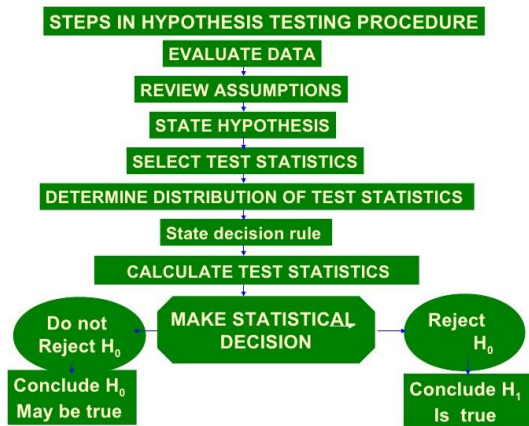
- ▶ Cuanto mayor sea la variabilidad inexplicada de los datos, mayor será la incertidumbre estadística.
- ▶ Cuanto mayor es el tamaño de la muestra, menor es la incertidumbre estadística.

ALGUNOS CONCEPTOS ESTADÍSTICOS

Toni Monleón-Getino

Junio de 2020

Test de hipótesis en estadística.



Errores en la toma de decisiones estadística

Tipos de error

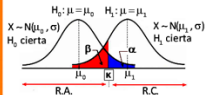
- Error tipo I: ocurre cuando la hipótesis nula es rechazada cuando en realidad es verdadera (riesgo del productor α)
- Error tipo II: ocurre cuando la hipótesis nula es aceptada cuando en realidad debió ser rechazada (riesgo del consumidor β)

| | | Hipótesis nula | |
|--------------------|-------------------|-----------------------------------|----------------------------------|
| | | Verdadero | Falso |
| La decisión tomada | No rechazar H_0 | $p=1-\alpha$ Decisión correcta | $p=\beta$ Error tipo II |
| | Rechazar H_0 | $p=\alpha$ Error tipo I | $p=1-\beta$ Decisión correcta |

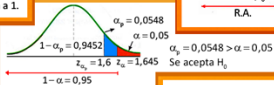
p-valor

Cuanto más lejos se encuentra H_2 de H_1 , menor es la probabilidad de cometer un Error Tipo II (β) y, por tanto, la Potencia se aproxima a 1.

$\alpha = P[\text{Error tipo I}]$
 $\beta = P[\text{Error tipo II}]$
Potencia = $1 - \beta$



$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ es cierta}]$
 $\beta = P[\text{Aceptar } H_0 \mid H_0 \text{ es falsa}]$



Potencia = $1 - \beta = P[\text{Aceptar } H_1 \mid H_1 \text{ es cierta}]$



La elección del nivel de significación α (riesgo) es en cierta forma arbitrario. El p-valor es una cantidad que resume de manera objetiva el resultado del experimento. El p-valor = α_p es el riesgo más pequeño con el que se acepta la hipótesis alternativa H_1 en las observaciones muestrales. La decisión a tomar con $\alpha_p = p$ -valor que se obtiene de la muestra:

- Se acepta la hipótesis nula H_0 : p-valor = $\alpha_p > \alpha$
- Se rechaza la hipótesis nula H_0 : p-valor = $\alpha_p \leq \alpha$

$\alpha_p = p\text{-valor} = P[\text{Rechazar estadístico muestra} \mid H_0 \text{ es cierta}]$

INTERVALOS DE CONFIANZA

Toni Monleón-Getino

Junio de 2020

Intervalos de confianza

Un **intervalo de confianza** es un rango de valores que cubre el parámetro de una población desconocida con alta probabilidad. Aproximadamente la la probabilidad es $100 - \alpha\%$ donde α es el nivel de significación.

Por ejemplo:

5 to 25

es el 95% de confianza para un intervalo para la diferencia promedio desconocida en peso entre hombres y mujeres.

Los intervalos de confianza tienen la ventaja sobre los valores p, ya que su valor absoluto tiene una interpretación directa.¹

¹Confidence intervals rather than P values: estimation rather than hypothesis testing. Statistics with Confidence, Altman et al.

Relación entre los p-valores y los intervalos de confianza

Si se estima un parámetro poblacional por ejemplo β ,

$$\beta = \text{mean}(\text{peso hombres}) - \text{mean}(\text{peso mujeres})$$

y se calcula el 95% del intervalo de confianza para ese parámetro,

$$[\text{inferior}_{95}, \text{superior}_{95}]$$

entonces la hipótesis nula:

$$\beta = 0 \quad \text{"No hay diferencias"}$$

podemos rechazarla con un 5% de nivel de significación, si el valor 0 **no** está incluida en el intervalo: $0 \notin [\text{inferior}_{95}, \text{superior}_{95}]$.

Construyendo intervalos de confianza

A **Intervalo de confianza del 95%** para el parámetro μ es [En general μ representará la media poblacional]:

$$[\hat{\mu} - 1.96 * SE; \hat{\mu} + 1.96 * SE]$$

Ejemplo: un intervalo de confianza para la media de un determinado órgano de un ratón del grupo "Control" está dado por:

$$\begin{aligned} 95\%CI &= [0.0091 - 1.96 * 0.001054; 0.0091 + 1.96 * 0.001054] \\ &= [0.007; 0.011]. \end{aligned}$$

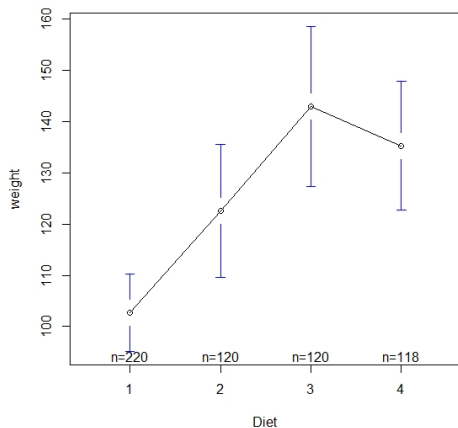
El error estándar (standard error) SE mide la variabilidad de la media $\hat{\mu}$ alrededor del valor (desconocido) de la población β , bajo la asunción de que el modelo está correctamente especificado.

Construyendo intervalos de confianza

```
#Calculation and representation of CI at 95% for a
mean (Example: chickens' weight with different
diets).
library(datasets)
data(ChickWeight)
data(state)
#Graphic representation for each diet.
plotmeans(weight ~ Diet,data =ChickWeight )
#Calculation of CI at 95% for the mean of all diets
.
library(DescTools)
MeanCI(ChickWeight$weight,na.rm = T)
```

```
> MeanCI(ChickWeight$weight,na.rm = T)
mean   lwr.ci   upr.ci
121.8183 116.0121 127.6246
```

Intervalo de confianza del 95%



Esperamos, al menos, que 5 de cada 100 intervalos de confianza no incluyan el verdadero valor del parámetro. Ha de mirarse en este caso para cada dieta.

Intervalo de confianza para el peso de unos pollos sometidos a dieta

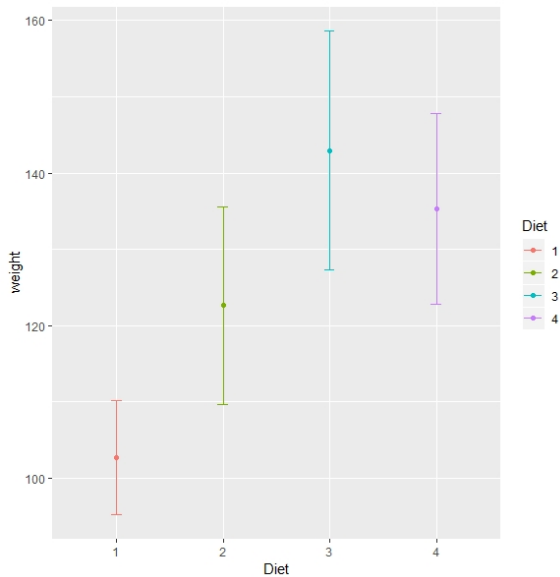
```
library(datasets)
data(ChickWeight)
library(Rmisc)
res.RMISC1 <- summarySE(ChickWeight, measurevar="
weight", groupvars=c("Diet"),na.rm = T)
res.RMISC1
```

| | Diet | N | weight | sd | se | ci |
|---|------|-----|----------|----------|----------|-----------|
| 1 | 1 | 220 | 102.6455 | 56.65655 | 3.819784 | 7.528242 |
| 2 | 2 | 120 | 122.6167 | 71.60749 | 6.536840 | 12.943596 |
| 3 | 3 | 120 | 142.9500 | 86.54176 | 7.900146 | 15.643078 |
| 4 | 4 | 118 | 135.2627 | 68.82871 | 6.336197 | 12.548506 |

Representación de los intervalos de confianza

```
#R-code
pd <- position_dodge(0.1) #move them 0.05 to the
left and right
gi <- ggplot(res.RMISC1, aes(x = Diet, y = weight,
colour = Diet, group=Diet))
#add x and y labs
gi <- gi + xlab("Diet") + ylab("weight")
#add the summary statistic to plot: mean. Draw a
point and join it by a line
gi <- gi + stat_summary(fun.y=mean, geom="point")+
  stat_summary(fun.y=mean, geom="line") +
  geom_errorbar(aes(ymin=weight-ci, ymax=weight+ci)
, width=.1, position=pd)
#plot
plot(gi)
```

Representación de los intervalos de confianza



Límites para el intervalo de confianza de la media geométrica

```
#Geometric mean
#reference: https://rdrr.io/cran/survJamda/man/ci.gm.html
v = c(1.5,2.5,7,4) #example
ci.gm(v) #calculation

#The function is currently defined as:
ci.gm<- function(x){
  gm1 = mean(log(x), na.rm = T)
  cil = exp(gm1-(1.96*(sd(log(x), na.rm = T)/sqrt(
    length(x))))))
  ciupp = exp(gm1+(1.96*(sd(log(x), na.rm = T)/sqrt(
    length(x))))))
  vec = c(exp(gm1), round(cil,2), round(ciupp,2))
  return (vec)
}
```

```
> ci.gm(v) #calculo
[1] 3.201086 1.680000 6.100000
```

Parámetros

En general, es difícil interpretar un valor p sin una cuantificación adicional del parámetro de interés (ej: media).

Los parámetros son características interpretables de una población que deben estimarse en función de los datos.

Algunos ejemplos de parámetros son:

- ▶ Medias
- ▶ Diferencia de medias
- ▶ Probabilidades
- ▶ Risk ratios, odds ratios, hazard ratios
- ▶ Parámetros de asociación, coeficientes de regresión

ANOVA: Introducción

Toni Monleón-Getino

Junio de 2020

Análisis de la varianza (ANOVA)

- ▶ El ANOVA es una técnica estadística paramétrica utilizada para comparar conjuntos de datos.
- ▶ Esta técnica la inventó R.A. Fisher (Estadístico y biólogo, 1890 – 1962), y es en general conocido como ANOVA de Fisher.
- ▶ RA Fisher en 1919 comomenzó a trabajar en Rothamsted Research, una estación agrícola experimental donde desarrolló el ANOVA para analizar la inmensa cantidad de datos de los cultivos, disponibles desde 1840, y donde en los siguientes años se ganó una gran reputación como bioestadístico.



Análisis de la varianza (ANOVA)

- ▶ Esta técnica es similar a técnicas como la prueba T y la prueba Z, ya que se utiliza para comparar medias y su variación.
- ▶ Sin embargo, el análisis de la varianza (ANOVA) se aplica mejor donde se quiere comparar más de 2 poblaciones o muestras.
- ▶ Solemos ligar esta técnica al concepto de modelo estadístico.
- ▶ Cada diseño experimental está ligado a un modelo estadístico

Diseño de experimentos y ANOVA

- ▶ El diseño de experimentos es una disciplina propia en la ciencia experimental, con el que se intenta disminuir la incertidumbre de los factores que intervienen en una respuesta.
- ▶ La planificación de un experimento para obtener datos apropiados y sacar inferencia de los datos con respecto a cualquier problema bajo investigación se conoce como diseño y análisis de experimentos.
- ▶ Esto puede variar desde las formulaciones de los objetivos del experimento en términos claros hasta la etapa final de la redacción de los informes que incorporan los hallazgos importantes de la investigación.
- ▶ La estructuración de las variables dependientes e independientes, la elección de sus niveles en el experimento, el tipo de material experimental que se utilizará, el método de manipulación de las variables en el material experimental, el método de registro y tabulación de datos, el modo de análisis del material y la inferencia válida, etc., son todos detalles intermedios que van con el diseño y el análisis de un experimento.
- ▶ **Diseño:** Los experimentos suelen tener asociado un modelo aleatorio, compuesto por una parte determinista y una parte aleatoria (error).
- ▶ Es frecuente utilizar en diseño experimental modelos lineales y especialmente ANOVA para la estimación de los parámetros de los mismos y el estudio de su validez.

Ejemplo de experimento: Pollos sometidos a dieta

```
library(datasets)
data(ChickWeight)
```

ChickWeight es un data frame con 578 filas y 4 columns de un experimento del efecto de una dieta en el crecimiento de unos pollos.

Outcome o respuesta (Y): Peso de los pollos (gr)

| Grupo | Peso | N | Mean | Sd |
|-------|---------|---|-------|------|
| I | Dieta 1 | 8 | 316.6 | 58.7 |
| II | Dieta 2 | 9 | 256.4 | 37.1 |
| III | Dieta 3 | 5 | 278.0 | 33.8 |

ANOVA

```
#R-code for ANOVA  
anova(lm(weight~Diet,data=ChickWeight))
```

```
> anova(lm(weightDiet,data=ChickWeight))
```

```
Analysis of Variance Table
```

```
Response: weight
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|-----|---------|---------|---------|---------------|
| Diet | 3 | 155863 | 51954 | 10.81 | 6.433e-07 *** |
| Residuals | 574 | 2758693 | 4806 | | |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Tabla ANOVA para el peso de los pollos

| Source of variation | Degrees of freedom | Sum of squares | Mean squares | <i>F</i> | <i>P</i> |
|---------------------|--------------------|----------------|--------------|----------|----------|
| Between groups | 3 | 155863 | 51954 | 10.81 | 6.4e-7 |
| Within groups | 574 | 2758693 | 4806 | | |
| Total | 577 | 2914556 | | | |

¿Qué representan la suma de cuadrados y los grados de libertad?

Recordad la definición de varianza para una muestra de N valores X_1, \dots, X_N con media $= \bar{X}$:

$$Var = \frac{1}{N-1} \{(X_1 - \bar{X})^2 + \dots + (X_N - \bar{X})^2\}$$

$$Var = \frac{1}{\underbrace{N-1}_{\text{degrees of freedom}}} \underbrace{\{(X_1 - \bar{X})^2 + \dots + (X_N - \bar{X})^2\}}_{\text{Sum of squares}}$$

En la terminología ANOVA, la varianza se conoce como un cuadrado medio que es la abreviatura de: desviación cuadrática media de la media.

Métodos ANOVA

Existen muchos experimentos y análisis que pueden solucionarse mediante ANOVA, como:

- ▶ Observaciones independientes:
 - ▶ T-test para dos grupos
 - ▶ One-way ANOVA para varios grupos
 - ▶ More-way ANOVA para varios grupos variables
- ▶ Observaciones dependientes:
 - ▶ Diseño de medidas repetidas
 - ▶ Modelos de efectos mixtos
- ▶ Estadísticos de rango (tests ANOVA no paramétricos)
 - ▶ Anova no paramétrico (test Kruskal-Wallis)
- ▶ Combinación de factores discretos y continuos:
 - ▶ Ancova
- ▶ Modelos de comparación y modelos de selección ...

Hipótesis típicas para los F-test

| | | |
|-------|----------------------------|---|
| H_0 | Hipótesis nula | <i>El peso de los pollos no depende del tipo de tratamiento (dieta)</i> |
| H_1 | Hipótesis alter- nativa | <i>El peso de los pollos depende del tipo de tratamiento</i> |

Esto significa que:

H_0 : Media grupo I = Media grupo II = Media grupo III

H_1 : Media grupo I \neq Media grupo II

o Media grupo III \neq Media grupo II

o Media grupo I \neq Media grupo III

Normalmente deseamos conocer **qué** tratamiento presenta **mejor** respuesta.

El estadístico F (F-test)

Idea central: la desviación de la respuesta de un sujeto de la gran media de todas las respuestas es atribuible a una desviación de ese valor de su media grupal más la desviación de esa media grupal de la gran media.

$$F = \frac{\text{Variabilidad inter-grupo (between-group)}}{\text{Variabilidad intra-grupo (within-group)}}$$
$$= \frac{\text{Varianza de los valores medios de respuesta entre grupos}}{\text{Variación de los valores dentro de los grupos}}$$

Si la variabilidad entre grupos es grande en relación con la variabilidad dentro del grupo, entonces el factor de agrupación contribuye a la parte sistemática de la variabilidad de los valores de respuesta.

Conclusiones de la tabla ANOVA

| Source of variation | Degrees of freedom | Sum of squares | Mean squares | F | P |
|---------------------|--------------------|----------------|--------------|-------|---------|
| Between groups | 3 | 155863 | 51954 | 10.81 | 6.4e-07 |
| Within groups | 574 | 2758693 | 4806 | | |
| Total | 577 | 2914556 | | | |

Conclusión: El peso de los pollos depende **significativamente** del tipo de dieta al que han sido sometidos (tratamiento).

- ▶ La variación de datos se puede descomponer en una parte sistemática y una aleatoria.
- ▶ La desviación estándar cuantifica la variabilidad de los datos..
- ▶ El error estándar cuantifica la incertidumbre de las conclusiones estadísticas..
- ▶ ANOVA es una técnica estadística antigua y general con muchas aplicaciones diferentes.

ONE-WAY ANOVA

Toni Monleón-Getino

Junio de 2020

¿Qué situación estudiamos?

- ▶ Queremos estudiar si un único factor con k niveles influye en una determinada variable de respuesta:
 - ▶ p.e. ¿La variedad del tomate (A, B, C, D) influye en la producción?
 - ▶ p.e. ¿El crecimiento micelial en términos de diámetro de la colonia (mm) de cultivos aislados de *Rizoctonia solani* después de 14 horas de incubación se ve afectado por unos tratamientos inhibidores?
- ▶ Más precisamente: ¿La media poblacional de la variable de respuesta es diferente según el nivel del factor?

Análisis de la varianza (ANOVA)

- ▶ **RECUERDA:** Un factor es una variable **cualitativa**, que está compuesta de uno o varios niveles y que permite diferenciar un grupo de otro grupo.
- ▶ Por ejemplo son factores: tratamiento, lugar de muestreo, etc.
- ▶ Ejemplo: Cuatro niveles de dosis de nemátodos en un experimento en un experimento en planteles.

Ejemplo de comparación de medias entre grupos

¿Las diferencias entre medias son cercanas a las que esperaríamos por azar en un muestreo aleatorio?

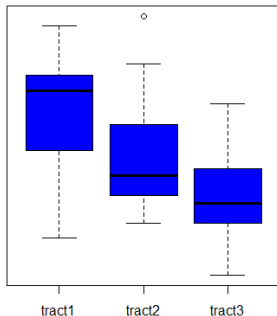
```
#Simulation of one factor with 3 levels
```

```
a1<-rnorm(20, 500,400); a2<-rnorm(20, 250,400)
```

```
a3<-rnorm(20, 100,400);a<-as.data.frame(c(a1,a2,a3))
```

```
a$group<-rep(c("tract1","tract2","tract3"),c(20,20,20))
```

```
colnames(a)<- c("Response","Factor"); boxplot(a$Response~a$Factor, yaxt='n', ann=FALSE, col="blue")
```



Diseño paralelo completamente aleatorizado

El experimento

Cada unidad experimental se asigna, al azar, a exactamente una condición experimental o tratamiento, es decir, a cada nivel del factor.

- ▶ Como consecuencia, el *factor* individuo está jerárquicamente dentro de cada grupo de tratamiento
- ▶ Este diseño (muy ampliamente utilizada) recibe varios nombres:
 - ▶ diseño de un solo factor (*one-way layout*)
 - ▶ diseño paralelo para un solo factor
 - ▶ Diseño de un factor totalmente aleatorizado (CRD)

Completely Randomized Design (CRD)

Ventajas

- ▶ Simplicidad
- ▶ Optimización del tiempo de duración del experimento
- ▶ Mínimo peligro de pérdida de casos

Inconvenientes

- ▶ Mayor número de unidades experimentales (que a otros diseños)
- ▶ Comparación *entre* sujetos, no *dentro* de los sujetos
- ▶ Movilización de muchos recursos en poco tiempo

Modelo lineal asociado al CRD

- ▶ Hay dos fuentes de variación entre las N observaciones obtenidas en un CRD:
 - ▶ 1) la variación debida a los tratamientos (niveles del factor)
 - ▶ 2) el error experimental

El tamaño relativo de (1) y (2) se utiliza para indicar si la diferencia observada entre los tratamientos es real o se debe al azar. Se dice que la diferencia de tratamiento es real si la variación del tratamiento es bastante mayor que el error experimental. Un modelo estadístico es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales (Ex: ANOVA, regresión) para indicar los diferentes factores que modifican la variable de respuesta.

Definición del modelo lineal asociado

$$\text{Respuesta}(Y) = \text{Causado por el factor}(F) + \text{error}$$

Modelo lineal asociado al CRD

El modelo estadístico más simple es el usado en los diseños completos aleatorizados (DCA), también denominados en inglés CRD.

Definición

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

donde:

Y_{ij} es la observación j por el nivel i del factor

μ_i es la media en cada población o nivel del factor

ϵ_{ij} es el error aleatorio correspondiente a la observación i, j

Supongamos que los errores ϵ_{ij} són i.i.d con $E(\epsilon_{ij}) = 0$ y $\text{var}(\epsilon_{ij}) = \sigma^2$.

Modelo lineal asociado al CRD II

Definición alternativa

Este modelo también puede ser escrito como:

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

donde:

Y_{ij} es la observación j por el nivel i del factor

μ es la media general

α_i es un parámetro que mide el efecto del nivel i del factor

ϵ_{ij} es el error aleatorio correspondiente a la observación i, j

Supongamos que los errores ϵ_{ij} son i.i.d con $E(\epsilon_{ij}) = 0$ y $\text{var}(\epsilon_{ij}) = \sigma^2$ y suponemos también que $\sum_{i=1}^a \alpha_i = 0$.

Primeros resultados

Suma y media para cada nivel

$$Y_{i.} = \sum_{j=1}^n Y_{ij} \quad \bar{Y}_{i.} = \frac{1}{n} \sum_{j=1}^n Y_{ij}$$

Suma y media global con $N = an$

$$Y_{..} = \sum_{i=1}^a \sum_{j=1}^n Y_{ij} \quad \bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^a \sum_{j=1}^n Y_{ij}$$

Estimadores puntuales de los parámetros del modelo

$$\hat{\mu} = \bar{Y}_{..} \quad \hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}$$

Prueba de igualdad de medias

Prueba de hipótesis

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a (= 0)$$

$$H_1 : \alpha_i \neq \alpha_{i'} \text{ por algún } i \neq i'$$

Descomposición de la suma de cuadrados

$$SS_T = SS_A + SS_E$$

Suma de cuadrados total $SS_T = \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2$

$$SS_A = n \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2 \quad SS_E = \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2$$

Ejemplo: Tratamiento del fitopatógeno *Rizoctonia solani*

Utilizamos un interesante experimento comentado por la FAO y que hemos adaptado a unos datos que ya conocemos. Crecimiento micelial en términos de diámetro de la colonia (mm) de aislados de *R. solani* en el medio PDA tras 14 horas de incubación.

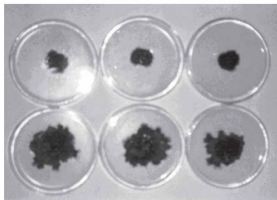
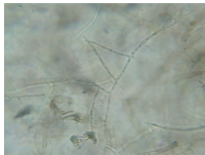


Figura 7. Colonias de *G. citricarpa* (hileras superior) y *G. mangiferae* (hileras inferior) en medio PDA a los siete días.

Font: Wikipedia and <http://scielo.sld.cu/>

Cálculos con R:

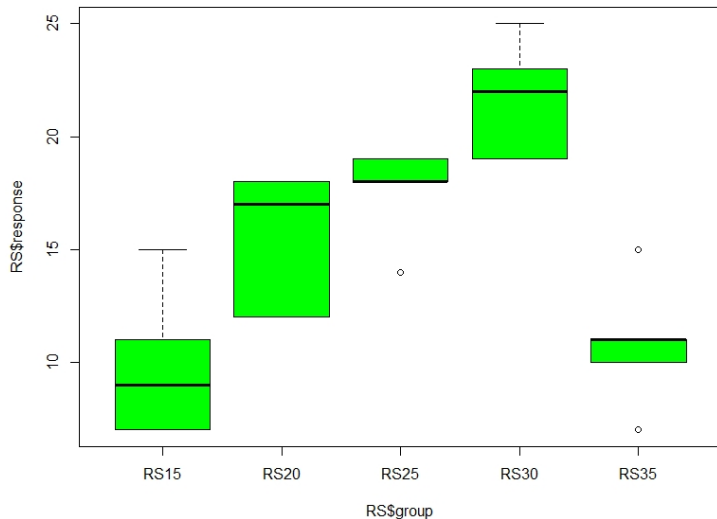
```
#Dataframe example:
grupo<-c("RS15", "RS15", "RS15", "RS15", "RS15", "RS20"
, "RS20", "RS20", "RS20", "RS20", "RS25", "RS25", "
RS25", "RS25", "RS25", "RS30", "RS30", "RS30", "
RS30", "RS30", "RS35", "RS35", "RS35", "RS35", "
RS35")
medida<-c(7, 7, 15, 11, 9, 12, 17, 12, 18, 18, 14,
18, 18, 19, 19, 19, 25, 22, 19, 23, 7, 10, 11, 15,
11)

#Data frame
RS<- data.frame(group=as.factor(grupo),response=medida)

#Box-plot
boxplot(RS$response~RS$group,col="green")
```

Ejemplo: Tratamiento del fitopatógeno *Rizoctonia solani*

Representación gráfica de los niveles del factor TRATAMIENTO:



Datos originales en formato tabular

Crecimiento micelar en términos del diámetro de la colonia (mm) de aislados del hongo *R. solani*, utilizando el medio PDA tras 14 horas de incubación.

| Tratamiento | | | | | | $y_{i\cdot}$ | $\bar{y}_{i\cdot}$ | $\hat{\alpha}_i$ |
|-------------|--------|--------|--------|----|----|------------------|------------------------|------------------|
| RS15 | 7 | 7 | 15 | 11 | 9 | 49 | 9.8 | -5.24 |
| RS20 | 12 | 17 | 12 | 18 | 18 | 77 | 15.4 | 0.36 |
| RS25 | 14 | 18 | 18 | 19 | 19 | 88 | 17.6 | 2.56 |
| RS30 | 19 | 25 | 22 | 19 | 23 | 108 | 21.6 | 6.56 |
| RS35 | 7 | 10 | 11 | 15 | 11 | 54 | 10.8 | -4.24 |
| | | | | | | 376 | 15.04 | 0.00 |
| | | | | | | $y_{\cdot\cdot}$ | $\bar{y}_{\cdot\cdot}$ | suma |
| | SS_T | SS_A | SS_E | | | | | |
| | 637 | 476 | 161 | | | | | |

¿Qué ocurre si sumamos o restamos datos?

1er nivel - 5, 4to nivel + 5: idéntica varianza (SS_E)

| Tractament | | | | | | $y_{i\cdot}$ | $\bar{y}_{i\cdot}$ | $\hat{\alpha}_i$ |
|------------|--------|--------|--------|----|----|------------------|------------------------|------------------|
| RS15 | 2 | 2 | 10 | 6 | 4 | 24 | 4.8 | -10.24 |
| RS20 | 12 | 17 | 12 | 18 | 18 | 77 | 15.4 | 0.36 |
| RS25 | 14 | 18 | 18 | 19 | 19 | 88 | 17.6 | 2.56 |
| RS30 | 24 | 30 | 27 | 24 | 28 | 133 | 26.6 | 11.56 |
| RS35 | 7 | 10 | 11 | 15 | 11 | 54 | 10.8 | -4.24 |
| | | | | | | 376 | 15.04 | 0.00 |
| | | | | | | $y_{\cdot\cdot}$ | $\bar{y}_{\cdot\cdot}$ | suma |
| | SS_T | SS_A | SS_E | | | | | |
| | 1477 | 1316 | 161 | | | | | |

¿Qué ocurre si aumentamos la varianza y mantenemos la media global?

1er y 4to nivel con más varianza, manteniendo igual media global, aumenta (SS_E)

| Tractament | | | | | | $y_{i.}$ | $\bar{y}_{i.}$ | $\hat{\alpha}_i$ |
|------------|--------|--------|--------|----|----|----------|----------------|------------------|
| RS15 | 2 | 7 | 22 | 4 | 14 | 49 | 9.8 | -5.24 |
| RS20 | 12 | 17 | 12 | 18 | 18 | 77 | 15.4 | 0.36 |
| RS25 | 14 | 18 | 18 | 19 | 19 | 88 | 17.6 | 2.56 |
| RS30 | 14 | 30 | 17 | 19 | 28 | 108 | 21.6 | 6.56 |
| RS35 | 7 | 10 | 11 | 15 | 11 | 54 | 10.8 | -4.24 |
| | | | | | | 376 | 15.04 | 0.00 |
| | | | | | | $y_{..}$ | $\bar{y}_{..}$ | suma |
| | SS_T | SS_A | SS_E | | | | | |
| | 1031 | 476 | 555 | | | | | |

Cuadrados medios

$$MS_E = \frac{SS_E}{N - a} \quad MS_A = \frac{SS_A}{a - 1}$$

$$E(MS_E) = \sigma^2 \quad E(MS_A) = \sigma^2 + \frac{n}{a - 1} \sum_{i=1}^a \alpha_i^2$$

Consecuencia

Si H_0 es cierta, MS_E y MS_A estiman el mismo σ^2 .

Distribución de referencia del estadístico F

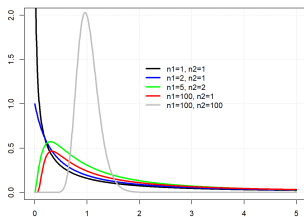
Utilizaremos el estadístico F para comparar los cuadrados medios, este estadístico tiene distribución F .

La distribución F es una distribución de probabilidad continua, muy utilizada en teoría de probabilidad y estadística. También se la conoce como distribución F de Fisher-Snedecor.

Una variable aleatoria de distribución F se construye como el cociente de :

$$F = \frac{U_1/d_1}{U_2/d_2}$$

donde U_1 y U_2 siguen una distribución chi-cuadrado con d_1 y d_2 grados de libertad respectivamente, y son estadísticamente independientes.



Función de densidad de probabilidad F

El estadístico F

Definición

El estadístico F se define como

$$F = \frac{MS_A}{MS_E} = \frac{SS_A/(a-1)}{SS_E/(N-a)}$$

Si H_0 es cierta, F tenderá a ser cercano a 1. Si H_1 es cierta, F tenderá a ser mayor que 1. Por lo tanto, un test muy razonable consiste en rechazar H_0 si F es suficientemente grande.

Test

Si $F \geq c_\alpha$, entonces rechazaremos H_0 .

Es necesario escoger c_α de forma que $P_{H_0}(F \geq c_\alpha) = \alpha$, donde α es el nivel de significación.

El modelo normal

Teorema

Si suponemos que los errores son normales y homocedásticos, es decir:

$$\epsilon_{ij} \sim N(0, \sigma^2), \text{ i.i.d.}$$

entonces bajo H_0 :

- $SS_A/\sigma^2 \sim \chi^2(a - 1)$
- $SS_E/\sigma^2 \sim \chi^2(N - a)$
- e independientes

como consecuencia:

$$F = \frac{\frac{SS_A}{\sigma^2}/(a - 1)}{\frac{SS_E}{\sigma^2}/(N - a)} = \frac{MS_A}{MS_E} \sim F(a - 1, N - a)$$

Tabla de análisis de la varianza (ANOVA)

| Fuente de variación | Suma de cuadrados | Grados de lib. | Cuadrados medias | Estadístico F |
|---------------------------------------|-------------------|----------------|-------------------------|-------------------|
| Entre trat. (niveles del factor A) | SS_A | $a - 1$ | $MS_A = SS_A / (a - 1)$ | $F = MS_A / MS_E$ |
| Error (dentro niveles de A) | SS_E | $N - a$ | $MS_E = SS_E / (N - a)$ | |
| Total | SS_T | $N - 1$ | | |

Criterio de decisión ANOVA

Suponemos que F es el valor calculado sobre los datos.

Si H_0 es cierta, entonces F es un valor de la v.a.

$$F_{N-a}^{a-1} \sim F(a-1, N-a)$$

Dado un nivel de significación α y sea F_α el valor crítico para la distribución de la variable F_{N-a}^{a-1} .

Criterio

Si $F > F_\alpha$, rechazamos H_0 o, de forma equivalente, rechazamos H_0 si el p -valor es suficientemente pequeño:

$$p\text{-valor} = P(F_{N-a}^{a-1} \geq F) \leq \alpha$$

Ejemplo: Tratamiento del fitopatógeno *Rizoctonia solani*

Cálculos con R:

```
#Data frame example
grup<-c("RS15", "RS15", "RS15", "RS15", "RS15", "RS20",
"RS20", "RS20", "RS20", "RS20", "RS25", "RS25", "RS25",
"RS25", "RS25", "RS30", "RS30", "RS30", "RS30", "RS30",
"RS35", "RS35", "RS35", "RS35", "RS35")
medida<-c(7, 7, 15, 11, 9, 12, 17, 12, 18, 18, 14,
18,
18, 19, 19, 19, 25, 22, 19, 23, 7, 10, 11, 15, 11)

#Data frame
RS<-data.frame(group=as.factor(grupo), response=medida)
#ANOVA
aov.res<-aov(response~group,data = RS)
summary(aov.res)
```

Resultados: Tabla ANOVA

| Fuente de variación | Suma de cuadrados | Grados de lib. | Cuadrados medias | Estadístico F |
|---------------------|-------------------|----------------|------------------|---------------|
| Tratamiento | 475.76 | 4 | 118.94 | 14.76 |
| Error | 161.20 | 20 | 8.06 | |
| Total | 636.96 | 24 | | |

El valor crítico es $F_{0.05}(4, 20) = 2.87$ y el estadístico $F = 14.76$
Como $14.76 > 2.87$, rechazamos H_0 . También como el p -valor(9.13×10^{-6}) < 0.05 , rechazamos H_0 .

Intervalos de confianza para las medias

Bajo la hipótesis de normalidad

$$\hat{\mu}_i = \bar{Y}_i \sim N(\mu_i, \sigma^2/n) \implies \frac{\bar{Y}_i - \hat{\mu}_i}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Si sustituimos σ^2 por su estimador MS_E , entonces:

$$\frac{\bar{Y}_i - \hat{\mu}_i}{\sqrt{MS_E/n}} \sim t(N - a)$$

Intervalo de confianza para μ_i

$$\bar{Y}_i \pm t_{\alpha}(N - a) \sqrt{\frac{MS_E}{n}}$$

Ejemplo: Para nivel del antifúngico RS30, un IC al 95% es:

$$21.60 \pm 2.086 \sqrt{\frac{8.06}{5}} = 21.60 \pm 2.65 \implies 18.95 \leq \mu_{RS30} \leq 24.25$$

Validez del modelo

Como son los errores?

Es cierto que los errores verifican las condiciones de Gauss-Markov?

1. $E(\epsilon_{ij}) = 0$
2. $\text{var}(\epsilon_{ij}) = \sigma^2$
3. los errores son independientes
4. los errores son normales

En resumen:

$$\epsilon_{ij} \sim N(0, \sigma^2) \quad \text{i.i.d.}$$

La respuesta es el análisis de los residuos:

- ▶ Análisis gráfico: estructura, gráficos de probabilidad, boxplots, . . .
- ▶ Pruebas de significación: homocedasticidad (prueba de Bartlett), normalidad (Kolmogorov-Smirnov), . . .

Homocedasticidad

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_a^2$$
$$H_1 : \sigma_i^2 \neq \sigma_{i'}^2 \text{ per algun } i \neq i'$$

Prueba de Bartlett

Si las poblaciones son normales, para contrastar las hipótesis el estadístico es:

$$\chi_B^2 = 2.3026 \frac{(N - a) \log_{10} \hat{S}^2 - \sum_{i=1}^a (n_i - 1) \log_{10} \hat{S}_i^2}{1 + \frac{1}{3(a-1)} \left(\sum_{i=1}^a \left(\frac{1}{n_i - 1} \right) - \frac{1}{N - a} \right)}$$

$$\hat{S}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \quad \hat{S}^2 = \frac{1}{N - a} \sum_{i=1}^a (n_i - 1) \hat{S}_i^2$$

Si H_0 es cierta, el estadístico χ_B^2 sigue una $\chi^2(a - 1)$ (aprox.)
Rechazaremos H_0 si $\chi_B^2 \geq \chi_\alpha^2(a - 1)$.

¿Qué niveles son realmente diferentes?

Suponemos que el test ANOVA ha dado un valor significativo.

Dos cuestiones problemáticas:

- ▶ Hacer simples pruebas t de nivel α en todas las parejas hace que se incremente las probabilidades de error de tipo I, el nivel de significación real es superior al nominal α .

¿Cómo controlamos este desajuste de la probabilidad de error?

- ▶ Hace falta comparaciones planificadas previamente?
Si solamente comparamos los más diferentes, distorsionamos el nivel de significación.

Probabilidad global de error de tipo I

Indicamos para E_i el hecho de cometer el error de tipo I en la i -ésima vez que repetimos una prueba.

Repetido k veces, la probabilidad *global* es:

$$\begin{aligned} P(E_1 \cup E_2 \cup \dots \cup E_k) = & \\ & P(E_1) + P(E_2) + \dots + P(E_k) \\ & - P(E_1 \cap E_2) - P(E_1 \cap E_3) - \dots - P(E_{k-1} \cap E_k) \\ & \vdots \\ & + (-1)^{k+1} P(E_1 \cap E_2 \cap \dots \cap E_k) \end{aligned}$$

Método de Bonferroni

Utiliza la desigualdad de Bonferroni:

$$P(E_1 \cup E_2 \cup \dots \cup E_k) \leq P(E_1) + P(E_2) + \dots + P(E_k)$$

Definición

Un nivel de significación global inferior a α queda asegurado si el nivel de significación de las k pruebas individuales se fija en:

$$P(E_i) = \alpha/k \quad \text{per a tot } i = 1, \dots, k$$

Es un método muy general a costa de una gran pérdida de potencia.

Diferencia mínima significativa (LSD)

Definición

La diferencia mínima significativa o Least Significant Difference (LSD) es el estadístico t de comparación de dos muestras cuando tenemos información de a grupos o poblaciones:

$$t_{ik} = \frac{\bar{Y}_i - \bar{Y}_k}{\sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}}$$

Rechazaremos $H_0 : \mu_i = \mu_k$ si $|t_{ik}| \geq t_{\alpha}(N - a)$

Este método estima mejor σ^2 y con mas grados de libertad (es mas potente que la t sencilla), pero el nivel simultaneo real es superior a α (en múltiples comparaciones).

Ejemplo

En los datos de crecimiento (diámetro) del hongo según tratamiento: $a = 5$, $n = 5$, $N - a = 20$, $t_{0.05}(20) = 2.086$, $MS_E = 8.06$. Por lo que el ajustado es:

$$\begin{aligned}t_{\alpha}(N - a)\sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_k} \right)} &= t_{\alpha}(N - a)\sqrt{\frac{2MS_E}{n}} \\ &= 2.086\sqrt{\frac{2 \times 8.06}{5}} = 3.75\end{aligned}$$

Entonces comparamos todas las diferencias con este valor.

Por ejemplo, μ_{RS25} con μ_{RS30} : $|21.6 - 17.6| = 4 > 3.75 \Rightarrow$ son diferentes.

Solamente aceptamos las igualdades $\mu_{RS15} = \mu_{RS35}$ y

$\mu_{RS20} = \mu_{RS25}$.

Prueba de Duncan

La prueba de Duncan o de los rangos múltiples tiene el siguiente procedimiento:

1. Ordenamos las medias en orden creciente

$$\bar{Y}_{(1)} < \bar{Y}_{(2)} < \dots < \bar{Y}_{(a)}$$

2. Estimamos el error estándar de las medias para un valor común

$$S_{\bar{Y}_i} = \sqrt{\frac{MS_E}{n_0}} \quad n_0 = \begin{cases} n & \text{si } n_1 = \dots = n_a = n \\ \frac{a}{\sum_{i=1}^a \frac{1}{n_i}} & \text{si } n_i \neq n_k \end{cases}$$

3. Con la tabla $r_\alpha(p, f)$, donde $p = 2, 3, \dots, a$; $f = g.l.l.$

$$\bar{Y}_{(1)} < \dots < \underbrace{\bar{Y}_{(h)} < \dots < \bar{Y}_{(k)}}_p < \dots < \bar{Y}_{(a)}$$

rango mínimo significativo = $R_p = r_\alpha(p, f)S_{\bar{Y}_i}$.

Prueba de Duncan

1. Se comprueba si es cierto que:

$$\begin{array}{ccccccc} \bar{Y}_{(a)} - \bar{Y}_{(1)} > R_a, & \bar{Y}_{(a)} - \bar{Y}_{(2)} > R_{a-1}, & \dots & \bar{Y}_{(a)} - \bar{Y}_{(a-1)} > R_1 \\ & \bar{Y}_{(a-1)} - \bar{Y}_{(1)} > R_{a-1}, & \dots & \bar{Y}_{(a-1)} - \bar{Y}_{(a-2)} > R_1 \\ & & \ddots & & \vdots \\ & & & & \bar{Y}_{(2)} - \bar{Y}_{(1)} > R_1 \end{array}$$

2. Si $\bar{Y}_{(i)} - \bar{Y}_{(k)} > R_{i-k+1}$ es rechaza $H_0 : \mu_i = \mu_k$.

Para evitar contradicciones, si esta comparación no es significativa, por definición todas las “interiores” para i', k' , con $i < i' < k' < k$, tampoco lo son.

Propiedades de la prueba de Duncan

- ▶ Cuantas más medias haya en el grupo, más grande es R_p y mas difícil es encontrar una diferencia significativa.
- ▶ Para dos consecutivas es equivalente al LSD.
- ▶ $r_\alpha(p, f)$ garantiza “nivel de protección”: todas las comparaciones dentro de un grupo de p consecutivos, tienen una significación simultánea $1 - (1 - \alpha)^{p-1}$.
- ▶ Nivel de significación real bastante mas grande que α (gran riesgo de error I) pero también alta potencia.

Ejemplo de trabajo con el hongo *Rhizoctonia solani*

- ▶ Las medias muestrales ordenadas son:

$$\begin{aligned}\bar{y}_{(1)} = \bar{y}_{1.} = 9.8 & \quad \bar{y}_{(2)} = \bar{y}_{5.} = 10.8 & \quad \bar{y}_{(3)} = \bar{y}_{2.} = 15.4 \\ \bar{y}_{(4)} = \bar{y}_{3.} = 17.6 & \quad \bar{y}_{(5)} = \bar{y}_{4.} = 21.6\end{aligned}$$

- ▶ La estimación común del error estándar es:

$$MS_E = 8.06, \quad N = 25, \quad n = 5 \Rightarrow S_{\bar{Y}_i} = \sqrt{8.06/5} = 1.27$$

- ▶ En la tabla de Duncan, para $N - a = 20$ g.ll.:

$$\begin{aligned}r_{0.05}(2, 20) = 2.95 & \quad r_{0.05}(3, 20) = 3.10 \\ r_{0.05}(4, 20) = 3.18 & \quad r_{0.05}(5, 20) = 3.25\end{aligned}$$

$$R_p = r_\alpha(p, f) S_{\bar{Y}_i} = r_{0.05}(p, 20) \times 1.27$$

$$R_2 = 3.75 \quad R_3 = 3.94 \quad R_4 = 4.04 \quad R_5 = 4.13$$

Ejemplo de trabajo con el hongo *Rhizoctonia solani*

Todas las diferencias:

$$\bar{y}_{(5)} - \bar{y}_{(1)} = \bar{y}_{4.} - \bar{y}_{1.} = 21.6 - 9.8 = 11.8 > R_5 = 4.13 \Rightarrow \text{signif.}$$

$$\bar{y}_{(5)} - \bar{y}_{(2)} = \bar{y}_{4.} - \bar{y}_{5.} = 21.6 - 10.8 = 10.8 > R_4 = 4.04 \Rightarrow \text{signif.}$$

$$\bar{y}_{(5)} - \bar{y}_{(3)} = \bar{y}_{4.} - \bar{y}_{2.} = 21.6 - 15.4 = 6.2 > R_3 = 3.94 \Rightarrow \text{signif.}$$

$$\bar{y}_{(5)} - \bar{y}_{(4)} = \bar{y}_{4.} - \bar{y}_{3.} = 21.6 - 17.6 = 4.0 > R_2 = 3.75 \Rightarrow \text{signif.}$$

$$\bar{y}_{(4)} - \bar{y}_{(1)} = \bar{y}_{3.} - \bar{y}_{1.} = 17.6 - 9.8 = 7.8 > R_4 = 4.04 \Rightarrow \text{signif.}$$

⋮

$$\bar{y}_{(2)} - \bar{y}_{(1)} = \bar{y}_{5.} - \bar{y}_{1.} = 10.8 - 9.8 = 1.0 > R_2 = 3.75 \Rightarrow \text{no signif.}$$

Resultat

Las únicas diferencias no significativas son 3 con 2 y 5 con 1.

Ejemplo de trabajo con el hongo *Rhizoctonia solani*

¡Alerta!

¿Qué nivel de significación tenemos en realidad al hacer todas las comparaciones dentro del grupo de la (4) hasta la (1)?

Respuesta:

Como está formado por $p = 4$ medias, tenemos que el verdadero nivel de significación, es decir, la probabilidad de encontrar alguna falsa diferencia es:

$$1 - (1 - \alpha)^{p-1} = 1 - (1 - 0.05)^{4-1} = 0.1426$$

Prueba de Newman-Keuls

Esta prueba es muy parecida a la de Duncan pero basada en $q_\alpha(p, f)$, percentil α superior del *rango studentizado* para un grupo de p medias:

$$\frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{MS_E/n_0}}$$

Las diferencias de medias se comparan con $K_p = q_\alpha(p, f)S_{\bar{Y}_i}$, como se hace con R_p .

Prueba HSD de Tukey

Es una de las más utilizadas, y frecuentemente se toma como referencia.

La prueba se llama HSD de Tukey o Tukey-Cramér por las siglas:
Honestly **S**ignificant **D**ifference

- ▶ El error estándar de cada \bar{Y}_i . se estima de forma común como

$$S_{\bar{Y}_i} = \sqrt{\frac{MSE}{n_0}} \quad \text{on} \quad n_0 = \begin{cases} n & \text{si } n_1 = \dots = n_a \\ \frac{a}{\sum_{i=1}^a \frac{1}{n_i}} & \text{si } n_i \neq n_j \end{cases}$$

- ▶ Hay un único valor crítico $T = q_\alpha(a, f)S_{\bar{Y}_i}$. en la tabla de Newman-Keulen con lo comparamos todas las diferencias $|\bar{Y}_i - \bar{Y}_j|$.
- ▶ Prueba para todas las parejas de medias (no para establecer grupos homogéneos como las pruebas de Duncan y de Newman-Keulen).

Ejemplo: RS tratamiento hongo

Determinación del valor crítico:

$$q_{\alpha}(a, f) = q_{0.05}(5, 20) = 4.23$$

$$S_{\bar{Y}_i} = \sqrt{\frac{MSE}{n_0}} = \sqrt{\frac{8.060}{5}} = 1.27$$

$$T = q_{\alpha}(a, f)S_{\bar{Y}_i} = 4.23 \times 1.27 = 5.37$$

Resultat

Cualquier diferencia, para ser significativa, debe superar el valor crítico de 5.37.

Prueba de Scheffé

- ▶ La prueba de Scheffé utiliza el siguiente valor crítico:

$$c_{\alpha}(i, k) = S_{ik} \sqrt{(a-1)F_{\alpha}(a-1, N-a)}$$

on $S_{ik} = \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}$

y se rechaza $H_0 : \mu_i = \mu_k$ si $|\bar{y}_{j\cdot} - \bar{y}_{k\cdot}| \geq c_{\alpha}(i, k)$.

- ▶ Es un test con nivel de significación simultáneo $\leq \alpha$.
- ▶ Los intervalos de confianza de la forma

$$\bar{y}_{j\cdot} - \bar{y}_{k\cdot} \pm c_{\alpha}(i, k)$$

son también I.C. simultáneos para las diferencias $\mu_i - \mu_k$ con nivel de confianza, como mínimo de, $1 - \alpha$.

Prueba de Dunnet

- ▶ Comparación de los otros tratamientos contra un tratamiento control (digamos que es el último):

$$H_0(i) : \mu_i = \mu_a, \quad i = 1, \dots, a - 1$$

- ▶ Rebutjarem $H_0(i)$ si

$$|\bar{y}_{i\cdot} - \bar{y}_{a\cdot}| \geq d_\alpha(a - 1, N - a) \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}$$

donde $d_\alpha(a - 1, f)$ es un valor crítico encontrado por Dunnet que garantiza un nivel de significación simultáneo α para los $a - 1$ contrastes.

Comparación entre métodos

Hay diversidad de opiniones entre los expertos.

Alguno criterios son:

- ▶ De menor a mayor probabilidad de error tipo I:
Scheffé, Tukey, Newman-Keuls, Duncan
- ▶ Potencia estadística: parece que el mismo orden.
- ▶ Pocas comparaciones previamente planificadas: posiblemente preferible más potencia aunque el nivel de significación sea alto. Muchas: al contrario.
- ▶ También se han de valorar las consecuencias de error I y II.
- ▶ Utilizaremos Dunnet para la comparación de un control contra los otros tratamientos.

False Discovery Ratio (FDR)

Imaginemos ahora que tenemos múltiples variables (Y_i) y tenemos sólo 2 grupos. Pensemos por un momento que sucede si sólo disponemos de una única variable con **un único test estadístico**.

$H_0 : \mu_T = \mu_R \rightarrow$ no diferencia

$H_1 : \mu_T \neq \mu_R \rightarrow$ diferencia

Un valor p (p-value) es la probabilidad de que el estadístico de prueba esté más lejos en la cola de la distribución (de referencia) de lo que has observado, *si la hipótesis nula es verdadera*.

Procedimiento de Test: Escogemos un valor (cut-off) ' α ' para nuestro p-value ' p ', y concluir la importancia (por ejemplo, efecto del tratamiento o diferencia de grupos) para $p < \alpha$.

Esto se justifica como dar sólo α a la probabilidad del falso positivo (probabilidad del error de tipo I). Por ejemplo en ANOVA, $\hat{\alpha}_i \neq 0$ (Efecto de uno de los tratamientos) sólo si su p-value es menor que el riesgo aceptado de un falso descubrimiento (false discovery) para cada coeficiente. Este problema es muy común en Big Data, donde se trata de tomar muchas decisiones difíciles.

El problema de la multiplicidad de tests estadísticos

- ▶ La probabilidad α sólo debe utilizarse para un único test estadístico. Si (incorrectamente) realizamos muchos tests estadísticos, sobre un $\alpha \times 100\%$ de los test nulos aparecerán erróneamente como significativos.
- ▶ Suponga que debemos comparar las medias de dos grupos (T test) para 100 variables de respuesta diferentes (experimento con sólo dos grupos pero muchas variables!), supongamos que 5 de 100 coeficientes de regresión son en realidad influyente y que cuando hacemos los test nos dan significativos.
- ▶ Calculemos el test estadístico (T test) para el resto ($n=95$) utilizando $\alpha = 0.05$:
- ▶ Cuando se rechaza H_0 para 5% de las 95 variables descartadas anteriormente, $4.75/9.75 \approx 50\%$ de los test significativos obtenidos son descubrimientos falsos (false discoveries)!

$$4.75/9.75 = 95 * 0.05 / (95 * 0.05 + 100 * 0.05)$$

- ▶ Este ratio es lo que se denomina "False Discovery Proportion" (FDP). Puede ser realmente grande con una pequeña tasa verdadera no nula (true non-Null rate).

False Discovery Rate (FDR)

En lugar de centrarnos en pruebas estadísticas individuales, consideraremos:

$$\text{False Discovery Proportion (FDP)} = \frac{\text{falsos positivos}}{\text{numero test significativos}}$$

Al ser FDP una propiedad de nuestro modelo ajustado, no podemos conocerlo, pero si podemos controlar su esperanza (cantidad media producida):

$$\text{False Discovery Rate, FDR} = E[\text{FDP}].$$

FDR es el análogo de α que hemos utilizado para los tests pero de manera multivariante (agregado de tests) . Pero, ¿Cómo lo controlamos?

Control del False Discovery Rate (FDR)

Supongamos que queremos asegurar que $FDR \leq q$ (por ejemplo, 0.1).

El algoritmo de Benjamini + Hochberg (BH) permite el control del FDR y puede describirse como:

- Ordenar (Rank) los N p-values obtenidos, de menor a mayor, $P_{(1)} \cdots P_{(N)}$.
- Calcular el punto de corte (cut-off) para el p-value como $p^* = \max \left\{ p_{(k)} : p_{(k)} \leq q \frac{k}{N} \right\}$.

Si la región de rechazo es p-values $\leq p^*$, entonces $FDR \leq q$. q es la pendiente de una línea que define nuestra región de rechazo.

Precaución: se supone independencia (aproximada) entre las pruebas estadísticas.

Ejemplo de uso del FDR: RNAseq

Scan large DNA sequences for association with disease.

La secuenciación de ARN es una técnica de secuenciación basada en tecnología particular que utiliza la secuenciación de próxima generación (NGS) para revelar la presencia y cantidad de ARN en una muestra biológica en un momento dado, analizando el transcriptoma celular que cambia continuamente.

El uso más simple pero a menudo más potente de RNA-Seq es encontrar diferencias en la expresión génica entre dos o más afecciones (p. Ej., Tratado frente a no tratado); Este proceso se llama expresión diferencial. Los resultados se denominan con frecuencia genes expresados diferencialmente (DEG) y estos genes pueden estar regulados hacia arriba o hacia abajo (es decir, más o menos en la condición de interés).

Algunas preguntas que podemos plantearnos utilizando esta tecnología:
¿Qué genes están diferencialmente expresados? Pensemos que pueden ser miles. risk?

Levaduras sometidas a stress abiótico

Monleón (2020) en su trabajo sobre RNAseq describe un experimento de expresión diferencial de genes de la levadura de cerveza (*Saccharomyces cerevisiae*). Se toman diferentes réplicas de un cultivo de levadura y unas cuantas se someten a stress abiótico y otras no (control). Se secuencia el DNA de las levaduras y se obtienen los análisis de la expresión de los genes mediante RNAseq, finalmente se obtienen los resultados de unos 2000-3000 genes de los grupos estrés y control.

Así obtendremos 2000-3000 p-values procedentes de 2000-3000 test estadísticos, uno para cada gen (variables Y) entre los grupos control y estrés (Debe elegirse el test más adecuado en función de las unidades de las variables, ej: t-test, test de proporciones, etc)
RNAseq P-values: ¿Cuáles son significativos? Recordar: los valores p de la distribución nula son uniformes.

Controlando el False Discovery Rate (FDR) en nuestro análisis

La pendiente de la línea del FDR es $q/[\# \text{ de las variables}]$.

Tendremos que comprobar cuantos del total de tests (2000-3000) según el FDR (por ejemplo de $1e-3$) son falsos descubrimientos

Todo lo que hay que saber sobre FDR

Los valores p de la distribución nula son uniformes y N de ellos clasificados y graficados deben estar a lo largo de una línea con pendiente $1/N$.

FDP es el número de descubrimientos falsos dividido por el número de resultados de prueba llamados significativos. No conoces el FDP.

Podemos controlar $FDR = E[FDP]$ para que sea $\leq q$ entre N tests

- ▶ clasificar y trazar los valores p contra el rango/ N
- ▶ dibujar la línea con pendiente q/N
- ▶ encontrar el punto máximo donde los valores p cruzan esta línea, y usa ese punto para establecer tu región de rechazo.

Modelo con un factor aleatorio

Toni Monleón-Getino

Junio de 2020

Un ejemplo: peso de de biomasa en un bioreactor (kg)

Queremos comprobar la fiabilidad de un nuevo bioreactor para la producción de un cultivo celular con microorganismos de importancia biológica. Realizamos cada día y con las mismas condiciones experimentales 3 repeticiones al atzar y en días sucesivos fabricamos 5 lotes de producto mediante el bioreactor:

| Lote fabricado | Peso en kg | | |
|----------------|------------|-------|-------|
| 1 | 3.205 | 4.100 | 3.550 |
| 2 | 2.300 | 4.300 | 3.435 |
| 3 | 3.300 | 2.560 | 3.600 |
| 4 | 4.560 | 3.440 | 4.340 |
| 5 | 3.560 | 4.550 | 2.550 |

Lote es un factor aleatorio (factor batch o lote) con 5 niveles.

Concepto y definición de factor aleatorio

Definición

Los a niveles del factor estudiado son una muestra aleatoria de los muchos posibles.

El modelo es ahora:

$$Y_{ij} = \mu + A_i + \epsilon_{ij}$$

donde

- ▶ μ es, como antes, una media general,
- ▶ A_i es una variable aleatoria, no una constante,
- ▶ ϵ_{ij} es el error no explicado por el modelo.

Las condiciones para poder aplicar el modelo son:

$E(A_i) = 0$, $E(\epsilon_{ij}) = 0$, $\text{var}(A_i) = \sigma_A^2$, $\text{var}(\epsilon_{ij}) = \sigma^2$, A_i i ϵ_{ij} son independientes (i.i.d)

Por tanto, $\text{var}(Y_{ij}) = \sigma_A^2 + \sigma^2$. σ_A^2 , σ^2 son las componentes de la varianza del modelo.

Pruebas de significación

Ahora el contraste de hipótesis más importante es:

$$H_0 : \sigma_A^2 = 0$$

$$H_1 : \sigma_A^2 > 0$$

La descomposición de la suma de cuadrados continúa siendo válida, $SS_T = SS_A + SS_E$, pero los cuadrados medios tienen otro sentido:

$$E(MS_E) = \sigma^2 \quad E(MS_A) = \sigma^2 + n\sigma_A^2$$

El estadístico $F = MS_A/MS_E$ continúa siendo adecuado: Si H_0 es cierta, F será cercano a 1, si H_0 es falsa, F tenderá a ser grande.

Pruebas de significación

Bajo condiciones de normalidad

$$A_i \sim N(0, \sigma_A^2) \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

Test F

Si H_0 es cierta, $F \sim F(a - 1, N - a)$ y la tabla ANOVA, el p -valor, etc. son igual de válidos que el caso de un factor fijo.

Alerta!

La interpretación de los resultados, en este caso es diferente..

Ejemplo de los lotes de producción de biomasa (Efecto Batch)

Cálculos con R:

```
lot <- gl(5,3)
pes <- c(3.205,4.1,3.55,2.3,4.3,3.435,3.3,2.56,
        3.6,4.56,3.44,4.34,3.56,4.55,2.55)
model <- lm(pes ~ lot)
anova(model)
```

Analysis of Variance Table

Response: pes

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|--------|
| lot | 4 | 1.5635 | 0.39088 | 0.6861 | 0.6177 |
| Residuals | 10 | 5.6971 | 0.56971 | | |

Resultado: Parece que la variación en el peso de la biomasa de las replicas atribuible al factor lote no es significativa ($p=0.6177$).

Correlació intraclásica

En un modelo de un sólo factor aleatorio como el presentado en el ejemplo, puede haber correlación intraclásica:

$$\text{cov}(Y_{ij}, Y_{ik}) = \text{cov}(\mu + A_i + \epsilon_{ij}, \mu + A_i + \epsilon_{ik}) = \text{var}(A_i) = \sigma_A^2$$

Definición

$$\text{cor}(Y_{ij}, Y_{ik}) = \frac{\text{cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{var}(Y_{ij})\text{var}(Y_{ik})}} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}$$

Si la componente de la varianza asociada al factor no es significativa, entonces la correlación intraclásica no es significativa y al revés.

Estimación puntual

Los estimadores puntuales de las componentes de la varianza son:

$$\begin{aligned}\hat{\sigma}^2 &= MS_E \\ \hat{\sigma}_A^2 &= \frac{MS_A - MS_E}{n_0}\end{aligned}$$

donde

$$n_0 = \begin{cases} n & \text{caso balanceado} \\ \frac{1}{a-1} \left(\sum_{i=1}^a n_i - \frac{\sum_{i=1}^a n_i^2}{\sum_{i=1}^a n_i} \right) & \text{pel caso no balanceado} \end{cases}$$

Alerta!

Numéricamente, pueden obtenerse posibles valores absurdos (negativos) de $\hat{\sigma}_A^2$.

Intervalos de confianza para las componentes de la varianza y correlación

Del hecho que (distribución de la varianza):

$$\frac{(N - a)MS_E}{\sigma^2} \sim \chi^2(N - a)$$

obtenemos el intervalo de confianza al $100(1 - \alpha)\%$

$$\frac{(N - a)MS_E}{\chi_{\alpha/2}^2(N - a)} \leq \sigma^2 \leq \frac{(N - a)MS_E}{\chi_{1-\alpha/2}^2(N - a)}$$

Un intervalo similar para σ_A^2 es difícil de obtener.

En cambio, para la proporción $\text{cor}(Y_{ij}, Y_{ik}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}$ tenemos

$$\frac{c(\alpha/2)}{1 + c(\alpha/2)} \leq \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2} \leq \frac{c(1 - \alpha/2)}{1 + c(1 - \alpha/2)}$$

donde $c(p) = \frac{1}{n} \left(\frac{MS_A}{MS_E} F_{p(a-1, N-a)} - 1 \right)$.

ANOVA N-WAY

Toni Monleón-Getino

Junio de 2020

Estructura de los datos (2 factores, fijos y cruzados (interacción))

La teoría anterior puede generalizarse a dos o más factores, que pueden estar combinados (interacción, jerárquización) o no (bloques).

- ▶ Diseño no balanceado de dos factores A i B , cruzados, con a y b niveles respectivamente,

| | | | |
|----------|--------------------------------|-----|--------------------------------|
| | B_1 | ... | B_b |
| A_1 | $y_{111}, \dots, y_{11n_{11}}$ | ... | $y_{1b1}, \dots, y_{1bn_{1b}}$ |
| \vdots | \vdots | | \vdots |
| A_a | $y_{a11}, \dots, y_{a1n_{a1}}$ | ... | $y_{ab1}, \dots, y_{abn_{ab}}$ |

- ▶ Si el diseño es balanceado,

$$n_{ij} = n, \text{ para todo } i = 1, \dots, a \text{ y } j = 1, \dots, b$$

Modelo lineal

Modelo lineal asociado

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

para $i = 1, \dots, a$, $j = 1, \dots, b$ y $k = 1, \dots, n_{ij}$

Con las restricciones:

$$\sum_{i=1}^a \alpha_i = 0 \quad \sum_{j=1}^b \beta_j = 0$$

$$\sum_{j=1}^b \gamma_{ij} = 0 \quad \text{per a } i = 1, \dots, a \quad \sum_{i=1}^a \gamma_{ij} = 0 \quad \text{per a } j = 1, \dots, b$$

y las condiciones:

$$E(\epsilon_{ijk}) = 0 \quad \text{var}(\epsilon_{ijk}) = \sigma^2 \quad \text{para todo } i, j, k$$

Ejemplo numérico de este modelo

Variable de respuesta (Y) productividad agrícola. Datos de 3 fertilizantes (Factor A) y 4 variedades (Factor A):

| Fertilizante | Variedad | | | | | | | |
|--------------|----------|----|----|----|----|----|----|----|
| | A | | B | | C | | D | |
| 1 | 35 | 26 | 45 | 39 | 24 | 23 | 55 | 48 |
| | 38 | 20 | 39 | 43 | 36 | 29 | 39 | 49 |
| 2 | 55 | 44 | 64 | 57 | 58 | 74 | 68 | 61 |
| | 68 | 64 | 62 | 61 | 49 | 69 | 60 | 75 |
| 3 | 97 | 89 | 93 | 91 | 89 | 98 | 82 | 78 |
| | 92 | 99 | 82 | 98 | 85 | 87 | 89 | 92 |

Sumas, medias y estimaciones

$$Y_{i..} = \sum_{j=1}^b \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{i..} = \frac{Y_{i..}}{bn} \quad i = 1, \dots, a$$

$$Y_{.j.} = \sum_{i=1}^a \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{.j.} = \frac{Y_{.j.}}{an} \quad j = 1, \dots, b$$

$$Y_{ij.} = \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{ij.} = \frac{Y_{ij.}}{n} \quad i = 1, \dots, a \quad j = 1, \dots, b$$

$$Y_{...} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{...} = \frac{Y_{...}}{abn}$$

Estimaciones de los parámetros

$$\hat{\mu} = \bar{Y}_{...} \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...} \quad \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$$

$$\hat{\gamma}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

Descomposición de la suma de cuadrados

$$SS_A = bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 \quad a - 1 \text{ g.d.ll.}$$

$$SS_B = an \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \quad b - 1 \text{ g.d.ll.}$$

$$SS_{AB} = n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \quad (a - 1)(b - 1) \text{ g.d.ll.}$$

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij.})^2 \quad ab(n - 1) \text{ g.d.ll.}$$

$$SS_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2 = SS_A + SS_B + SS_{AB} + SS_E$$

con $abn - 1$ g.d.ll.

Cuadrados medios y esperanzas

$$MS_A = \frac{SS_A}{a-1}$$

$$E(MS_A) = \sigma^2 + \frac{bn \sum_{i=1}^a \alpha_i^2}{a-1}$$

$$MS_B = \frac{SS_B}{b-1}$$

$$E(MS_B) = \sigma^2 + \frac{an \sum_{j=1}^b \beta_j^2}{b-1}$$

$$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$$

$$E(MS_{AB}) = \sigma^2 + \frac{n \sum_{i=1}^a \sum_{j=1}^b \gamma_{ij}^2}{(a-1)(b-1)}$$

$$MS_E = \frac{SS_E}{ab(n-1)}$$

$$E(MS_E) = \sigma^2$$

Contrastes sobre los parámetros del modelo

- ▶ ¿Es significativo el efecto del factor A ?

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a = 0$$

$$H_1 : \alpha_i \neq 0$$

- ▶ ¿Es significativo el efecto del factor B ?

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_b = 0$$

$$H_1 : \beta_j \neq 0$$

- ▶ ¿Es significativa la interacción?

$$H_0 : \gamma_{11} = \gamma_{12} = \dots = \gamma_{ab} = 0$$

$$H_1 : \gamma_{ij} \neq 0$$

Tabla ANOVA

La tabla ANOVA asociada a este modelo es la siguiente:

| Fuente de variación | Suma de cuadrados | Grados de libertad | Cuadrados medios | Estadístico F |
|----------------------------|--------------------------|---------------------------|-------------------------|-----------------------------------|
| Trat. A | SS_A | $a - 1$ | MS_A | MS_A/MS_E |
| Trat. B | SS_B | $b - 1$ | MS_B | MS_B/MS_E |
| Interacción | SS_{AB} | $(a - 1)(b - 1)$ | MS_{AB} | MS_{AB}/MS_E |
| Residuo | SS_E | $ab(n - 1)$ | MS_E | |
| Total | SS_T | $abn - 1$ | | |

Modelo normal

Si los errores son i.i.d. i todos $\epsilon_{ijk} \sim N(0, \sigma)$,

- ▶ Significación del factor A : $F \sim F(a - 1, ab(n - 1))$
- ▶ Significación del factor B : $F \sim F(b - 1, ab(n - 1))$
- ▶ Significación de la interacción $A*B$:
 $F \sim F((a - 1)(b - 1), ab(n - 1))$

Modelo normal

En un modelo lineal normal los valores críticos y los p -valors se obtienen de la tabla de la distribución F .

Ejemplo Fertilizante*Variedad

¿Cómo deben introducirse los datos en R?

```
Fertilizante <- c("F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1",  
"F1", "F1", "F1", "F1", "F1", "F1", "F1", "F2", "F2", "F2", "F2", "F2",  
"F2", "F2", "F2", "F2", "F2", "F2", "F2", "F2", "F2", "F2", "F3",  
"F3", "F3", "F3", "F3", "F3", "F3", "F3", "F3", "F3", "F3", "F3", "F3",  
"F3", "F3", "F3")
```

```
Variedad <- c("A", "A", "A", "A", "B", "B", "B", "B", "C", "C", "C", "C",  
"D", "D", "D", "D", "A", "A", "A", "A", "B", "B", "B", "B", "C", "C",  
"C", "C", "D", "D", "D", "D", "A", "A", "A", "A", "B", "B", "B", "B",  
"C", "C", "C", "C", "D", "D", "D", "D")
```

```
Produccion <- c(35, 26, 38, 20, 45, 39, 39, 43, 24, 23, 36, 29, 55, 48,  
39, 49, 55, 44, 68, 64, 64, 57, 62, 61, 58, 74, 49, 69, 68, 61, 60,  
75, 97, 89, 92, 99, 93, 91, 82, 98, 89, 98, 85, 87, 82, 78, 89, 92)
```

#Data frame blat

```
blat <- data.frame(Fertilizante=as.factor(Fertilizante),  
Variedad=as.factor(Variedad), Produccion=Produccion)
```


Ejemplo de Fertilizante*Variedad

Cálculos amb R:

```
#Tabla ANOVA
res.aov <- aov(Produccion ~ Fertilizante * Variedad, data=blat)
summary(res.aov)
```

Analysis of Variance Table

Response: Produccio

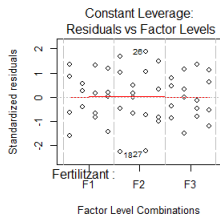
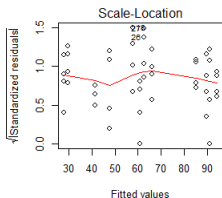
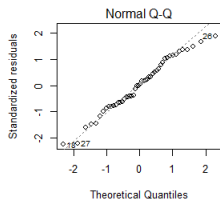
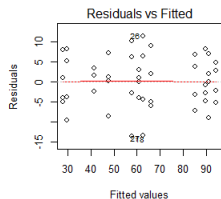
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------------|----|---------|---------|----------|---------------|
| Fertilizante | 2 | 22764.9 | 11382.4 | 230.6601 | < 2.2e-16 *** |
| Variedad | 3 | 331.7 | 110.6 | 2.2409 | 0.100189 |
| Fertilit:Variedad | 6 | 1052.1 | 175.4 | 3.5535 | 0.007246 ** |
| Residuals | 36 | 1776.5 | 49.3 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

El factor Fertilizante y la interacción son claramente significativas ($p < 0.05$).

Comprobación del modelo

En estos gráficos pueden verse los diagramas de dispersión y de normalidad de los errores. De una manera visual puede comprobarse su normalidad y la homocedasticidad entre niveles del factor principal.



Estimación de los parámetros del modelo

Estimaciones de los coeficientes (medias) del modelo lineal

Produccion ~ *Fertilizante* * *Varietad*:

```
model.tables(res.aov, type="means", se=TRUE) #medias
```

```
Tables of means
```

```
Grand mean
```

```
62.875
```

```
Fertilizant
```

```
  F1  F2  F3
```

```
36.75 61.81 90.06
```

```
Varietat
```

```
  A  B  C  D
```

```
60.58 64.50 60.08 66.33
```

```
Fertilizant:Varietat
```

```
      A      B      C      D
```

```
F1 29.75 41.50 28.00 47.75
```

```
F2 57.75 61.00 62.50 66.00
```

```
F3 94.25 91.00 89.75 85.25
```

```
Standard errors for differences of means
```

```
  Fertilizant  Varietat  Fertilizant:Varietat
```

```
    2.484    2.868    4.967
```

```
replic.          16          12          4
```

Estimación de los parámetros del modelo

Estimaciones de los coeficientes (efectos) del model lineal

Produccion ~ *Fertilizante* * *Varietat*:

```
model.tables(res.aov, type="effects", se=TRUE) #Efectos (alfa,  
beta, gamma) #efectos del modelo
```

Tables of effects

Fertilizant

| F1 | F2 | F3 |
|---------|--------|--------|
| -26.125 | -1.063 | 27.187 |

Varietat

| A | B | C | D |
|--------|-------|--------|-------|
| -2.292 | 1.625 | -2.792 | 3.458 |

Fertilizante:Varietat

| | A | B | C | D |
|----|--------|--------|--------|--------|
| F1 | -4.708 | 3.125 | -5.958 | 7.542 |
| F2 | -1.771 | -2.437 | 3.479 | 0.729 |
| F3 | 6.479 | -0.687 | 2.479 | -8.271 |

Standard errors of effects

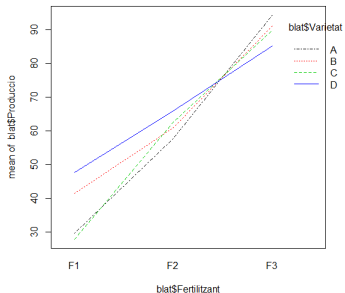
| | Fertilizant | Varietat | Fertilizante:Varietat |
|---------|-------------|----------|-----------------------|
| replic. | 1.756 | 2.028 | 3.512 |
| | 16 | 12 | 4 |

Interacción: Interaction plot

Gráfico de interacción entre 'Variedad' y 'Producción'.

```
interaction.plot(blat$Fertilizante, blat$Variedad, blat$Produccion,  
col=1:length(levels(blat$Variedad)))
```

Gráfico "Interaction plot":



El carácter no paralelo de las líneas sugiere interacción (que se confirma con la prueba ANOVA) ▶

Comparaciones múltiples: condición experimental óptima

Parece que la condición más favorable (condición óptima) a la producción de trigo es la combinación del fertilizante 3 con la variedad A. De todas formas habría que determinar si las diferencias que observamos entre condiciones experimentales son significativas o podrían ser debidas meramente al azar. Esta es una cuestión de "comparaciones múltiples".

Cuando hay interacción, hacer comparaciones múltiples marginalmente por cada factor no tiene mucho sentido. Por ejemplo se pueden hacer comparaciones entre fertilizantes por cada variedad:

```
# Código R
library(agricolae)
blatA.aov <- aov(Produccion ~ Fertilizante, data = blat, subset = blat$Variedad == "A")
comp.mult <- TukeyHSD(blatA.aov, which="Fertilizant")
comp.mult
```

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

```
Fit: aov(formula = Produccion ~ Fertilizante, data = blat, subset = blat$Variedad == "A")
```

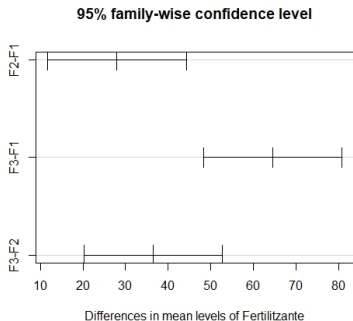
```
$Fertilizante
      diff      lwr      upr      p adj
F2-F1 28.0 11.7699 44.2301 0.0024527
F3-F1 64.5 48.2699 80.7301 0.0000040
F3-F2 36.5 20.2699 52.7301 0.0003799
```

Como puede verse para la variedad A todos los fertilizantes presentan diferencias estadísticamente significativas, siendo su combinación con F3 el que presenta la máxima producción.

Comparaciones múltiples: condición experimental óptima

También pueden presentarse los resultados anteriores (comparaciones múltiples) mediante intervalos de confianza de las diferencias:

```
# Código R
library(agricolae)
blatA.aov <- aov(Produccion ~ Fertilizante, data = blat, subset = blat$Variedad == "A")
comp.mult <- TukeyHSD(blatA.aov, which="Fertilizant")
#comp.mult
plot(comp.mult)
```



Comparaciones múltiples: condición experimental óptima

Aunque posiblemente la comparación más interesante, dado que hay interacción significativa, es entre las celdas (combinaciones de los niveles de cada factor):

```
# Código R
library(agricolae)
inter.aov = aov(Produccion ~ Fertilizante:Variedad, data = blat)
model.tables(inter.aov, type="means", cterms = "Fertilizante:Variedad", se=TRUE)
comp.mult <- TukeyHSD(inter.aov)
comp.mult
```

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

```
Fit: aov(formula = Produccion ~ Fertilizante:Variedad, data = blat)
```

```

$`Fertilizante:Variedad`
      diff      lwr      upr      p adj
F2:A-F1:A 28.00 10.6626868 45.337313 0.0001207
F3:A-F1:A 64.50 47.1626868 81.837313 0.0000000
F1:B-F1:A 11.75 -5.5873132 29.087313 0.4528652
F2:B-F1:A 31.25 13.9126868 48.587313 0.0000168
F3:B-F1:A 61.25 43.9126868 78.587313 0.0000000
F1:C-F1:A -1.75 -19.0873132 15.587313 0.9999999
F2:C-F1:A 32.75 15.4126868 50.087313 0.0000067
F3:C-F1:A 60.00 42.6626868 77.337313 0.0000000
F1:D-F1:A 18.00  0.6626868 35.337313 0.0360457
...
```

Como hay tantas combinaciones se recomienda realizar un análisis de homogeneidad de grupos. ▶

Temas adicionales en varios factores con interacción

Pensemos por un momento en los experimentos que realizamos, cuando combinamos, por ejemplo, dos factores experimentales:

- ▶ ¿Han de tener más de una réplica por cada condición experimental?
- ▶ ¿Que consecuencias tiene que una combinación de factores sólo tenga una réplica?
- ▶ ¿Qué es el test de Tuckey de aditividad?

Caso de una réplica por casilla en la combinación de factores

La importancia de las interacciones:

Alerta!

Si $n = 1$, la suma de cuadrados SS_E té 0 g.d.ll. y σ^2 no es estimable a no ser que supongamos que no hay interacción.

- ▶ Si no hay interacción, utilizaremos $SS_E = SS_T - (SS_A + SS_B)$ con $(a - 1)(b - 1)$ g.d.ll. y no podemos separar el residuo de las posibles interacciones.
- ▶ El estadístico

$$F = \frac{MS_A}{SS_E / ((a - 1)(b - 1))}$$

tiene distribución $F(a - 1, (a - 1)(b - 1))$ y permite probar la significación de A (y de forma similar para B).

Significación de la interacción

Cuando $n = 1$ y sólo disponemos de una réplica por casilla, contrastar la significación de la interacción es un problema difícil.

Prueba de Tukey de aditividad: Tukey's test of additivity

En un modelo restrictivo donde las interacciones són $\gamma_{ij} = \gamma\alpha_i\beta_j$, si $H_0 : \gamma = 0$ es cierta,

$$F = \frac{SS_N}{SS_{\text{Error}} / [(a-1)(b-1) - 1]} \sim F(1, (a-1)(b-1) - 1)$$

donde

$$SS_N = \frac{\left[\sum_{i=1}^a \sum_{j=1}^b Y_{ij} Y_{i.} Y_{.j} - Y_{..} \left(SS_A + SS_B + \frac{Y_{..}^2}{ab} \right) \right]^2}{abSS_ASS_B}$$

$$SS_{\text{Error}} = SS_T - (SS_A + SS_B) - SS_N$$

Ejemplo de la aplicación del test de Tuckey de aditividad

Veamos como aplicar con R el test de aditividad de Tuckey en el caso de querer realizar un contraste de la interacción con una única réplica.

```
# R
# An experiment with one factor in blocks (one replication)
treatment <- as.factor(c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5))
block <- as.factor(c(rep(1,5),rep(2,5),rep(3,5)))
response <- c(7.62,8.14,7.76,7.17,7.46,8,8.15,7.73,7.57,7.68, 7.93,
7.87,7.74, 7.8,7.21)
```

```
> summary(aov(response ~ treatment + block)) #ANOVA
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|----------|
| treatment | 4 | 0.7324 | 0.18311 | 4.192 | 0.0404 * |
| block | 2 | 0.0971 | 0.04856 | 1.112 | 0.3750 |
| Residuals | 8 | 0.3495 | 0.04369 | | |

```
#Tukey's test of additivity to detect interaction with one replication
library(asbio)
tukey.add.test(strength, treatment, block)
```

```
Tukey's one df test for additivity
F = 0.2962553   Denom df = 7   p-value = 0.6031373
```

No se detecta interacción en este caso ($p > 0.05$), aunque no podemos rechazar que si exista (H_0)

Los diseños experimentales: ¿son completamente aleatorizados?

A menudo no es posible aleatorizar totalmente, queremos controlar factores adicionales no directamente interesantes o tenemos restricciones experimentales.

Veamos un ejemplo

El diseño del ejemplo Fertilizante * Variedad no es, en realidad, totalmente aleatorizado:

En el artículo original indica:

- ▶ Se considera un área de siembra muy grande que se divide en 12 zonas igual de grandes.
- ▶ Las 12 combinaciones de fertilizante y variedad se asignan al azar en las zonas.
- ▶ Para medir el error experimental, cada zona se divide en cuatro subzonas que reciben todas el mismo tratamiento.

Un diseño no completamente aleatorizado

Fijémonos que no es totalmente aleatorizado. Cada zona i, j es un bloque que puede tener su efecto, descrito por un parámetro δ_{ij} . Así, un modelo más realista (pero más complicado) sería:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta_{ij} + \epsilon_{ijk}$$

¡Alerta: problemas!

Al depender de los mismos índices, δ_{ij} no se puede estimar separadamente de la interacción. Si δ_{ij} no es constante y nulo (cosa que no podemos probar) tenemos una fuente de sesgo y / o variabilidad no medible, confundida con la interacción.

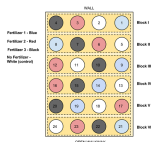
DISEÑOS CON BLOQUES FACTORIALES

Toni Monleón-Getino

Junio de 2020

Bloques en diseños factoriales

- ▶ Con un diseño aleatorio de bloques, el experimentador divide los sujetos en subgrupos llamados bloques, por lo que la variabilidad dentro de los bloques es inferior a la variabilidad entre bloques.
- ▶ A continuación, los sujetos dentro de cada bloque se asignan aleatoriamente a condiciones de tratamiento.
- ▶ En comparación con un diseño completamente aleatorio, este diseño reduce la variabilidad dentro de las condiciones de tratamiento y las posibles confusiones, produciendo una **mejor estimación de los efectos del tratamiento**.
- ▶ En un diseño aleatorio de bloques, sólo hay un factor principal a considerar en el experimento.
- ▶ Cada bloque se prueba con todos los niveles de tratamiento del factor primario en orden aleatorio.
- ▶ Con este diseño se pretende eliminar posibles influencias de otros factores externos.



Ejemplo del diseño: bioremediación de los ríos

- ▶ Tres productos obtenidos a partir de bacterias permiten la biorremediación de los ríos después de un vertido. La bioremediación se mide en % de eficacia.
- ▶ Para saber si tienen la misma eficacia (porcentaje 0-100%), 6 ríos más o menos similares se escogen aleatoriamente para participar en el estudio.
- ▶ De acuerdo con el diseño de bloques aleatorios, en cada ríos se realizará pruebas de eficacia de los tres nuevos productos.
- ▶ Además, en un río se testará sólo un producto a la semana y se tardan tres semanas en probar todos los productos.
- ▶ El orden de prueba de los productos para cada río también asigna aleatoriamente.

Al nivel de significación .05, comprueba si la eficacia media de los 3 nuevos productos son iguales.

#DATOS:

```
Accuracy <- c(31, 27, 24, 31, 28, 31, 45, 29, 46, 21, 18,
48 , 42, 36, 46, 32, 17, 40)
Product <- c(1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2,
3, 1, 2, 3)
blk <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6)
experiment <- data.frame(Accuracy,
Product=as.factor(Product), blk=as.factor(blk))
```

Estructura de los datos

- ▶ Diseño no balanceado de dos factores A y B (donde B es bloque), con a y b niveles respectivamente,

$$\begin{array}{cccc} & B_1 & \dots & B_b \\ A_1 & y_{111}, \dots, y_{11n_{11}} & \dots & y_{1b1}, \dots, y_{1bn_{1b}} \\ \vdots & \vdots & & \vdots \\ A_a & y_{a11}, \dots, y_{a1n_{a1}} & \dots & y_{ab1}, \dots, y_{abn_{ab}} \end{array}$$

- ▶ Si es balanceado,

$$n_{ij} = n, \text{ para todo } i = 1, \dots, a \text{ i } j = 1, \dots, b$$

Modelo lineal

Modelo lineal asociado a un factor principal fijo (A) y un factor bloque (B)

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

per a $i = 1, \dots, a$, $j = 1, \dots, b$ i $k = 1, \dots, n_{ij}$

Con las restricciones

$$\sum_{i=1}^a \alpha_i = 0 \quad \sum_{j=1}^b \beta_j = 0$$

y las condiciones

$$E(\epsilon_{ijk}) = 0 \quad \text{var}(\epsilon_{ijk}) = \sigma^2 \quad \text{para todo } i, j, k$$

Bloques en diseños factoriales: ejemplo biorremediación

```
#DATOS:
```

```
Accuracy <- c(31, 27, 24, 31, 28, 31, 45, 29, 46, 21, 18,  
48, 42, 36, 46, 32, 17, 40)
```

```
Product <- c(1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2,  
3, 1, 2, 3)
```

```
blk <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6)
```

```
experiment <- data.frame(Accuracy,
```

```
Product=as.factor(Product), blk=as.factor(blk))
```

```
#ANOVA
```

```
summary(aov(response ~ treatment + block))
```

```
#ANOVA
```

```
> av = aov(Accuracy ~ Product + blk, data=experiment)
```

```
> summary(av)
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|----------|
| Product | 2 | 538.8 | 269.39 | 4.959 | 0.0319 * |
| blk | 5 | 559.8 | 111.96 | 2.061 | 0.1547 |
| Residuals | 10 | 543.2 | 54.32 | | |

Como puede ser significativo el factor "Product" pero no el factor bloque (blk) por lo que se descarta su influencia en el resultado de la eficacia del bioremediador. Si que existe efecto del producto sobre la biorremediación ($p=0.319$).

Uso de bloques en diseños multifactoriales

En el experimento anterior (fertilizante y variedad), un diseño también con bloques, más adecuado, sería:

- ▶ Se considera un área de siembra muy grande que se divide en 4 zonas igual de grandes.
- ▶ Cada zona se divide en 12 subzonas. Por cada una de las 4 zonas, los 12 tratamientos se asignan al azar a las 12 subzonas.

Modelo con bloques

Cada una de las 4 "réplicas" se asocia a un "bloque zona". El modelo es ahora:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta_k + \epsilon_{ijk}$$

Las interacciones con el factor bloque se deben suponer inexistentes o confundidas con el error (1 sola réplica), pero el efecto principal δ_k es analizable.

DISEÑOS MULTIFACTORIALES CON UNO O MÁS FACTORES ALEATORIOS

Toni Monleón-Getino

Junio de 2020

Experimentos con factores aleatorios

Supongamos que A y B son factores aleatorios, es decir, sus niveles son muestras aleatorias de tamaño a y b , respectivamente, de poblaciones más grandes. [Apuntes factores aleatorios](#)

Modelo lineal asociado

$$Y_{ijk} = \mu + A_i + B_j + I_{ij} + \epsilon_{ijk}$$

donde $i = 1, \dots, a$, $j = 1, \dots, b$ i $k = 1, \dots, n$.

$$A_i \sim N(0, \sigma_A) \quad B_j \sim N(0, \sigma_B) \quad I_{ij} \sim N(0, \sigma_{AB})$$

$$\epsilon_{ijk} \sim N(0, \sigma)$$

con A_i , B_j , I_{ij} y ϵ_{ijk} v.a. independientes.

Componentes de la varianza

La independencia de las v.a. los factores y del residuo hace que la varianza de las observaciones se descomponga en las **componentes de la varianza**:

$$\text{var}(Y_{ijk}) = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma^2$$

Por otra parte existe dependencia entre las observaciones:

$$\text{cov}(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 & \text{si } i = i', j = j', k \neq k' \\ \sigma_A^2 & \text{si } i = i', j \neq j' \\ \sigma_B^2 & \text{si } i \neq i', j = j' \\ 0 & \text{si } i \neq i', j \neq j' \end{cases}$$

Significación de los factores y la interacción

Ahora los contrastes de mayor interés son:

$$\begin{array}{lll} H_0 : \sigma_A^2 = 0 & H_0 : \sigma_B^2 = 0 & H_0 : \sigma_{AB}^2 = 0 \\ H_1 : \sigma_A^2 \neq 0 & H_1 : \sigma_B^2 \neq 0 & H_1 : \sigma_{AB}^2 \neq 0 \end{array}$$

Las sumas de cuadrados y los cuadrados medios son los mismos que los del modelo de efectos fijos, pero:

$$\begin{array}{ll} E(MS_A) = \sigma^2 + n\sigma_{AB}^2 + bn\sigma_A^2 & E(MS_B) = \sigma^2 + n\sigma_{AB}^2 + an\sigma_B^2 \\ E(MS_{AB}) = \sigma^2 + n\sigma_{AB}^2 & E(MS_E) = \sigma^2 \end{array}$$

Y los respectivos estadísticos F adecuados son:

$$F = \frac{MS_A}{MS_{AB}} \sim F(a-1, (a-1)(b-1)) \quad F = \frac{MS_B}{MS_{AB}} \sim F(b-1, (a-1)(b-1))$$

$$F = \frac{MS_{AB}}{MS_E} \sim F((a-1)(b-1), ab(n-1))$$

Ejemplo de ANOVA varios factores aleatorios

Análisis de diferentes Interleucinas en granjas animales

- ▶ Las interleucinas o citocinas son proteínas de bajo peso molecular esenciales en la intercomunicación celular.
- ▶ En un experimento realizado sobre animales de granja, se quería estudiar el efecto de tres soluciones proteínicas (S1, S2 y S3) cogidas al azar, de entre muchas otras posibles, sobre la respuesta inmunitaria.
- ▶ La respuesta a estos tratamientos se midió en forma de un recuento de células precursoras de los linfocitos (variable "Compt.prec").
- ▶ Se sospechaba que la respuesta podía variar según la procedencia de los animales, por lo que se seleccionaron al azar 3 granjas, G1, G2 y G3, de entre muchas posibles.
- ▶ De cada granja se tomó una muestra aleatoria de 9 animales (por lo tanto se trabajó con una muestra total de 27 animales), que se asignaron al azar y de forma balanceada a cada tratamiento.

¿Es significativo el factor "solución proteínica"? Y el factor " granja"?
¿Hay interacción?

Tabla de datos: ejemplo interleucinas (2 factores mixtos y cruzados)

Tabla y datos

A= solución proteixa, B= Granja

| A/B | G1 | | | G2 | | | G3 | | |
|-----|----|----|----|----|----|----|----|----|----|
| S1 | 45 | 47 | 39 | 47 | 51 | 39 | 40 | 42 | 35 |
| S2 | 52 | 52 | 53 | 54 | 49 | 49 | 48 | 51 | 52 |
| S3 | 66 | 64 | 62 | 61 | 58 | 57 | 56 | 59 | 59 |

Observad que el diseño contiene dos factores cruzados (interacción), con 3 réplicas por condición experimental.

Tabla de datos: ejemplo interleucinas

#Dades en R:

```
Solucio<- c(rep(1,9),rep(2,9),rep(3,9))  
Granja<- c(rep(1,3), rep(2,3), rep(3,3),  
rep(1,3), rep(2,3), rep(3,3), rep(1,3), rep(2,3),  
rep(3,3))  
Compt.prec <- c(45, 47, 39, 47, 51, 39, 40, 42, 35, 52, 52,  
53, 54, 49, 49, 48, 51, 52, 66, 64, 62, 61, 58, 57, 56, 59, 59)
```

#DATOS:

```
interleucines<- data.frame(Solucio=as.factor(Solucio),  
Granja=as.factor(Granja), Compt.prec)
```

#ANOVA:

```
interleucines.aov <- aov(Compt.prec ~ Solucio * Granja,  
data = interleucines)
```

Para poder hacer este ejercicio con R de manera eficiente, es necesario utilizar la librería AnovaTools, desarrollada por el profesor Jordi Ocaña y que encontrareis en: [ANOVATOOLS](#)

Ejemplo de interleucinas

Los cálculos de la tabla ANOVA deben realizarse correctamente para poder calcular adecuadamente los valores F y p de los diferentes factores (A, B, A*B), de acuerdo con el diseño y su naturaleza (fija o aleatoria). En este caso los factores "granja" y "solución", así como la interacción son considerados aleatorios.

```
#Análisis mediante la funcion anova2cross()
library(AnovaTools)
taula.anova <- anova2cross(interleucines.aov,
random = c("Solucio","Granja"))
taula.anova
```

Analysis of Variance Table

Response: Compt.prec

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|----|---------|---------|---------|-------------|
| Solucio | 2 | 1370.30 | 685.15 | 45.2298 | 0.001793 ** |
| Granja | 2 | 81.41 | 40.70 | 2.6870 | 0.182080 |
| Solucio:Granja | 4 | 60.59 | 15.15 | 1.4819 | 0.249180 |
| Residuals | 18 | 184.00 | 10.22 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Puede verse que solo es significativo el factor "solución" ($p < 0.05$), por lo que este aporta varianza a la variable de respuesta.

Estimación de las componentes de la varianza

Estimadores puntuales:

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_{AB}}{bn} \quad \hat{\sigma}_B^2 = \frac{MS_B - MS_{AB}}{an}$$
$$\hat{\sigma}_{AB}^2 = \frac{MS_{AB} - MS_E}{n} \quad \hat{\sigma}^2 = MS_E$$

En el ejemplo anterior (Interleucinas en granjas)

$$\hat{\sigma}_A^2 = \hat{\sigma}_{\text{solucio}}^2 = \frac{685.15 - 15.15}{3 \times 3} = 74.4 \quad \hat{\sigma}_B^2 = \hat{\sigma}_{\text{granja}}^2 = 2.83$$
$$\hat{\sigma}_{AB}^2 = \hat{\sigma}_{\text{solucio,granja}}^2 = 1.64 \quad \hat{\sigma}^2 = 10.22$$

y la covarianza entre animales de la misma solución y granja:

$$\hat{\sigma}_{ijk,ijk'} = \hat{\sigma}_{\text{solucio}}^2 + \hat{\sigma}_{\text{granja}}^2 + \hat{\sigma}_{\text{solucio,granja}}^2 = 78.87$$

Ejemplo Solucion * Granja

Cálculos con R de las componentes de la varianza: Utilizar la función lmer de R (library(lme4)) para calcular las componentes de la varianza

```
library(lme4)
res.lmer<-lmer(Compt.prec~(1|Solucio)+(1|Granja)+(1|Solucio:Granja),
data=interleucines)
summary(res.lmer)
confint(res.lmer)
```

```
> summary(res.lmer)
Random effects:
  Groups      Name      Variance Std.Dev.
Solucio:Granja (Intercept)  1.647   1.283
Granja        (Intercept)  2.838   1.685
Solucio       (Intercept) 74.437   8.628
Residual                               10.220   3.197
> confint(res.lmer)
Computing profile confidence intervals ...
                2.5 %   97.5 %
.sig01      0.000000  4.897943
.sig02      0.000000  7.395252
.sig03      3.535681 21.181271
.sigma      2.381301  4.569152
(Intercept) 39.699961 63.040971
```

Como puede verse se obtienen los mismos resultados que manualmente, aunque con más precisión y también se pueden calcular los IC95% de las estimaciones de las componentes de la varianza.

Un factor aleatorio y un factor fijo: modelo no restringido (¡lo adoptamos por defecto!)

Modelo lineal asociado

$$Y_{ijk} = \mu + \alpha_i + B_j + I_{ij} + \epsilon_{ijk}$$

per a $i = 1, \dots, a$, $j = 1, \dots, b$ i $k = 1, \dots, n$

Con las restricciones

$$\sum_{i=1}^a \alpha_i = 0$$

$$B_j \sim N(0, \sigma_B) \quad I_{ij} \sim N(0, \sigma_{AB})$$

Toda interacción con un término aleatorio siempre es aleatoria.

Un factor aleatorio y un factor fijo: modelo no restringido

Consideramos ahora que el factor A es fijo y el B es aleatorio.

Contrastes:

$$\begin{array}{lll} H_0 : \alpha_1 = \dots = \alpha_a = 0 & H_0 : \sigma_B^2 = 0 & H_0 : \sigma_{AB}^2 = 0 \\ H_1 : \alpha_i \neq 0 & H_1 : \sigma_B^2 > 0 & H_1 : \sigma_{AB}^2 > 0 \end{array}$$

Esperanzas de los cuadrados medios:

$$\begin{array}{ll} E(MS_A) = \sigma^2 + n\sigma_{AB}^2 + \frac{bn \sum_{i=1}^a \alpha_i^2}{a-1} & E(MS_B) = \sigma^2 + n\sigma_{AB}^2 + an\sigma_B^2 \\ E(MS_{AB}) = \sigma^2 + n\sigma_{AB}^2 & E(MS_E) = \sigma^2 \end{array}$$

Estadísticos F:

$$\begin{array}{ll} F = \frac{MS_A}{MS_{AB}} \sim F(a-1, (a-1)(b-1)) & F = \frac{MS_B}{MS_{AB}} \sim F(b-1, (a-1)(b-1)) \\ F = \frac{MS_{AB}}{MS_E} \sim F((a-1)(b-1), ab(n-1)) & \end{array}$$

Estimación de las componentes de la varianza en el modelo no restringido (ejemplo)

Estimadores puntuales:

$$\hat{\sigma}_B^2 = \frac{MS_B - MS_E - n\hat{\sigma}_{AB}^2}{an} \quad \hat{\sigma}_{AB}^2 = \frac{MS_{AB} - MS_E}{n}$$
$$\hat{\sigma}^2 = MS_E$$

En el ejemplo

$$\hat{\sigma}_B^2 = \hat{\sigma}_{granja}^2 = 2.84$$
$$\hat{\sigma}_{AB}^2 = \hat{\sigma}_{solucio,granja}^2 = 1.64 \quad \hat{\sigma}^2 = 10.22$$

y la covarianza entre animales de la misma solución y granja:

$$\hat{\sigma}_{ijk,ijk'} = \hat{\sigma}_{granja}^2 + \hat{\sigma}_{solucio,granja}^2 = 4.48$$

Ejemplo Solucion * Granja: factores mixtos

Cálculos con R de las componentes de la varianza en el caso de A fijo y B aleatorio (modelo no restringido): Utilizar la función lmer de R (library(lme4)) para calcular las componentes de la varianza

```
library(lme4)
res.lmer<-lmer(Compt.prec~Solucio+(1|Granja)+(1|Solucio:Granja),
data=interleucines)
summary(res.lmer)
confint(res.lmer)
```

```
> summary(res.lmer)
Random effects:
Groups          Name          Variance Std.Dev.
Solucio:Granja (Intercept)  1.642   1.282
Granja          (Intercept)  2.840   1.685
Residual                               10.222  3.197
> confint(res.lmer)
Computing profile confidence intervals ...
              2.5 %    97.5 %
.sig01      0.000000  3.129234
.sig02      0.000000  5.016098
.sigma      2.465064  4.361990
(Intercept) 39.683085 45.872503
Solucio2     5.268712 11.397952
Solucio3    14.379823 20.509063
```

Como puede verse se obtienen los mismos resultados que manualmente, aunque con más precisión y también se pueden calcular los IC95% de las estimaciones de las componentes de la varianza.

Modelo restringido

Modelo lineal asociado

$$Y_{ijk} = \mu + \alpha_i + B_j + I_{ij} + \epsilon_{ijk}$$

per a $i = 1, \dots, a$, $j = 1, \dots, b$ i $k = 1, \dots, n$

Con las restricciones

$$\sum_{i=1}^a \alpha_i = 0 \quad B_j \sim N(0, \sigma_B)$$

$$I_{ij} \sim N\left(0, \sqrt{\frac{a-1}{a}} \sigma_{AB}\right) \quad \sum_{i=1}^a I_{ij} = 0, \quad j = 1, \dots, b \quad (1)$$

Toda interacción con un término aleatorio siempre es aleatoria.

(1) hace que algunas expresiones sean más sencillas; debido a (1) el modelo se conoce con *modelo restringido*.

Modelo restringido

Contrastes:

$$\begin{array}{lll} H_0 : \alpha_1 = \dots = \alpha_a = 0 & H_0 : \sigma_B^2 = 0 & H_0 : \sigma_{AB}^2 = 0 \\ H_1 : \alpha_i \neq 0 & H_1 : \sigma_B^2 > 0 & H_1 : \sigma_{AB}^2 > 0 \end{array}$$

Esperanzas de los cuadrados medios:

$$\begin{array}{ll} E(MS_A) = \sigma^2 + n\sigma_{AB}^2 + \frac{bn \sum_{i=1}^a \alpha_i^2}{a-1} & E(MS_B) = \sigma^2 + an\sigma_B^2 \\ E(MS_{AB}) = \sigma^2 + n\sigma_{AB}^2 & E(MS_E) = \sigma^2 \end{array}$$

Estadísticos F :

$$\begin{array}{ll} F = \frac{MS_A}{MS_{AB}} \sim F(a-1, (a-1)(b-1)) & F = \frac{MS_B}{MS_E} \sim F(b-1, ab(n-1)) \\ F = \frac{MS_{AB}}{MS_E} \sim F((a-1)(b-1), ab(n-1)) & \end{array}$$

Modelo restringido vs modelo mixto sin restricciones

El quocient F correcto sobre el efecto del **factor aleatorio** (aquí designado como B) es:

- ▶ En el modelo no restringido: dividir por MS_{AB} .
- ▶ En el modelo restringido: dividir por MS_E .

La decisión puede cambiar y se puede rechazar H_0 en uno y no al otro.

¿Qué modelo es más correcto?

Una consecuencia de las restricciones impuestas sobre las interacciones es que $\text{cov}(I_{ij}, I_{i'j}) = -\sigma_{AB}^2/a$, de forma que las interacciones para el mismo nivel aleatorio están correlacionadas y no son independientes. De modo que la principal razón para considerar el modelo restringido es la observación de correlación entre las interacciones. De lo contrario es preferible el modelo no restringido.

Además, el modelo restringido es problemático sin balanceo.

Un último ejemplo de dos factores mixtos

Supondremos un experimento donde se obtiene un modelo lineal de 2 factores (A = fijo, B = aleatorio) y cruzado (interacción, aleatoria)

- ▶ Asumimos que el factor A es fijo y que el factor B es aleatorio, el orden en que especificamos estos dos factores es crucial en la fórmula siguiente.
- ▶ El modelo es el restringido, es decir, las interacciones suman 0 respecto al índice del factor fijo
- ▶ Modelo balanceado

```
#simulamos unos datos, 2 factores y una variable de respuesta Y=resp
factorA <- c(rep(1,4),rep(2,4),rep(3,4),rep(4,4))
factorB <- c(rep(1,2),rep(2,2),rep(1,2),rep(2,2))
Data.set<- data.frame(factorA=factorA, factorB=factorB,
  resp=c(rnorm(2,33,3),rnorm(2,28,3), >rnorm(2,14,2),rnorm(2,10,2),
  rnorm(2,18,3), rnorm(2,12,3), rnorm(2,25,3), rnorm(2,20,3)))

# nombre de las columnas
colnames(Data.set) <- c('factorA', 'factorB', 'resp')
```

Un último ejemplo de dos factores mixtos

Calcular la tabla ANOVA correcta

```
#ANOVA: generamos el objeto data.aov, un ANOVA 2F completo
data.aov <- aov(resp ~ factorA*factorB, Data.set)
# Transformaciones para obtener la tabla ANOVA correcta
auxAnova <- summary(data.aov)[[1]] # el segundo factor es el aleatorio
, sino debemos girarlo
auxAnova[1,4] <- auxAnova$Mean[1]/auxAnova$Mean[3] # calcula
la F fent el quocient MSA/MSAB (model >restringit)
auxAnova[1,5] <- 1-pf(auxAnova[1,4], auxAnova$Df[1],auxAnova$Df[3])
```

```
#Calculo de la tabla ANOVA
> auxAnova

          Df Sum Sq Mean Sq F value    Pr(>F)
factorA    3  903.44  301.148  33.9579 0.008143 **
factorB    1  123.43  123.430  10.6217 0.011543 *
factorA:factorB  3   26.60   8.868   0.7632 0.545768
Residuals   8   92.96  11.621
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Es significativo el factor A, el B, pero no la interacción. B es aleatorio por tanto B aporta varianza en la variable de respuesta.

Ejemplo: Estimación de las componentes de la varianza

Utilizar la función lmer es adecuado para calcular las componentes de la varianza en R

```
library(lme4)
res.lmer<-lmer(resp~factorA+(1|factorB)+(1|factorA:factorB),
               data=Data.set)

summary(res.lmer)
confint(res.lmer)
```

#Componentes de la varianza

Random effects:

| Groups | Name | Variance | Std.Dev. |
|-----------------|-------------|-----------|----------|
| factorA:factorB | (Intercept) | 4.573e-04 | 0.02138 |
| factorB | (Intercept) | 1.415e+01 | 3.76167 |
| Residual | | 1.086e+01 | 3.29616 |

Number of obs: 16, groups: factorA:factorB, 8; factorB, 2

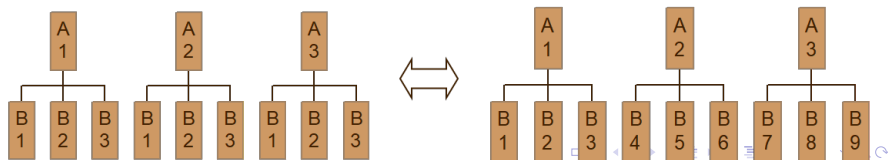
DISEÑOS JERÁRQUICOS

Toni Monleón-Getino

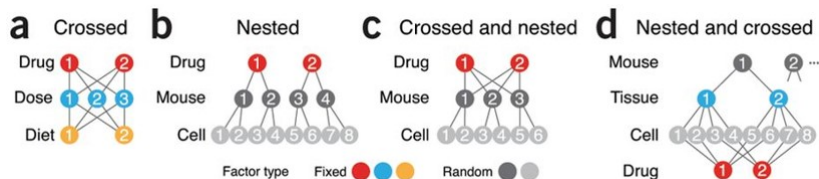
Junio de 2020

Concepto de factor jerárquico

- ▶ Diseño multifactorial al que los niveles de un factor son similares pero no idénticos para cada nivel de otro factor.
- ▶ Criterio: El factor B se sitúa jerárquicamente dentro del factor A si por cada nivel de A podemos cambiar arbitrariamente la numeración de los niveles de B
- ▶ Introducción muy moderna en <http://www.flutterbys.com.au/stats/tut/tut9.2a.html>
- ▶ En ocasiones, también se denominan "diseños anidados"



Tipos de diseños: algunos ejemplos combinados



- ▶ (a) Un diseño cruzado examina cada combinación de niveles para cada factor fijo.
- ▶ (b) El diseño anidado puede subreplicar progresivamente un factor fijo con niveles anidados de un factor aleatorio que son exclusivos del nivel dentro del cual están anidados.
- ▶ (c) Si un factor aleatorio puede reutilizarse para diferentes niveles del tratamiento, puede cruzarse con el tratamiento y modelarse como un bloque.
- ▶ (d) Un diseño de parcelas divididas en el que se cruzan los efectos fijos (tejido, fármaco) (se prueba cada combinación de tejido y fármaco) pero se anidan dentro de las réplicas

(de Nature Methods)

Modelo lineal jerárquico

Modelo de asociado lineal (2 factores fijos y B jerarquizado a A: B (A))

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

per a $i = 1, \dots, a$, $j = 1, \dots, b$ i $k = 1, \dots, n_{ij}$

Con las restricciones

$$\sum_{i=1}^a \alpha_i = 0 \quad \sum_{j=1}^b \beta_{ji} = 0, \quad \text{per a } i = 1, \dots, a$$

y las condiciones

$$E(\epsilon_{ijk}) = 0 \quad \text{var}(\epsilon_{ijk}) = \sigma^2 \quad \text{per a tot } i, j, k$$

Caso totalmente balanceado ($n_i = n$), pero puede haber falta de balanceo a cualquier nivel jerárquico. No hay interacciones, no tienen sentido en no haber cruce entre factores.

Modelo lineal jerárquico

Modelo mixto, A fijo, B aleatorio

$$Y_{ijk} = \mu + \alpha_i + B_{j(i)} + e_{k(ij)}$$

para $i=1, \dots, a$; $j=1, \dots, b$; $k=1, \dots, n$

$$\sum_{i=1}^a \alpha_i = 0; B_{j(i)} \sim N(0, \sigma_B); e_{k(ij)} \sim N(0, \sigma)$$

Modelo aleatorio, A aleatorio, B aleatorio:

$$Y_{ijk} = \mu + A_i + B_{j(i)} + e_{k(ij)}$$

para $i=1, \dots, a$; $j=1, \dots, b$; $k=1, \dots, n$

$$A_i \sim N(0, \sigma_B); B_{j(i)} \sim N(0, \sigma_B); e_{k(ij)} \sim N(0, \sigma)$$

Descomposición de la suma de cuadrados

Común a todos los modelos (factores: Fijos, Mixtos, Aleatorios, ...)

$$SS_T = SS_A + SS_{B(A)} + SS_E,$$

donde,

$$SS_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2 \quad abn - 1 \text{ g.d.l.}$$

$$SS_A = bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 \quad a - 1 \text{ g.d.l.}$$

$$SS_{B(A)} = n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..})^2 \quad a(b - 1) \text{ g.d.l.}$$

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij.})^2 \quad ab(n - 1) \text{ g.d.l.}$$

Esperanzas de los cuadrados medios

| $E(MS)$ | A fijo B fijo | A fijo B aleatorio | A aleatorio B aleatorio |
|---------|------------------|-----------------------|----------------------------|
|---------|------------------|-----------------------|----------------------------|

| | | | |
|-----------|---|---|---|
| $E(MS_A)$ | $\sigma^2 + \frac{bn \sum_{i=1}^a \alpha_i^2}{a-1}$ | $\sigma^2 + n\sigma_B^2 + \frac{bn \sum_{i=1}^a \alpha_i^2}{a-1}$ | $\sigma^2 + n\sigma_B^2 + bn\sigma_A^2$ |
|-----------|---|---|---|

| | | | |
|----------------|--|--------------------------|--------------------------|
| $E(MS_{B(A)})$ | $\sigma^2 + \frac{\sum_{i=1}^a \sum_{j=1}^b \beta_{j(i)}^2}{a(b-1)}$ | $\sigma^2 + n\sigma_B^2$ | $\sigma^2 + n\sigma_B^2$ |
|----------------|--|--------------------------|--------------------------|

| | | | |
|-----------|------------|------------|------------|
| $E(MS_E)$ | σ^2 | σ^2 | σ^2 |
|-----------|------------|------------|------------|

Pruebas F por los diversos contrastes

A fijo
B fijo

A fijo
B aleatorio

A aleatorio
B aleatorio

$$H_0 \quad \alpha_1 = \dots = \alpha_a$$

$$H_0 \quad \alpha_1 = \dots = \alpha_a$$

$$H_0 \quad \sigma_A^2 = 0$$

Análisis del
factor A

$$F = \frac{MS_A}{MS_E}$$
$$F(a-1, ab(n-1))$$

$$F = \frac{MS_A}{MS_{B(A)}}$$
$$F(a-1, a(b-1))$$

$$F = \frac{MS_A}{MS_{B(A)}}$$
$$F(a-1, a(b-1))$$

$$H_0 \quad \beta_1 = \dots = \beta_b$$

$$H_0 \quad \sigma_B^2 = 0$$

$$H_0 \quad \sigma_B^2 = 0$$

Análisis del
factor B

$$F = \frac{MS_{B(A)}}{MS_E}$$
$$F(a(b-1), ab(n-1))$$

$$F = \frac{MS_{B(A)}}{MS_E}$$
$$F(a(b-1), ab(n-1))$$

$$F = \frac{MS_{B(A)}}{MS_E}$$
$$F(a(b-1), ab(n-1))$$

Otros análisis de interés

Estimación puntual de los parámetros:

$$\hat{\mu} = \bar{Y}_{...}, \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_{j(i)} = \bar{Y}_{ij.} - \bar{Y}_{i..}$$

Análisis de los residuos

$$\hat{e}_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - \left(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_{j(i)} \right) = Y_{ijk} - \bar{Y}_{ij.}$$

Estimación de las componentes de la varianza:

$$\hat{\sigma}^2 = MS_E$$

$$\hat{\sigma}_B^2 = \frac{MS_{B(A)} - MS_E}{n}$$

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_{B(A)}}{bn}$$

Modelo de 3 factores fijos

Modelo:

$$Y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + e_{l(ijk)}$$

para $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; l = 1, \dots, n$

$$\sum_{i=1}^a \alpha_i = 0$$

$$\sum_{j=1}^b \beta_{j(i)} = 0; i = 1, \dots, a$$

$$\sum_{k=1}^c \gamma_{k(ij)} = 0; i = 1, \dots, a; j = 1, \dots, b$$

$$e_{l(ijk)} \sim N(0, \sigma) \text{ i.i.d.}$$

Modelo mixto: A y B fijos, C aleatorio

Modelo:

$$Y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + C_{k(ij)} + e_{l(ijk)}$$

para $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; l = 1, \dots, n$

$$\sum_{i=1}^a \alpha_i = 0$$

$$\sum_{j=1}^b \beta_{j(i)} = 0; i = 1, \dots, a$$

$$C_{k(ij)} \sim N(0, \sigma_C) \text{ i.i.d.}$$

$$e_{l(ijk)} \sim (0, \sigma) \text{ i.i.d.}; C_{k(ij)} \text{ y } e_{l(ijk)} \text{ independientes}$$

Modelo mixto: A fijo, B y C aleatorios

Modelo:

$$Y_{ijkl} = \mu + \alpha_i + B_{j(i)} + C_{k(ij)} + e_{l(ijk)}$$

per $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; l = 1, \dots, n :$

$$\sum_{i=1}^a \alpha_i = 0$$

$$B_{j(i)} \sim (0, \sigma_B) \text{ iid}$$

$$C_{k(ij)} \sim (0, \sigma_C) \text{ iid}$$

$$e_{l(ijk)} \sim (0, \sigma) \text{ iid}$$

$B_{j(i)}, C_{k(ij)}$ y $e_{l(ijk)}$ independientes

Modelo con tres factores aleatorios

Modelo:

$$Y_{ijkl} = \mu + A_i + B_{j(i)} + C_{k(ij)} + e_{l(ijk)}$$

per $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; l = 1, \dots, n$:

$$A_i \sim (0, \sigma_A) \text{ iid}$$

$$B_{j(i)} \sim (0, \sigma_B) \text{ iid}$$

$$C_{k(ij)} \sim (0, \sigma_C) \text{ iid}$$

$$e_{l(ijk)} \sim (0, \sigma) \text{ iid}$$

$A_i, B_{j(i)}, C_{k(ij)}$ i $e_{l(ijk)}$ independents

Descomposición de la suma de cuadrados (común a todos los diseños anteriores de 3 factores jerarquizados)

| Fuente de variación | Suma de cuadrados | Grados de libertad | Cuadrados medios |
|---------------------|--|--------------------|---|
| A | $SS_A = bcn \sum_i (\bar{Y}_{i\dots} - \bar{Y}_{\dots})^2$ | $a - 1$ | $MS_A = \frac{SS_A}{a-1}$ |
| B | $SS_{B(A)} = cn \sum_i \sum_j (\bar{Y}_{ij\dots} - \bar{Y}_{i\dots})^2$ | $a(b - 1)$ | $MS_{B(A)} = \frac{SS_{B(A)}}{a(b-1)}$ |
| C | $SS_{C(AB)} = n \sum_i \sum_j \sum_k (\bar{Y}_{ijk\dots} - \bar{Y}_{ij\dots})^2$ | $ab(c - 1)$ | $MS_{C(BA)} = \frac{SS_{C(AB)}}{ab(c-1)}$ |
| Error | $SS_E = \sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y}_{ijk\dots})^2$ | $abc(n - 1)$ | $MS_E = \frac{SS_E}{abc(n-1)}$ |
| Total | $SS_T = \sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y}_{\dots})^2$ | $abcn - 1$ | |

Modelo con tres factores fijos. Esperanzas de los cuadrados medios

$$E(MS_A) = \sigma^2 + \frac{bcn \sum_i \alpha_i^2}{a-1}$$

$$E(MS_{B(A)}) = \sigma^2 + \frac{cn \sum_i \sum_j \beta_{j(i)}^2}{a(b-1)}$$

$$E(MS_{C(BA)}) = \sigma^2 + \frac{n \sum_i \sum_j \sum_k \gamma_{k(ij)}^2}{ab(c-1)}$$

$$E(MS_E) = \sigma^2$$

Modelo mixto: A y B fijos, C aleatorio. Esperanzas de los cuadrados medios

$$E(MS_A) = \sigma^2 + n\sigma_C^2 + \frac{bcn \sum_i \alpha_i^2}{a-1}$$

$$E(MS_{B(A)}) = \sigma^2 + n\sigma_C^2 + \frac{cn \sum_i \sum_j \beta_{j(i)}^2}{a(b-1)}$$

$$E(MS_{C(BA)}) = \sigma^2 + n\sigma_C^2$$

$$E(MS_E) = \sigma^2$$

Modelo mixto: A fijo, B y C aleatorios. Esperanzas de los cuadrados medios

$$E(MS_A) = \sigma^2 + n\sigma_C^2 + cn\sigma_B^2 + \frac{bcn \sum_i \alpha_i^2}{a-1}$$

$$E(MS_{B(A)}) = \sigma^2 + n\sigma_C^2 + cn\sigma_B^2$$

$$E(MS_{C(BA)}) = \sigma^2 + n\sigma_C^2$$

$$E(MS_E) = \sigma^2$$

Modelo con tres factores aleatorios. Esperanzas de los cuadrados medios

$$E(MS_A) = \sigma^2 + n\sigma_C^2 + cn\sigma_B^2 + bc n\sigma_A^2$$

$$E(MS_{B(A)}) = \sigma^2 + n\sigma_C^2 + cn\sigma_B^2$$

$$E(MS_{C(BA)}) = \sigma^2 + n\sigma_C^2$$

$$E(MS_E) = \sigma^2$$

Ejemplo de diseño jerárquico: Salinidad en los pozos



Estamos realizando un estudio sobre la salinidad del agua de unos pozos en una determinada zona agrícola cercana al mar (zona mediterránea), donde se ha observado un efecto de fitotoxicidad en las plantas regadas. Vamos a medir la salinidad de unos pozos elegidos al azar dos veces cada día. Por motivos logísticos y de personal los pozos muestreados son diferentes en cada una de las horas.

- Especificad el modelo y su parametrización
- resolved el modelo e indicad las conclusiones a las que llegáis.
- Estimad todos sus parámetros

Datos del problema propuesto

Cálculos con R: ¿Cómo introducir los datos?

```
#DATOS:
```

```
Hora<- c("7h","7h","7h","7h","7h","7h","21h","21h","21h","21h","21h","21h")
```

```
Pozo <- c("Pou1","Pou1","Pou2","Pou2","Pou3","Pou3","Pou1","Pou1","Pou2","Pou2","Pou3","Pou3")
```

```
Salinidad<-c(7.17,7.19,7.33,7.45,7.56,7.52,4.65,4.69,4.71,4.78,4.73,4.87)
```

```
#DATA-FRAME:
```

```
pous <-data.frame(Hora, Pozo, Salinidad)
```

```
#TIPO DE FACTOR:
```

```
class(pous$Hora)
```

```
class(pous$Pozo)
```

Ejemplo Pozos y salinidad

Realizad los cálculos del ANOVA mediante R:

```
#) Modelo con 2 factores: 1 fijo, 1 aleatorio, jerrquico o anidado

#imaginamos que son tratados como factores fijos incorrectamente y calculemos ANOVA:
pous1.aov <- aov(Salinidad ~ Hora/Pozo, data=pous)
# La tabla ANOVA no es correcta !!!!!!!!!!!!!!!!.
#anova(pous1.aov) # factor hora p-valor = 3.334e-10; factor Pou(hora) p-valor =0.007192

# Tabla ANOVA correcta dado que el primer factor es fijo y el otro Aleatorio.

# Utilizad la funcion 'anova2nest' de la libreria ANOVATOOLS
library(AnovaTools) #https://github.com/amonleong/Statistical_tools
# Tabla ANOVA correcta:
anova2nest(pous1.aov, random = c("Pozo"))
```

Analysis of Variance Table

Response: Salinidad

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|---------|---------|---------|---------------|
| Hora | 1 | 20.7770 | 20.7770 | 562.174 | 1.876e-05 *** |
| Hora:Pozo | 4 | 0.1478 | 0.0370 | 10.435 | 0.007192 ** |
| Residuals | 6 | 0.0212 | 0.0035 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ejemplo Pozos y salinidad

Modelo con 2 factores: 1 fijo, 1 aleatorio, jerárquico o anidado
Cálculos con R

```
# Ahora calculemos la estimacion de las componentes de la varianza del modelo ANOVA
# En futuros calculos, nos resultara mas comodo indicar como a, b y n
# el nmero de niveles de cada factor y el tamaño muestral en cada casilla:
a <- length(levels(pous[,"Hora"])) # niveles de hora
b <- length(levels(pous[,"Pozo"])) # niveles de pozo
# Solament vliid per aquest cas balancejat:
n <- nrow(pous) %/% (a*b) # son las replicas
a
b
n
#ANOVA
anovaTab <- anova2nest(pous1.aov, random = c("Pou"))
anovaTab
# Sumas de cuadrados medios
ms.pou_aniuat_hora <- anovaTab[2,3]
ms.resid <- anovaTab[3,3]

# Componentes de la varianza (factor POZO anidado a HORA):
var.pou_aniuat_hora <- (ms.pou_aniuat_hora - ms.resid) / n
var.pou_aniuat_hora #componente de la varianza de POZO anidado a HORA
var.resid <- ms.resid
var.resid #componente de la varianza de RESIDUO
```

```
> var.pou_aniuat_hora #componente de la varianza de POZO ANIDADO EN HORA
[1] 0.01670833
> var.resid <- ms.resid
> var.resid #componente de la varianza de RESIDUO
[1] 0.003541667
```

Pozos y salinidad

Cálculos con R

```
# Calculo manual de las componentes de la varianza en forma porcentual
var.total <- var.pou_aniuat_hora + var.resid
round(100 * var.pou_aniuat_hora / var.total, 2)
round(100 * var.resid / var.total, 2)
```

```
> var.total <- var.pou_aniuat_hora + var.resid #100%
> round(100 * var.pou_aniuat_hora / var.total, 2)
[1] 82.51 %% (POZO ANIDADO EN HORA)
> round(100 * var.resid / var.total, 2)
[1] 17.49 %% RESIDUAL
```

```
# Alternativa moderna al calculo de las componentes de varianza
library(nlme)
data.lme <- lme(Salinitat ~ 1, random = ~ 1|Hora/Pou, data = pou, method = "REML")
varcomp <- VarCorr(data.lme)
varcomp
```

```
> varcomp
```

| | Variance | StdDev |
|-------------|--------------|-----------|
| Hora = | pdLogChol(1) | |
| (Intercept) | 3.456674988 | 1.8592135 |
| Pozo = | pdLogChol(1) | |
| (Intercept) | 0.016708333 | 0.1292607 |
| Residual | 0.003541667 | 0.0595119 |

Pozos y salinidad

Cálculos con R

```
#Estimacin de la media global y los efectos del factor hora:  
coef(pous1.aov)  
dummy.coef(pous1.aov)  
model.tables(pous1.aov)  
model.tables(pous1.aov, type = "means")
```

```
> model.tables(pous1.aov, type = "means")
```

```
Tables of means
```

```
Grand mean
```

```
6.054167
```

```
> model.tables(pous1.aov,type = "effects")
```

```
Tables of effects
```

```
  Hora
```

```
  Hora
```

```
    21h    7h
```

```
-1.3158  1.3158
```

Ejemplo: un experimento espacial con tres factores jerarquizados

Imaginad que hemos diseñado un experimento en el que pretendemos medir una respuesta (Y) a de unos tratamientos (tres niveles; 'a1', 'a2' y 'a3') en un bosque.

Con las siguientes características:

- ▶ Los tratamientos se realizan a escala espacial (sobre un área) que excede con creces la escala logística de las unidades de muestreo (se tardaría demasiado tiempo en la muestra en la escala en que se aplicaron los tratamientos).
- ▶ Los tratamientos se hicieron a escala de hectáreas, mientras que sólo era posible tomar muestras mediante cuadrados de 1 m²



Ejemplo: Un experimento espacial con tres factores jerarquizados



Algunas características de este diseño espacial son:

- ▶ Los lugares en los que pretendía muestrear eran muy desiguales
- ▶ No es posible disponer de más de cinco réplicas para de cada tipo de tratamiento
- ▶ Para evaluar el efecto del tratamiento, se decidió finalmente utilizar un diseño anidado (jerarquizado) en el que se ubicaban de forma aleatoria 10 cuadrados dentro de cada uno de los cinco sitios de muestreo (al azar) para cada uno de los tres tratamientos utilizados.
- ▶ Ver un ejemplo simulado en: [Link](#)

Un experimento espacial con tres factores jerarquizados

Datos en R (datos simulados)

```
#Simular los datos
set.seed(1)
nTreat <- 3
nSites <- 15
nSitesPerTreat <- nSites/nTreat
nQuads <- 10
site.sigma <- 12
sigma <- 5
n <- nSites * nQuads
sites <- gl(n = nSites, k = nQuads, lab = paste0("S", 1:nSites))
A <- gl(nTreat, nSitesPerTreat * nQuads, n, labels = c("a1", "a2", "a3"))
a.means <- c(40, 70, 80)
## the site means (treatment effects) are drawn from normal distributions with means of 40,
70 and 80
## and standard deviations of 12
A.effects <- rnorm(nSites, rep(a.means, each = nSitesPerTreat), site.sigma)
Xmat <- model.matrix(~sites - 1)
lin.pred <- Xmat %*% c(A.effects)
## the quadrat observations (within sites) are drawn from normal distributions with means
according to
## the site means and standard deviations of 5
y <- rnorm(n, lin.pred, sigma)
data.nest <- data.frame(y = y, A = A, Sites = sites, Quads = 1:length(y))
head(data.nest) ##print out the first six rows of the data set

#Variable de respuesta (Y=medida ambiental) . A = factor tratamiento, B (sites=lugares) i
C #(quads=cuadrados de muestreo 1m2)
#B y C factores aleatorios.
```

Un experimento espacial con tres factores jerarquizados

Obtención de la tabla ANOVA:

```
# Generamos un objeto data.aov, un ANOVA 3F jerarquico
data.aov <- aov(Y ~ A + Sites %in% A + Quads %in% (A/Sites), data.nest)
# o bien
# sentencia alternativa para ANOVA
# data.aov <- aov( resp ~ factorA + factorA/factorB + factorA/FactorB/
  factorC,data.nest)
summary(data.aov) #la tabla obtenida es incorrecta (solo considera factores fijos
)
```

#Se obtiene una tabla ANOVA incorrecta. Solo considera factores fijos

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---------------|-----|--------|---------|---------|------------|
| A | 2 | 38547 | 19273 | 971.561 | <2e-16 *** |
| A:Sites | 12 | 22615 | 1885 | 95.001 | <2e-16 *** |
| A:Sites:Quads | 15 | 200 | 13 | 0.673 | 0.806 |
| Residuals | 120 | 2380 | 20 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ha de calcularse la tabla correcta utilizando AnovaTools

Un experimento espacial con tres factores jerarquizados

Script R para obtener una tabla ANOVA correcta (2 factores aleatorios jerarquizados uno en el otro):

```
library(AnovaTools)
# La tabla ANOVA correcta dado que el primer factor es fijo y los otros aleatorios
anova3nest(data.aov, random = c("Quads", "Sites"))
```

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---------------|-----|--------|---------|---------|---------------|
| A | 2 | 38547 | 19273.3 | 10.227 | 0.002556 ** |
| A:Sites | 12 | 22615 | 1884.6 | 141.170 | 3.672e-13 *** |
| A:Sites:Quads | 15 | 200 | 13.3 | 0.673 | 0.806473 |
| Residuals | 120 | 2380 | 19.8 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
#Para calcular las componentes de la varianza podemos hacer
```

```
library(nlme)
data.nest1.lme <- lme(y ~ A, random = ~1 | Sites/Quads, data.nest, method = "REML")
VarCorr(data.nest1.lme)
```

| | Variance | StdDev |
|-------------|--------------|-----------|
| Sites = | pdLogChol(1) | |
| (Intercept) | 186.546440 | 13.658200 |
| Quads = | pdLogChol(1) | |
| (Intercept) | 14.309919 | 3.782845 |
| Residual | 4.806668 | 2.192411 |

#o manualment (veure formules)

ANOVA BAYESIANO

Toni Monleón-Getino

Junio de 2020

Fiolosofía de los métodos bayesianos

"En nuestros razonamientos sobre los hechos, se dan todos los grados imaginables de seguridad, desde la certeza más alta hasta la especie más baja de evidencia moral. Un hombre sabio, por lo tanto, proporciona su creencia a la evidencia".

David Hume (Filósofo, 1711-1776)

Ver introduccion a los métodos estadísticos bayesianos en:

- ▶ WIN-BUGS: un software para el análisis de modelos bayesianos usando MCMC (A. Monleón-Getino, C. Crespo-Palomo, 2017)
- ▶ Ver un ejemplo en el siguiente link: ANOVA bayesian hypothesis

¿Tiene ventajas el uso de métodos bayesianos?

- ▶ Representa diversas fuentes de incertidumbre de parámetros, especialmente para componentes de varianza, mientras el enfoque frecuentista considera solo la incertidumbre del muestreo.
- ▶ La inferencia se basa en la distribución a posteriori completa de los parámetros del modelo que pueden evaluarse con la suficiente precisión mediante simulación MCMC. Por lo tanto, las medidas descriptivas y los intervalos se pueden obtener de la distribución a posteriori de manera precisa y confiable independientemente del tamaño muestral (UNA GRAN VENTAJA)
- ▶ Podemos introducir información a priori a nuestro modelo, siempre cuando tengamos información.

Métodos bayesianos: ¿Por qué utilizarlos? ¿Por qué se han popularizado?

- ▶ Los métodos bayesianos se han vuelto cada vez más populares en casi todas las disciplinas científicas, debido principalmente a los métodos computacionales desarrollados a finales del S. XX y sus lenguajes/librerías asociados (BUGS, JAGS, STAN).
- ▶ Una razón importante para esta ganancia en popularidad es la facilidad con que los métodos bayesianos pueden aplicarse a problemas relativamente complejos que involucran, por ejemplo, el modelado jerárquico o la comparación entre modelos no anidados.
- ▶ Sin embargo, los métodos bayesianos también se pueden aplicar en escenarios estadísticos más simples, como los que presentan procedimientos de prueba básicos. Ejemplos prominentes de tales procedimientos incluyen análisis de varianza (ANOVA) y regresión. Estas pruebas son la piedra angular del análisis de datos en campos como la biología y la medicina.
- ▶ En los últimos años todos los paquetes estadísticos importantes de análisis de datos (ej: R, SPSS, SAS, etc) han creado librerías para el análisis ANOVA utilizando métodos bayesianos (ej: rstanarm en R).
- ▶ Antes de entrar continuar en este capítulo se debería repasar los conceptos de probabilidad simple de un suceso, probabilidad condicionada, teorema de las probabilidades condicionadas y Teorema de Bayes.

¿Qué es la inferencia bayesiana?

- ▶ Supongamos que salimos una mañana y notamos que la acera está mojada, y nos preguntamos por qué. Consideramos todas las posibles causas de la humedad, incluidas las posibilidades como lluvia reciente, riego reciente de jardines, un manantial subterráneo recientemente erupcionado, una tubería de alcantarillado rota, un transeúnte que derramó una bebida, etc.
- ▶ Si todo lo que sabemos hasta este punto es que alguna parte de la acera está mojada, entonces todas esas posibilidades tendrán cierta credibilidad previa basada en el conocimiento previo. Por ejemplo, la lluvia reciente puede tener mayor probabilidad previa que una bebida derramada de un transeúnte. Continuando nuestro viaje exterior, miramos a nuestro alrededor y recopilamos nuevas observaciones.
- ▶ Si observamos que la acera está mojada hasta donde podemos ver, al igual que los árboles y los automóviles estacionados, entonces reasignamos la credibilidad a la hipotética causa de la lluvia reciente. Las otras causas posibles, como un transeúnte que derrama una bebida, no darían cuenta de las nuevas observaciones.
- ▶ Por otro lado, si en cambio observamos que la humedad estaba localizada en un área pequeña, y había una taza de bebida vacía a unos metros de distancia, entonces reasignaríamos credibilidad a la hipótesis de la bebida derramada, a pesar de que tenía relativamente baja probabilidad previa



Este tipo de reasignación de credibilidad a través de las posibilidades es la esencia de la inferencia bayesiana. Es un proceso natural y lógico ¿por qué no utilizarlo como método estadístico? (John Kruschke, Doing Bayesian Data Analysis)

Análisis bayesiano

- ▶ El análisis bayesiano se basa en la teoría de probabilidad
- ▶ La incertidumbre del ítem se puede representar mediante probabilidades $P(A)$
- ▶ Las probabilidades se actualizan en función de nueva información que va apareciendo
- ▶ Thomas Bayes (170? - 1761) fue un inglés inconformista, ministro presbiteriano y matemático que desarrolló el denominado "Teorema de Bayes"
- ▶ Richard Price publicó el artículo de Bayes sobre el uso de las probabilidades condicionales en 1763 después de la muerte de Bayes
- ▶ Bayes consideró el problema de la probabilidad inversa, elemento muy significativo de la teoría bayesiana
- ▶ Bayes no inventó la probabilidad bayesiana, pero fue el primero en resolver el problema de probabilidad inversa en casos especiales
- ▶ La Teoría bayesiana moderna se basa en pruebas rigurosas desarrolladas en el siglo XX

Base matemática de los métodos bayesianos

¿Que tratamiento en el más eficaz?

- ▶ Los métodos bayesianos se basan en la teoría de probabilidad (Teorema de Bayes) y en la denominada "Teoría de la decisión"
- ▶ En busca de un complemento al método frecuentista que permita la deducción de las técnicas propuestas por Fisher y sus colegas se desarrollaron axiomas conocidas como la teoría de la decisión (relacionadas directamente con la teoría de la probabilidad, convirtiéndose en la base de la inferencia bayesiana.
- ▶ La teoría de la decisión es un método para la toma de decisiones el cual se caracteriza por hacer elecciones de forma coherente cuando se presentan varias opciones ante un problema dado.
- ▶ La toma de decisiones consiste, básicamente, en elegir una alternativa entre las disponibles, a los efectos de resolver un problema actual o potencial.
- ▶ Al tener un problema de decisión se deberá asignar una distribución de probabilidad para realizar la elección de la opción más probable.

El Teorema de Bayes



- ▶ Thomas Bayes plantea un teorema de probabilidad que cumple los supuestos de los sucesos disjuntos y exhaustivos, presentando la proporción de un evento aleatorio
- ▶ El teorema de Bayes es de una enorme importancia puesto que relaciona la probabilidad de A dado B con la probabilidad de B dado A.
- ▶ Es decir, por ejemplo, que sabiendo la probabilidad de tener un dolor de cabeza dado que se tiene gripe, se podría saber (si se tiene algún dato más), la probabilidad de tener gripe si se tiene un dolor de cabeza.
- ▶ Muestra este sencillo ejemplo la alta relevancia del teorema en cuestión para la ciencia en todas sus ramas, puesto que tiene vinculación íntima con la comprensión de la probabilidad de aspectos causales dados los efectos observados.

Teorema de Bayes

El Teorema de Bayes plantea la probabilidad de que se haya dado un determinado suceso (A) sabiendo que como resultado final del experimento se ha obtenido otro determinado suceso (B).

$$P(A_k/B) = P(B/A_k)P(A_k) / \sum_{i=1}^n P(B/A_i)P(A_i), \quad k = 1, \dots, n$$

- ▶ Observa que se intenta calcular una probabilidad “antinatural”, pues se pretende expresar lo que ocurre antes, A_k , en función de lo que ocurre después, B .
- ▶ De todas formas, lo anterior tiene sentido porque en algunas ocasiones se conoce el resultado final de un experimento, pero se desconocen algunos de los pasos intermedios, en los que se está interesado.
- ▶ El teorema de Bayes resuelve esta cuestión, llevando el cálculo de las probabilidades a un terreno más natural, expresando las probabilidades a posteriori, $P(A_i/B)$, en función de las *verosimilitudes*, $P(B/A_i)$.
- ▶ En estadística, la función de verosimilitud (o verosimilitud) es una función de los parámetros de un modelo estadístico que permite realizar inferencias acerca de su valor a partir de un conjunto de observaciones.

Incertidumbre y modelado probabilístico

- ▶ Existen dos tipos de incertidumbres: aleatoria y epistemológica
- ▶ Representaremos la incertidumbre a través de probabilidades
- ▶ Actualizaremos la incertidumbre a medida que avanzamos en el conocimiento del fenómeno estudiado (ej: experimento para determinar el mejor tratamiento).

¿A que son debidas estos tipos de incertidumbres?

- ▶ Incertidumbre aleatoria es debida a la aleatoriedad de los fenómenos (ej: variabilidad biológica)
 - ▶ No podemos obtener observaciones que puedan reducir esta incertidumbre
- ▶ Incertidumbre epistémica es debida a la falta de conocimiento.
 - ▶ Si podemos obtener observaciones que pueden reducir esta incertidumbre
 - ▶ Dos observadores pueden tener diferente incertidumbre epistémica
 - ▶ Las probabilidades debidas a la incertidumbre epistémica cambian cuando cambia la información (ej: conocimiento de un experimento anterior, prueba piloto)

El Teorema de Bayes puede utilizarse fácilmente a la modelización de experimentos, como los que hemos estudiando donde tenemos una variable de respuesta (Y) y unos factores experimentales (X), realizados mediante un determinado modelo.

Patatas fritas



¿Cual es la probabilidad de encontrar patatas fritas de un cierto color en mi bolsa?

Actualizando la incertidumbre

Tomemos este ejemplo de Aki Vehtari de la Aalto University sobre un conteo de tipos de "patatas fritas". Imaginemos que queremos conocer la proporción de patatas fritas de color rojo (R), e imaginemos que en la bolsa hay una proporción de color rojo (R) y una de color amarillo (A).

- ▶ Probabilidad de rojo $P(R) = \frac{\#R}{\#R+\#A} = \theta$
- ▶ $p(y = R|\theta) = \theta$ puede definirse como incertidumbre aleatoria
- ▶ $p(\theta)$ podemos decir que es la incertidumbre epistemologica
- ▶ Elegir muchas patatas actualiza nuestra incertidumbre sobre la proporción de R y A
- ▶ $p(\theta|y = R, A, R, R, A, R, R, \dots) = ?$
- ▶ Así según el teorema de Bayes (Regla bayesiana o Bayes rule):
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

Modelo vs. verosimilitud (likelihood)

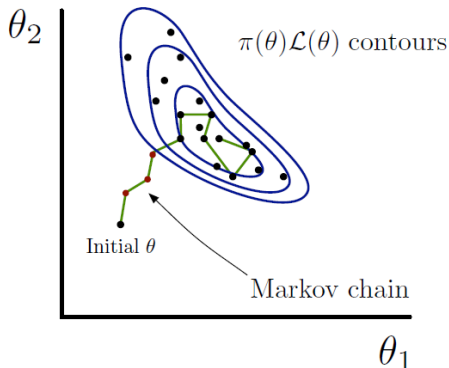
$\int p(y|\theta)p(\theta)d\theta$ es la probabilidad total (se conocen todas las probabilidades), suele obviarse en los cálculos bayesianos ya que suele ser muy complejo su cálculo.

- ▶ Utilizando la regla bayesiana (Bayes rule): $p(\theta|y) \propto p(y|\theta)p(\theta)$
- ▶ Modelo: $p(y|\theta)$ es una función de y dado fijo θ describe la incertidumbre aleatoria.
- ▶ Verosimilitud (Likelihood): $p(y|\theta)$ es una función de θ dado fijo y proporciona información sobre la incertidumbre epistémica, pero no es una distribución de probabilidad. Es la verosimilitud del evento y (conocer la proporción de patatas) dado θ (ser roja) , también llamada probabilidad condicional
- ▶ $p(\theta)$ es la probabilidad de los sucesos condicionados, también denominada probabilidad "a priori".
- ▶ La regla de Bayes combina la probabilidad con incertidumbre previa $p(\theta)$ y los transforma en incertidumbre posterior actualizada: $p(\theta|y)$ que nos permitirá conocer la respuesta a nuestra pregunta sobre la proporción de patatas fritas y hacer inferencia sobre ella.

El Teorema de Bayes posee una controversia, debido a que la estadística frecuentista solo admite probabilidades basadas en experimentos repetibles con confirmación empírica, mientras que para el análisis bayesiano la probabilidad es subjetiva debido a la información adicional que se posee del suceso.

Método de Markov Chain Monte Carlo (MCMC)

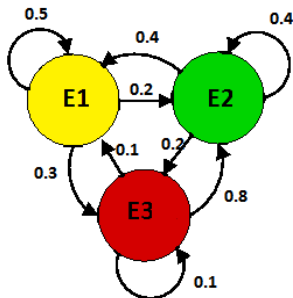
Para calcular la distribución de probabilidad de interés $p(\theta|y)$ se suele utilizar un método numérico computacional basado en las cadenas de Markov denominado Markov Chain Montecarlo (MCMC). Este método requiere muchos cálculos computaciones intensivos.



Cadenas de Markov

Las cadenas de Markov, son un sistema numérico para calcular probabilidades que sufren transiciones de un estado a otro dentro de un espacio de estados.

Link de ciencia



Cadena de Markov simple de 3 estados definidos. Por ejemplo, partiendo del estado E2 podemos ir al estado E1 con una probabilidad de 0.4, ir al E3 con una probabilidad de 0.2 y quedarse en el estado E2 con una probabilidad de 0.4. El siguiente estado depende del estado anterior pero no de ninguno previo, a esta cualidad la conocemos como propiedad de Markov y para la probabilidad a posteriori sería.

MCMC

- ▶ Las muestras aleatorias de las distribuciones marginales condicionales para cada variable aleatoria finalmente convergen en muestras aleatorias de la distribución posterior conjunta $p(\theta|y)$.
- ▶ La cadena de Markov es un proceso estocástico discreto en el cual la probabilidad posterior de un evento ($p(\theta|y)$) depende del suceso inmediatamente anterior. Esto se debe a que las cadenas tienen memoria corta que permiten condicionar las probabilidades futuras.
- ▶ Monte Carlo es un método no determinista o estadístico numérico usado para aproximar expresiones matemáticas complejas y costosas de evaluar mediante simulación. Su objetivo es realizar una serie de repeticiones de n puntos del espacio de varias dimensiones (M -dimensional) a través de un generador de números aleatorios reconociendo el comportamiento del sistema.
- ▶ El método MCMC nos permite resolver una multitud de problemas ya sea de optimización, minimización entre otras, fusionando el método de Montecarlo (simulación) con las cadenas de Markov.
- ▶ A pesar de que MCMC no es un método netamente de la inferencia bayesiana es muy utilizado en el área computacional, ya que permite generar muestras y estimar cantidades o parámetros de interés de la distribución a posteriori ($p(\theta|y)$).
- ▶ MCMC se basa en el Algoritmo de Metrópolis Hastings (MH) que simula una cadena de Markov con valor inicial θ_0 y distribución estacionaria $\pi(\theta|x)$. Utiliza el muestreo de Gibbs que es un caso particular del algoritmo MH en el cual las densidades propuestas coinciden con las distribuciones a posteriori ($p(\theta|y)$) con la probabilidad de aceptación de 1.

ANOVA bayesiano: ejemplo ratas y dieta de proteínas

ANOVA bayesiano, que tiene la ventaja de que pueden reutilizarse los parámetros previos, es robusto con la heterocedasticidad y no le afecta la multiplicidad de pruebas debido a su filosofía.

Veamos un ejemplo procedente de: Estimating ANOVA Models with library(rstanarm): [Bayesian ANOVA](#)

- ▶ Se analiza un experimento de dos factores cruzados con interacción y asignación aleatoria, en el que 40 ratas fueron sometidas a diferentes dietas para ver cuánto peso ganaron.
- ▶ Los factores experimentales (tipo fijo) fueron si su dieta tenía proteínas bajas o altas ("High", "Low") y si la proteína se derivaba de la carne o el cereal ("Beef", "Cereal").
- ▶ Antes de realizar el experimento, uno podría esperar que una proporción moderada de la variación en el aumento de peso se atribuya a la proteína (fuente) en la dieta.



ANOVA bayesiano: ejemplo ratas y dieta de proteínas

Cálculos con R tal como hemos realizado anteriormente (modo frecuentista):

```
library(HSAUR3)
data("weightgain", package = "HSAUR3")
#ANOVA
res1<-aov(weightgain ~ source * type, data = weightgain)
summary(res1) #ANOVA table
#medias para cada nivel del factor
model.tables(res.aov, type="means", se=TRUE) #Medias
```

```
          Df Sum Sq Mean Sq F value Pr(>F)
source     1    221   220.9    0.988 0.3269
type       1   1300  1299.6    5.812 0.0211 *
source:type 1    884   883.6    3.952 0.0545 .
Residuals 36   8049   223.6

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#MEANS OF THE INTERACTION
source:type
      type
source  High  Low
Beef   100.0  79.2
Cereal  85.9  83.9
```

Resultado: Parece que la variación en el peso de las ratas es explicada por el tipo de proteína ($p < 0.05$) y la interacción no es significativa, pero su valor nos hace dudar ($p = 0.0545$). Podemos obtener también los denominados coeficientes del modelo lineal (β_i) si tratamos este experimento como un modelo lineal

ANOVA bayesiano: ejemplo ratas y dieta de proteínas)

Debido a que ANOVA es un tipo de modelo lineal, podemos usar la función `lm()` alternativamente y obtener los coeficientes del modelo lineal (β_i), ligeramente diferente en cuanto a su significado a los efectos de los factores que ya estudiamos, ahora β_i pueden ser interpretados como una pendiente de cada nivel del factor de referencia (basal).

```
res.lm<-lm(weightgainsource * type, data =weightgain)
summary(res.lm)
#se puede hacer igual con
#coef(summary.lm(res.aov))
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------------|----------|------------|---------|-------------|
| (Intercept) | 100.000 | 4.729 | 21.148 | < 2e-16 *** |
| sourceCereal | -14.100 | 6.687 | -2.109 | 0.04201 * |
| typeLow | -20.800 | 6.687 | -3.110 | 0.00364 ** |
| sourceCereal:typeLow | 18.800 | 9.457 | 1.988 | 0.05447 . |

Residual standard error: 14.95 on 36 degrees of freedom

Multiple R-squared: 0.23, Adjusted R-squared: 0.1658

F-statistic: 3.584 on 3 and 36 DF, p-value: 0.02297

Resultado: Observamos que el modelo consta de diferentes coeficientes y que son significativos los de "sourceCereal", "typeLow" y el denominado "Intercept". Asumimos que el ajuste del modelo es razonable para nuestro propósito, el modelo es significativo ($p=0.02297$). `lm()` calcula una parametrización de efectos algo diferente a la del ANOVA, por lo que el primer grupo categórico, "sourceCereal", es el grupo de referencia para el factor "source": nivel ratas que toman "cereal". "typeLow" es el grupo de referencia para el factor "type": nivel ratas con dieta de nivel de proteínas "low". "Intercept" es la situación basal que presenta un coeficiente de 100 que corresponde a la estimación del peso ganado por aquellas ratas con "type" "High" y "source" "Beef". Los otros coeficientes pueden ser interpretados como la variación en cuanto al peso ganado por las ratas que han sido sometidas a tal o cual nivel del factor.

Análisis bayesiano: ANOVA como modelo lineal

Veamos ahora cómo realizar un ANOVA bayesiano y cómo se debe interpretar, así tomamos el ejemplo anterior e iremos realizándolo paso a paso.

Así hemos visto anteriormente que el ANOVA puede ser modelizado mediante un modelo lineal de variable de respuesta Y_i de tipo **continua/cuantitativa** de individuos i en relación a una o más variables explicativas específicas X_{i1}, \dots, X_{ip} que en este caso son cualitativas (factores) mediante el modelo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i.$$

- ▶ Intercept (término independiente): β_0
- ▶ Coeficientes de regresión: β_1, \dots, β_p
- ▶ Término de error ϵ_i con media 0, captura la variabilidad residual (no explicada por los factores).

Un objetivo típico es estudiar los cambios de la media de Y_i bajo los cambios de X_{i1}, \dots, X_{ip} .

Pasos en ANOVA bayesiano: ejemplo ratas y dieta de proteínas)

Los cuatro pasos de un análisis bayesiano (en general) son:

- ▶ 1. Especificar una distribución conjunta para el (los) resultado (s) Y_i y todas las incógnitas X_{1i}, X_{2i}, \dots , que generalmente toma la forma de una distribución previa marginal para las incógnitas (Ejemplo: distribución normal) multiplicada por una probabilidad de los resultados condicionales a las incógnitas.
- ▶ 2. Esta distribución conjunta es proporcional a una distribución posterior de las incógnitas condicionadas a los datos observados: $p(\theta|y) \propto p(y|\theta)p(\theta)$
- ▶ 3. Calcular la distribución posterior ($p(\theta|y)$) utilizando Markov Chain Monte Carlo (MCMC). Este paso no es completamente automático porque a veces es necesario especificar algunos parámetros de ajuste adicionales para obtener resultados óptimos y eficientes.
- ▶ 4. Evaluar qué tan bien el modelo se ajusta a los datos y posiblemente revisar el modelo.
- ▶ 4. Extraer de la distribución predictiva posterior de los resultados ($p(\theta|y)$) dados los valores interesantes de los predictores para visualizar cómo un cambio de un predictor afecta (una función de) los resultados.

Verosimilitud: ejemplo ratas y dieta de proteínas

Primeramente deberemos indicar la verosimilitud del modelo.

Verosimilitud del modelo lineal ANOVA (Likelihood)

La probabilidad de una observación bajo un modelo lineal se puede escribir como la función de densidad de probabilidad (PDF) condicionalmente normal (gausiana):

$$\frac{1}{\sigma_{\epsilon} \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{y - \mu}{\sigma_{\epsilon}} \right)^2}$$

donde $\mu = \alpha + \mathbf{x}^T \boldsymbol{\beta}$

es un predictor lineal y σ_{ϵ} es la desviación estándar del error al predecir el resultado, y . La probabilidad de toda la muestra es el producto de N contribuciones de probabilidad individuales.

Como hemos visto anteriormente, un modelo ANOVA puede considerarse un caso especial del modelo de regresión lineal en el que cada uno de los k predictores de X es una variable ficticia que indica la pertenencia a un grupo. Equivalentemente, se puede escribir un predictor como $\mu_j = \alpha + \alpha_j$, que expresa la esperanza (media) condicional del resultado en el grupo jj – *esimo* como la suma de una media general (o común), α y una desviación específica del grupo de media común,

$$\alpha_j$$

Priors: ejemplo ratas y dieta de proteínas

Priors del modelo lineal ANOVA: Esta especificación previa sobre R^2

- ▶ Si consideramos el modelo ANOVA como un caso especial de un modelo de regresión lineal con solo variables ficticias como predictores, entonces el modelo podría estimarse utilizando la especificación de distribuciones a priori.
- ▶ Las funciones para el cálculo bayesiano (ej: `stan_lm`, `stan_aov` de la `library(rstanarm)`) requieren que el usuario especifique un valor del conocimiento "a priori" por ejemplo del R^2 , la proporción de variación en el resultado atribuible a los predictores bajo un modelo lineal.
- ▶ Esta especificación previa sobre R^2 es atractiva en un contexto ANOVA debido a la identidad fundamental de la descomposición de la suma de cuadrados:
$$SS_{total} = SS_{model} + SS_{error},$$
donde SS es la suma de cuadrados. Si normalizamos esta identidad, obtenemos la tautología $1 = R^2 + (1 - R^2)$ pero es razonable esperar que un investigador tenga una suposición plausible para R^2 antes de realizar un ANOVA, así que podemos incorporar esta información a la verosimilitud.
- ▶ Por otra parte, si desconocemos esta suposición plausible para R^2 pero si disponemos de estudios previos y conocimiento de los parámetros, o suposiciones sobre los mismos podemos utilizar otro abordaje.

Priors del modelo lineal ANOVA: conocimiento previo de los parámetros del modelo

- ▶ Por otra parte, si desconocemos esta suposición plausible para R^2 pero si disponemos de estudios previos y conocimiento de los parámetros, o suposiciones sobre los mismos podemos utilizar otro abordaje.
- ▶ Si consideramos el modelo ANOVA como una diferencia de medias, entonces el modelo podría estimarse utilizando la especificación previa en la función `stan_lmer`.
- ▶ En la sintaxis popularizada por el paquete `lme4`, $y \sim 1 + (1||grupo)$ (generalmente utilizado para factores aleatorios) representa una probabilidad donde

$$\mu_j = \alpha + \alpha_j$$

se distribuye normalmente entre los grupos J con media cero y desviación estándar σ desconocida.

- ▶ La función `stan_lmer` especifica que esta desviación estándar σ tiene una distribución priori Gamma con, por defecto, sus parámetros de forma y escala iguales a 1, así es solo una distribución exponencial estándar.
- ▶ Sin embargo, los parámetros de forma y escala se pueden especificar como otros valores positivos.
- ▶ Este enfoque también requiere especificar una distribución previa sobre la desviación estándar de los errores que sea independiente de la distribución previa para cada α_j .
- ▶ Así cada parámetro de una distribución tiene asociada una distribución de probabilidades, no es algo fijo como en la estadística frecuentista

ANOVA bayesiano: ejemplo ratas y dieta de proteínas

Para obtener estimaciones bayesianas, podemos utilizar la función `stan_aov()` de la `library(rstanarm)` y especificar el conocimiento previo ("a priori") de R^2 , y otros argumentos si es necesario (argumentos computacionales).

Imaginemos que $R^2 = 0.5$, en R podríamos escribir utilizando `library(rstanarm)`:

```
library(rstanarm)
post1 <- stan_aov(weightgain ~ source * type, data = weightgain,
                 prior = R2(location = 0.5), adapt_delta = 0.999,
                 seed = 12345)
post1
```

Aquí hemos especificado el argumento `adapt_delta = 0.999` (tasa de optimización) para disminuir el tamaño de pasos y evitar en gran medida las transiciones divergentes. Los resultados pueden verse en la siguiente página.

ANOVA bayesiano: ejemplo ratas y dieta de proteínas

Los resultados obtenidos en el ANOVA bayesiano son los siguientes

```
stan_aov
family:      gaussian [identity]
formula:     weightgain ~ source * type
observations: 40
predictors:  4
```

```
-----
                Median MAD_SD
(Intercept)      98.8    4.4
sourceCereal    -12.9    6.3
typeLow         -18.6    6.3
sourceCereal:typeLow 16.7    9.0
```

Auxiliary parameter(s):

```
                Median MAD_SD
R2              0.2    0.1
log-fit_ratio  0.0    0.1
sigma          14.7    1.7
```

ANOVA-like table:

```
                Median MAD_SD
Mean Sq source  556.2  431.7
Mean Sq type   978.3  594.3
Mean Sq source:type 700.9  698.0
```

Resultado: Obtenemos un valor $R^2 = 0.2$. Podemos observar que nuestra suposición anterior de que $R^2 = 0.5$ era demasiado optimista. Sin embargo, las estimaciones frecuentistas presumiblemente sobreajustan los datos aún más y son parecidas a las anteriores (Intercept, sourceCereal, typeLow, sourceCereal:typeLow).

Otros diseños ANOVA y temas relacionados

Toni Monleón-Getino

Junio de 2020

Otros temas relacionados con ANOVA

Ver un ejemplo en el siguiente link: Temas avanzados en ANOVA Selecció de temes segons Nathaniel E. Helwig. Univ of Minesota (USA):

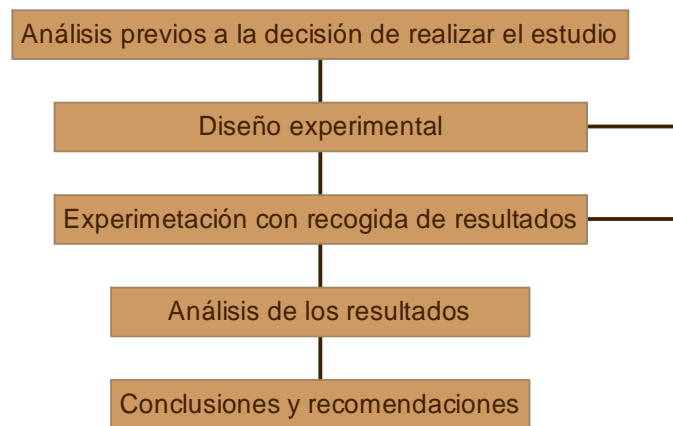
- ▶ One-Way Analysis of Variance (R code)
- ▶ Factorial and Unbalanced ANOVA (R code)
- ▶ Analysis of Covariance (R code)
- ▶ Effect Size and Power Analysis (R code)
- ▶ Linear Mixed-Effects Regression (R code)
- ▶ Permutation Tests (R code) Resampling (R code) Confidence Intervals (R code) Regression (R code)
- ▶ Smoothing Spline ANOVA (R code)

2-Diseño y análisis práctico de experimentos con R

2.1-El método científico

Una hipótesis es una explicación propuesta para un fenómeno. Para que una hipótesis sea una hipótesis científica, el método científico requiere que uno pueda probarla. Los científicos generalmente basan las hipótesis científicas en observaciones previas que no pueden explicarse satisfactoriamente con las teorías científicas disponibles (Monleón-Getino, 2017).

Pasos generales utilizados durante la aplicación del método científico:



2.2-¿Por qué realizar experimentos?

Un experimento se caracteriza por los tratamientos y las unidades experimentales que se utilizarán, la forma en que los tratamientos se asignan a las unidades y los resultados que se miden.

Los investigadores usan experimentos para responder preguntas. Algunas de las preguntas típicas que los Experimentos responden a las preguntas son:

¿Es una droga una cura segura y efectiva para una enfermedad? Esto podría ser una prueba de cómo el AZT afecta el progreso del SIDA.

¿Qué combinación de fuentes de proteínas y carbohidratos demuestra la mejor nutrición para el cultivo de corderos?

.¿Cómo cambiará el uso del teléfono de larga distancia si nuestra compañía ofrece una estructura de tarifas diferente a nuestros clientes?

. ¿Un helado fabricado con un nuevo tipo de estabilizador será tan apetecible como nuestro helado actual?

¿Bajo qué condiciones debo operar mi refinería química, dado el grado de materia prima de este mes?



Imagen de Wikipedia(<https://www.shutterstock.com/es/video/clip-8561515-scientist-doing-chemical-experiment-laboratory>)

2.3-Los experimentos nos ayudan a contestar preguntas

Los experimentos nos ayudan a responder preguntas, pero también hay técnicas no experimentales. ¿Qué tienen de especial los experimentos?

Ventajas de los experimentos

1. Los experimentos nos permiten establecer una comparación directa entre los tratamientos de interés.
2. Podemos diseñar experimentos para minimizar cualquier sesgo en la empresa.
3. Podemos diseñar experimentos para que el error en las comparaciones sea pequeño.

4. Lo más importante es que tenemos el control de los experimentos y el control que nos permite hacer inferencias más sólidas sobre la naturaleza de las diferencias que vemos en el experimento. Específicamente, podemos hacer inferencias sobre la causación.

2.4-Estadística frecuentista y bayesiana

Este manual se ha escrito eminentemente desde el punto de vista clásico, mal llamado “frecuentista”, aunque hace ya bastantes años que existe otro enfoque diferente llamado bayesiano, es todavía poco utilizado, aunque si me gustaría presentarlo y comentar algunas de sus peculiaridades.

La estadística frecuentista se desarrolla a partir de los conceptos de probabilidad y que se centra en el cálculo de probabilidades y los contrastes de hipótesis. Tiene como objetivo determinar una conclusión, sea en base a la significación estadística o a la aceptación/rechazo de hipótesis.

En el análisis estadístico de un experimento que pretende comparar la eficacia de un nuevo tratamiento, T, frente a otro estándar, S, se utiliza únicamente la información obtenida en el experimento. No existen subjetividades referentes a parámetros, puesto que se han fijado los criterios de decisión a priori y estos permanecen estáticos durante todo el experimento.

Los métodos bayesianos son un enfoque alternativo a la estadística frecuentista, se basan en la probabilidad obtenida mediante el teorema de Bayes, y que se diferencia de la estadística frecuentista básicamente en la incorporación de información externa al estudio que se esté realizando, de manera que, si conocemos la probabilidad de que ocurra un suceso, su valor será modificado cuando dispongamos de esa información. Así pues, las fuentes de información “a priori” se ven transformadas en probabilidad “a posteriori” y se utilizan a continuación para realizar la inferencia (contraste de hipótesis).

Podemos encontrar algoritmos que nos ayuden en el método estadístico utilizado. Obviamente todo ello basado en la teoría de la probabilidad y estadística. Una buena introducción puede ser en:

MONLEON T, RODRIGUEZ C. Probabilitat i estadística per a ciències I. PPU. BARCELONA. 2015.

MONLEON T, RODRIGUEZ C. Probabilitat i estadística per a ciències II. PPU. BARCELONA. 2017.

2.5-¿Cómo funciona la prueba de hipótesis estadística?

En muchos casos, el investigador tiene una creencia a priori, posiblemente basada en su experiencia previa con el fenómeno que se está estudiando en los valores numéricos de los parámetros de este proceso. En este caso, el investigador está interesado en estimar que los valores numéricos de estos parámetros son desconocidos, pero también puede querer comparar varias hipótesis sobre la distribución de probabilidad de la población que generó la muestra disponible.

En general, las suposiciones se refieren a si la información de la muestra es coherente con su creencia a priori sobre los valores de los parámetros, que formaron los llamados problemas de una muestra.

A veces, el investigador tiene dos muestras y se pregunta si la información proporcionada es consistente con la posibilidad que proviene de la misma población, en contra de la alternativa que proviene de diferentes poblaciones, estos problemas se denominan dos muestras.

Luego, una hipótesis es una afirmación sobre las características estadísticas de un proceso, que puede considerarse una hipótesis y una conjetura.

Por ejemplo, si un científico observa la energía en un determinado proceso metabólico durante varias horas, usted sabe el consumo promedio de las horas anotadas. Con la ayuda de la inferencia, puede avanzar un paso más y especular si el consumo promedio de todas las reacciones observadas es un valor determinado u otro. El proceso científico luego se prueba su hipótesis contra una hipótesis alternativa.

Un ejemplo de hipótesis podría ser la siguiente:

$$\begin{cases} H_o : \mu_A - \mu_B = 0 \\ H_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

Donde nos preguntamos si la media poblacional de la variable Y para la población A es igual a la media poblacional de la población B, para responder a esta pregunta y decidir entre la Hipótesis nula (H_o) o la hipótesis alternativa (H_1) podríamos utilizar la prueba estadística de comparación de medias de la T de Student.

2.6-Multiplicidad de pruebas estadísticas, riesgo alfa y potencia estadística ¿Cómo funciona?

2.6.1- Un poco de teoría

Repasemos algunos conceptos útiles estudiados anteriormente en cursos previos y que se desarrollan en la bibliografía, especialmente en MONLEON y RODRIGUEZ “Probabilitat i estadística per a ciències II” (2017).

Contraste de hipótesis

Los contrastes de hipótesis especifican siempre una posibilidad, llamada hipótesis nula, denotada por H_0 , que es aquella en la que el investigador está dispuesto a creer a priori.

Hay que especificar asimismo una hipótesis alternativa, denotada por H_1 , aquella que pasará a aceptar si rechaza la hipótesis nula. La idea previa a la contrastación estadística de hipótesis es que hay razones para creer que la hipótesis nula pueda ser cierta: es aquel suceso que parece más posible a priori lo que debe definir la hipótesis nula

Por otra parte, la hipótesis alternativa debe estar definida por aquellos sucesos, incompatibles con los que definen la hipótesis nula, que tienen probabilidad positiva. Un suceso de probabilidad nula no debe estar incluido ni en la hipótesis nula ni en la hipótesis alternativa.

Una hipótesis estadística es cualquier afirmación, ya sea verdadera o falsa, sobre alguna característica desconocida sobre la población, normalmente sobre sus parámetros (θ): esperanza, varianza, proporción o de otros.

Los análisis estadísticos pueden ayudarnos a responder a preguntas planteadas en las hipótesis, así si deseamos dilucidar entre las hipótesis:

$$\begin{cases} H_0 : \mu_A - \mu_B = 0 \\ H_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

Utilizando el método científico deberemos seguir el siguiente esquema:

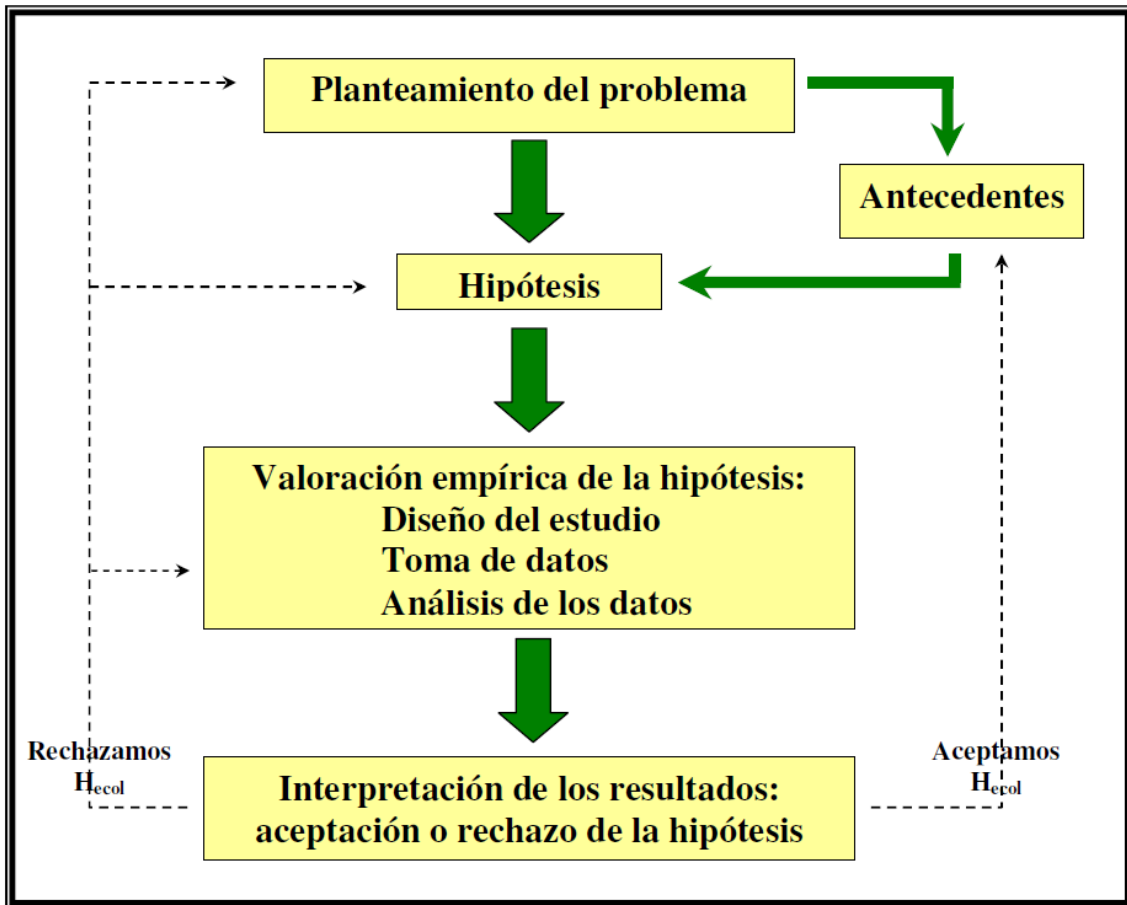


Imagen: Método científico resumido en un esquema (fuente: internet)

Y así deberemos recoger datos y realizar un análisis estadístico inferencial adecuado que permita decidir entre H_0 o H_1 , es lo que se denomina resolver el contraste de hipótesis. El análisis inferencial nos permitirá obtener conclusiones de aplicabilidad más general que la simple muestra y resolver el contraste de hipótesis.

Una vez realizado el contraste de hipótesis, se habrá optado por una de las dos hipótesis, la hipótesis nula o base (H_0) o la hipótesis alternativa (H_1) y la decisión escogida coincidirá o no con lo que en realidad es cierto, así que hay un cierto riesgo de equivocarnos.

Así cuando tras tomar datos experimentales y realizar un contraste estadístico de la hipótesis:

$$\begin{cases} H_0 : \mu_A - \mu_B = 0 \\ H_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

Decidimos H1 o Ho veamos que ocurre.

Etapas en la realización de un contraste¹

1. Describir el modelo y formular la hipótesis nula y la alternativa
2. Definir un estadístico de contraste que cuantifique la discrepancia entre los datos y la hipótesis nula, y cuya distribución sea conocida bajo H0
3. Definir la región crítica: ¿Qué valores del estadístico de contraste rechazan H0?
4. Determinar el valor crítico para un nivel de significación α dado
5. Tomar los datos y calcular el valor del estadístico de contraste
6. Calcular el p-valor
7. Tomar la decisión de rechazar o no H0

Tipos de errores cometidos cuando tomamos una decisión

En las pruebas de hipótesis estadísticas, el error tipo I es el rechazo incorrecto de una hipótesis nula verdadera (un "falso positivo"), mientras que un error tipo II retiene incorrectamente una hipótesis nula falsa (un "falso negativo").

Dicho de manera más simple, un error de tipo I está detectando un efecto que no está presente, mientras que un error de tipo II no puede detectar un efecto que está presente.

Los términos "error tipo I" y "error tipo II" a menudo se usan indistintamente con la noción general de falsos positivos y falsos negativos en la clasificación binaria, como

¹ Modificado de http://www5.uva.es/espartero/Tema8a_2012_alumno.pdf

las pruebas médicas, pero hablando estrictamente se refieren específicamente a las pruebas de hipótesis estadísticas en el marco de referencia de Neyman-Pearson, lema teórico fundamento de las mismas.

Tipos de error

- Error tipo I: ocurre cuando la hipótesis nula es rechazada cuando en realidad es verdadera (riesgo del productor α)
- Error tipo II: ocurre cuando la hipótesis nula es aceptada cuando en realidad debió ser rechazada (riesgo del consumidor β)

| | | Hipótesis nula | |
|--------------------|-------------------|-----------------------------------|----------------------------------|
| | | Verdadero | Falso |
| La decisión tomada | No rechazar H_0 | $p=1-\alpha$ Decisión correcta | $p=\beta$ Error tipo II |
| | Rechazar H_0 | $p=\alpha$ Error tipo I | $p=1-\beta$ Decisión correcta |

Nivel de significación (α) y potencia estadística ($1-\beta$)

Una cosa es nuestra decisión, otra es lo que realmente es verdad (y que nosotros desconocemos).

Nivel de significación: probabilidad de cometer el error de tipo I (rechazar la hipótesis nula cuando en realidad es cierta).

Potencia: probabilidad de no cometer el error de tipo II = 1 - probabilidad de cometerlo (β).

Así consideramos que rechazar H_0 si es cierta (error de tipo I) es el error más grave. Consideramos un nivel de significación α aceptable en el 5% (por consenso)

De todos los criterios (tests, pruebas estadísticas) que respetan α , procuraremos utilizar la que hace mínima la probabilidad de error de tipo II (máxima potencia estadística)

Sobre la hipótesis nula. ¿Es falsa la hipótesis nula?

La hipótesis nula es verdadera o falsa. Lamentablemente, no sabemos cuál es el caso y, en general, nunca lo sabremos. Es importante darse cuenta de que no hay probabilidad de que la hipótesis nula sea verdadera o de que sea falsa. No saber cuál es la correcta no significa que esté involucrada una probabilidad.

Por ejemplo, si se desea probar que un nuevo medicamento genera más respuesta inmune que el medicamento estándar, la hipótesis nula de que el medicamento nuevo produce la misma respuesta inmune que el estándar es verdadera o falsa. No hay ninguna probabilidad asociada con estos dos casos (en un sentido frecuentista) porque el medicamento ya provoca una cierta acción; no hay posibilidad de azar porque todo ya está establecido. Todo lo que tenemos es nuestra propia incertidumbre sobre la hipótesis nula. Deberíamos utilizar otro tipo de hipótesis para demostrar el caso

$$H_o : \mu_A - \mu_B = 0$$

Como la denominada hipótesis de equivalencia.

Esta falta de conocimiento sobre la hipótesis nula es la razón por la que necesitamos realizar una prueba estadística: queremos usar nuestros datos para hacer una inferencia sobre la hipótesis nula. Específicamente, debemos decidir si vamos a actuar como si la hipótesis nula fuera cierta o si fuera falsa. De nuestra prueba de hipótesis, por lo tanto, elegimos aceptar o rechazar la hipótesis nula. Si aceptamos la hipótesis nula, estamos afirmando que nuestros datos son consistentes con la hipótesis nula. Si rechazamos la hipótesis nula, estamos afirmando que nuestros datos son un resultado tan inesperado que son inconsistentes con la hipótesis nula.

IMPORTANTE

- No rechazar H_0 no implica que H_0 sea cierta (no se “demuestra” que H_0 sea cierta) sino que no hay evidencia suficiente en los datos muestrales para rechazarla.
- Rechazar H_0 no significa que H_0 sea falsa, sino que resulta muy difícil creer que se haya podido observar algo tan improbable bajo H_0 .

Debemos tomar una decisión

Nuestra decisión cambiará nuestro comportamiento. Si rechazamos la hipótesis nula, actuaremos como si la hipótesis nula fuera falsa, aunque no sabemos si ese es el caso. Si aceptamos la hipótesis nula, actuaremos como si la hipótesis nula fuera cierta, aun

cuando no hayamos demostrado que sea cierta. Este es un punto crítico: independientemente de los resultados de nuestra prueba estadística, nunca sabremos si la hipótesis nula es verdadera o falsa.

Debido a que tenemos dos posibilidades para la hipótesis nula (verdadero o falso) y dos posibilidades para nuestra decisión (aceptar o rechazar), hay cuatro escenarios posibles. Dos de estas son decisiones correctas: podríamos aceptar una hipótesis nula verdadera o podríamos rechazar una hipótesis nula falsa. Los otros dos son errores. Si rechazamos una hipótesis nula verdadera, hemos cometido un error tipo I. Si aceptamos una hipótesis nula falsa, hemos cometido un error de tipo II.

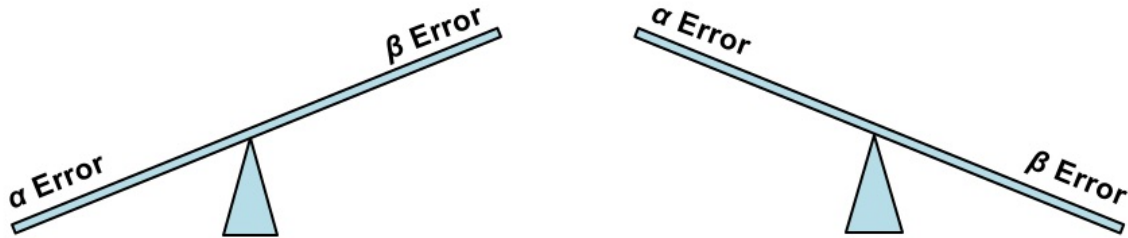
| | Es cierta H_0 | Es cierta H_1 |
|---------------------|-----------------------------------|------------------------------------|
| No rechazamos H_0 | Decisión correcta | Error de tipo II (o de 2ª especie) |
| Rechazamos H_0 | Error de tipo I (o de 1ª especie) | Decisión correcta |

Cada una de estas cuatro posibilidades tiene alguna probabilidad de ocurrir. Si la hipótesis nula es verdadera, solo hay dos posibilidades: elegiremos aceptar la hipótesis nula con probabilidad de $1-\alpha$, o la rechazaremos con probabilidad de α .

Si la hipótesis nula es falsa, solo hay dos posibilidades: elegiremos aceptar la hipótesis nula con una probabilidad de β , o la rechazaremos con una probabilidad de $1-\beta$ (potencia de test o potencia estadística). Debido a que las probabilidades dependen de si la hipótesis nula es verdadera o falsa, son las probabilidades en cada fila las que suman 100 (en porcentaje). Las probabilidades en cada columna no están limitadas a sumar 100.

Como no conocemos la verdad de una hipótesis nula, debemos cubrirnos y disminuir las posibilidades de cometer ambos tipos de error. Si la hipótesis nula es cierta, queremos disminuir nuestras posibilidades de cometer un error tipo I; en otras palabras, queremos encontrar formas de reducir alfa. Si la hipótesis nula es falsa, queremos reducir nuestras posibilidades de cometer un error tipo II: queremos encontrar formas de reducir la beta. Debido a que no sabemos si la hipótesis nula es

verdadera o falsa, necesitamos mantener simultáneamente las probabilidades de alfa y beta lo más pequeñas posible.



Reproduced by permission of John Wiley and Sons

Imagen: <http://daniellakens.blogspot.com/2016/12/why-type-1-errors-are-more-important.html>

Alfa y Beta son parte de un equilibrio y así cuando una aumenta y la otra disminuye.

p-valor

Este concepto se fundamenta en la teoría de test de hipótesis, por ello es necesario conocer bien cómo funciona un contraste estadístico, el concepto de región crítica y región de aceptación.

El p-valor se define, para una muestra concreta, como la probabilidad de observar, bajo H_0 , un valor del estadístico de contraste igual o más extremo (en la dirección de la alternativa) que el observado en la muestra \leftrightarrow probabilidad de obtener más discrepancia con H_0 que la obtenida con la muestra

Cuanto menor el p-valor \rightarrow más extremo el resultado muestral \rightarrow más evidencia contra H_0

Cuando tomamos una decisión en un contraste de hipótesis, se fija previamente el nivel de confianza $1 - \alpha$ (o el nivel de significación α). ¿Qué pasa si modificamos el nivel de confianza? En este caso deberíamos recalculamos las regiones de aceptación y rechazo (región crítica², W) y tomar una nueva decisión. Sin embargo, hay una manera

² Región crítica= W ={valores muestrales que conllevan rechazar H_0 }

de determinar la decisión en un contraste para diferentes niveles de confianza, usando lo que se conoce como p-valor. Los programas informáticos utilizan el p-valor para tomar la decisión en un contraste, de una manera precisa y eficiente.

Como se ha comentado anteriormente el nivel de significación (α) se establece al principio, antes de conocer el valor del estadístico de contraste U^3 . Nos sirve para decidir por qué valores del estadístico rechazaremos la hipótesis nula. En cambio el p-valor es una probabilidad (aceptando cierta hipótesis nula), de encontrar una muestra con un estadístico más extremo de lo que hemos calculado con nuestra muestra.

Podemos hablar del p-valor también como regla de decisión entre H_0 i H_1 .

Según la hipótesis alternativa (H_1) se puede calcular como:

- H_1 Bilateral
 $p - \text{valor} = P(|U| \geq |u_{exp}| \text{ bajo } H_0)$
- H_1 Unilateral por la derecha
 $p - \text{valor} = P(U \geq u_{exp} \text{ bajo } H_0)$
- H_1 Unilateral por la izquierda
 $p - \text{valor} = P(U \leq u_{exp} \text{ bajo } H_0)$

Para tomar la decisión, en lugar de comparar el valor crítico con el estadístico de la muestra, podemos comparar el p-valor con el nivel de significación y esto es independiente del independiente del valor de α elegido. Utilizaremos esta regla de decisión:

Regla de decisión del p-valor

$$p - \text{valor} > \alpha \rightarrow \text{no rechazar } H_0$$

$$p - \text{valor} \leq \alpha \rightarrow \text{rechazar } H_0$$

Valor(es) crítico(s)=C= valor(es) a partir del (de los) cual(es) se rechaza H_0 . La región complementaria a la región crítica es la Región de aceptación=A={valores muestrales que conllevan no rechazar H_0 }

³ Recordemos que la decisión entre H_0 o H_1 se basa en un estadístico de contraste $=U(X_1, \dots, X_n)$ que es una función conocida de la muestra.

Ejemplo de aplicación del p-valor: Si en la H_1 estamos haciendo un contraste unilateral por la derecha:

$$H_0: \theta = \theta_0$$
$$H_1: \theta > \theta_0$$

Se calcularà como: $p - \text{valor} = P(U \geq u_{exp} \text{ bajo } H_0)$

En el ejemplo anterior:

$$H_0: \mu = 400$$
$$H_1: \mu > 400$$

```
#Vamos a simular la situación anterior y después efectuaremos el  
contraste de la hipótesis utilizando un test T Student con una  
única población y hipótesis alternativa unilateral derecha H1: mu >  
400  
#Simulate a sample with mean = 400 and sd = 20  
Sample <- rnorm(n=10, mean =400 ,sd = 20)  
# One-sample t-test  
res <- t.test(Sample, mu = 400,alternative = "greater")  
# Printing the results  
res  
  
#      One Sample t-test  
# data: Sample  
# t = -0.48922, df = 9, p-value = 0.6818  
# alternative hypothesis: true mean is greater than 400  
# 95 percent confidence interval:  
# 388.9273      Inf  
# sample estimates:  
# mean of x  
# 397.6674
```

$$p - \text{valor} = P(U \geq \bar{X}_{exp} \text{ bajo } H_0)$$

No podemos rechazar $H_0: \mu = 400$ si el test es no significativo : $p_{valor} > \alpha$

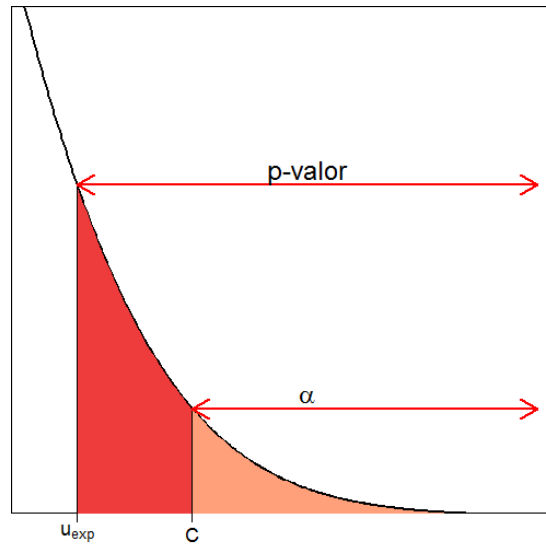


Figura: funcionamiento del p-valor, en comparación con el nivel de significación en el caso de un contraste unilateral derecho ($H_1: \theta > \theta_0$). En esta situación como $U_{exp} < C \rightarrow p_{valor} > \alpha \rightarrow H_0: \theta = \theta_0$. Donde U_{exp} es el valor del estadístico obtenido a partir de la muestra y C^4 el punto crítico del contraste para decidir entre H_0 o H_1 .

Rechazaremos H_0 y por tanto aceptaremos $H_1: \mu > 400$ como válido en el caso de test significativo $p_{valor} \leq \alpha$

⁴ Valor(es) crítico(s)= C=valor(es) a partir del (de los) cual(es) se rechaza H_0

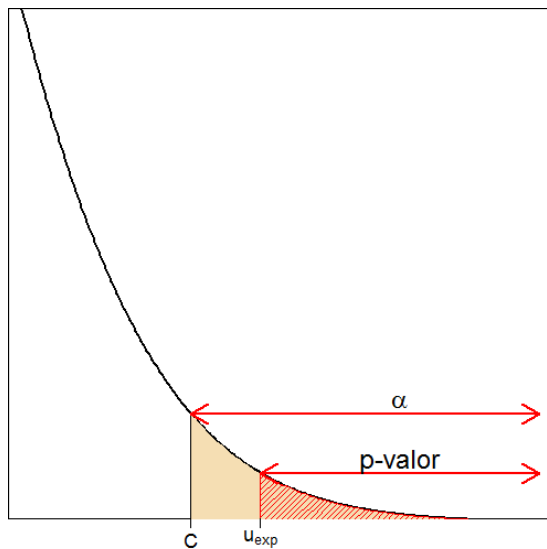


Figura: funcionamiento del p-valor, en comparación con el nivel de significación en el caso de un contraste unilateral derecho ($H_1: \theta > \theta_0$). En esta situación como $U_{exp} > C \rightarrow p_{valor} \leq \alpha \rightarrow H_1: \theta > \theta_0$. Donde U_{exp} es el valor del estadístico obtenido a partir de la muestra y C el punto crítico del contraste para decidir entre H_0 o H_1 .

El valor p-valor se calcula utilizando el estadístico de prueba calculado a partir de las muestras, la distribución asumida y el tipo de prueba que se realiza (prueba de cola inferior, prueba de cola superior o prueba de dos colas). Introduciendo estos valores en un ordenador se obtiene el p-valor.

Para calcular el p-valor sin ayuda de un ordenador se utilizan grandes tablas de valores que contienen aproximaciones a los valores de la función de densidad o distribución de probabilidad. En general, estas tablas se fijan con el eje vertical de la izquierda correspondiente a los grados de libertad y el eje horizontal en la parte superior correspondiente al valor p. Son sólo una aproximación difícil de manejar.

INTERPRETACION DEL P-VALOR

- p-valor muy pequeño \Rightarrow sería muy improbable observar lo observado si H_0 hubiera generado mis datos \Rightarrow los datos proporcionan evidencia suficiente en contra de $H_0 \Rightarrow$ rechazo H_0
- p-valor grande \Rightarrow nuestros datos no proporcionan evidencia suficiente en contra de H_0 (es probable que H_0 haya generado mis datos) y no rechazo.

p-valor en un contexto bayesiano

El p-valor, en los últimos años ha sido cuestionado, especialmente en el mundo clínico.

Desde un contexto bayesiano, un p-valor posterior es la probabilidad, dada la información, de que una observación futura sea más extrema (medida por alguna variable de prueba) que la información

Algunos expertos (ej: Matthews) sostienen que los p-valores exageran la significación real de los datos obtenidos, lo que no pocas veces ha conducido a la aprobación de conductas terapéuticas que poco tiempo después han sido fatalmente desacreditadas; a su juicio, la razón de estos escándalos, simplemente es, estadística. Según este autor, la razón más persuasiva para usar la inferencia bayesiana es la capacidad que tiene de proveer un nivel de protección mucho mayor que el enfoque frecuentista contra la posibilidad de “ver significación en hallazgos procedentes de la investigación científica que son enteramente espurios”.

Otro ejemplo de aplicación del p-valor: Si en la H_1 estamos haciendo un contraste bilateral para comparar las medias de dos poblaciones normales:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Se calculará como: $p - valor = P(U \geq u_{exp} \text{ bajo } H_0)$

La prueba t de dos muestras se utiliza para comparar directamente la media de dos grupos (X e Y). Se requiere que las mediciones en los dos grupos sean estadísticamente independientes. La hipótesis nula establece que las medias de dos grupos son iguales, o equivalentemente, la diferencia de las medias es cero:

$$H_0: \mu(X) = \mu(Y), \quad \text{or} \quad \mu(X) - \mu(Y) = 0$$

$$H_1: \mu(X) \neq \mu(Y), \quad \text{or} \quad \mu(X) - \mu(Y) \neq 0$$

Cuando se desconoce la desviación estándar, podemos usar:

$$t = \frac{\bar{X} - \bar{Y}}{S_x^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Una puntuación t grande indica que los grupos son diferentes. Una puntuación t pequeña indica que los grupos son similares.

#A t-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.
 #When the #scaling term is unknown and is replaced by an estimate based on the #data, the test #statistics (under certain conditions) follow a Student's #t distribution. The t-test can #be used, for example, to determine if #two sets of data are significantly different from #each other.
 #(see wikipedia https://en.wikipedia.org/wiki/Student%27s_t-test)
 #[Wikipedia] (https://en.wikipedia.org/wiki/Student%27s_t-test)

#Please read the introduction from:
 #[Introduction to t-test] #<<http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-t-tests-t-values-and-t-distributions>>

#see also at <https://www.investopedia.com/terms/t/t-test.asp>

#More examples and power calculation at:
<https://statistics.berkeley.edu/computing/r-t-tests>

1. Aquí tenemos datos sobre el tamaño del genoma (medido en picogramos de ADN por célula haploide) en dos grandes grupos de crustáceos (Decápodos e Isópodos). La causa de la variación en el tamaño del genoma ha sido un misterio durante mucho tiempo; Usaremos estos datos para responder la pregunta biológica de si algunos grupos de crustáceos tienen diferentes tamaños de genoma que otros.

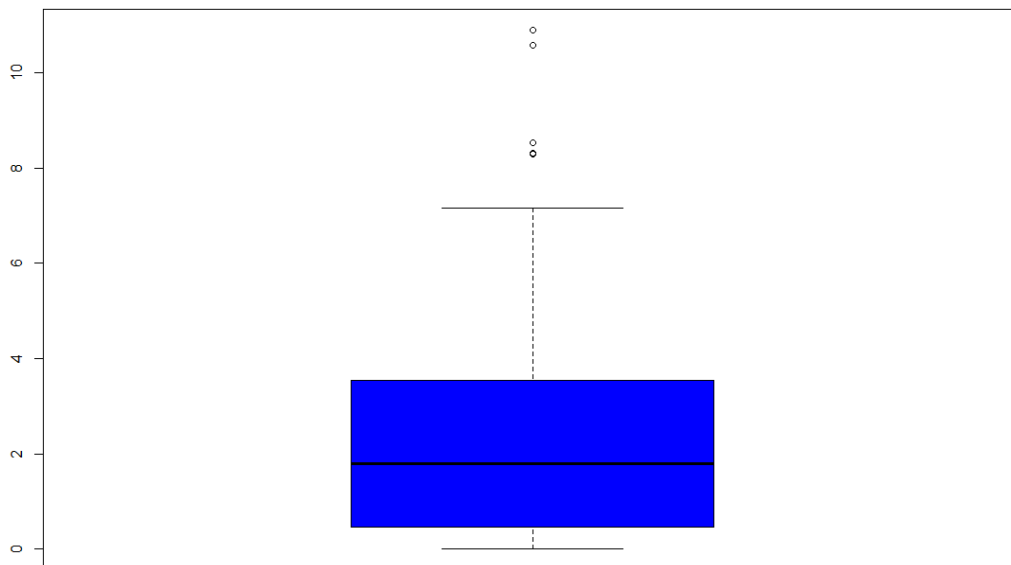
Primero debemos observar los datos, cargar el archivo en R y explorar gráficamente la dispersión y normalidad de todo el conjunto de datos. Mirando los histogramas, ¿crees que los datos son normales?

```
V1<-c("Decapods", "Decapods", "Decapods", "Decapods", "Decapods", "Decapods",
      "Decapods", "Decapods", "Decapods", "Decapods", "Decapods", "Decapods",
      "Decapods", "Decapods", "Decapods", "Decapods", "Decapods", "Decapods",
      "Decapods", "Decapods", "Decapods", "Decapods", "Decapods", "Decapods",
      "Decapods", "Decapods", "Isopods", "Isopods", "Isopods", "Isopods",
      "Isopods", "Isopods", "Isopods", "Isopods", "Isopods", "Isopods",
```

```

    "Isopods", "Isopods" , "Isopods", "Isopods", "Isopods", "Isopods", "Isopods",
    "Isopods", "Isopods", "Isopods", "Isopods", "Isopods",
    "Isopods", "Isopods", "Isopods" , "Isopods", "Isopods")
V2<-c(1.60, 1.65, 1.80, 1.90, 1.94, 0.27, 0.44, 2.66, 2.78, 0.01, 2.34, 1.50, 3.55,
3.66, 2.70, 2.75, 4.84,
    2.23, 6.20, 8.29, 8.53, 10.58, 6.56, 7.16, 8.30, 6.47, 10.89, 0.46, 0.70, 0.85, 1.47,
3.13, 0.19, 0.21,
    0.22, 0.23, 0.28, 0.30, 0.40, 0.47, 0.63, 0.87, 2.77, 2.91, 0.25, 0.26, 0.58, 0.97,
1.63, 1.77, 2.67,
    5.45, 6.81, 0.71)
genome_size<-data.frame(V1=as.factor(V1),V2=V2)
genome_size
#BOX PLOT IS A GOOD SOLUTION:
boxplot(genome_size[,2], col="blue")

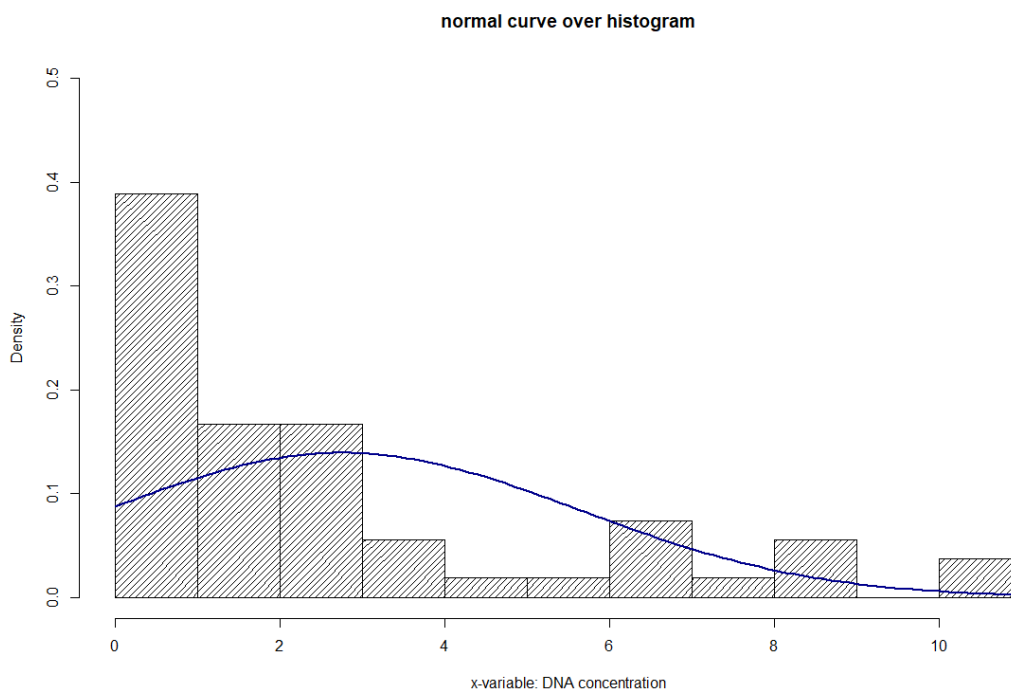
```



*#A good interpretation of a box-plot can be seen in the link
[#<https://www.r-graph-gallery.com/boxplot.html>](https://www.r-graph-gallery.com/boxplot.html)*

Otra solución para la variable [ADN] es representar la distribución empírica de la variable [ADN] mediante un histograma con una curva normal sobre:

```
#SOLUTION:
g = genome_size[,2]
m<-mean(g)
std<-sqrt(var(g))
hist(g, density=20, breaks=10, prob=T,
     xlab="x-variable: DNA concentration", ylim=c(0, 0.5),
     main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



- Si las varianzas no son similares entre los grupos y la variable no tiene una distribución normal, no podemos usar la prueba t de Student directamente. Para ello intentamos transformar los datos para que encajen en una distribución normal. Aplicaremos la transformación \log_{10} a nuestros datos. Utilice la función `log10()`. Calcule la media y la varianza de los datos recién transformados. ¿Son las variaciones más similares ahora?

```
#SOLUTION
genome_size$log10 <- log10(genome_size[,2])

mean(genome_size[genome_size[,1]=="Decapods",3])
## [1] 0.4020461

mean(genome_size[genome_size[,1]=="Isopods",3])
## [1] -0.1066846
```

```

var(genome_size[genome_size[,1]=="Decapods",3])
## [1] 0.3807424
var(genome_size[genome_size[,1]=="Isopods",3])
## [1] 0.2129233
# the means are different but the variances are more similar

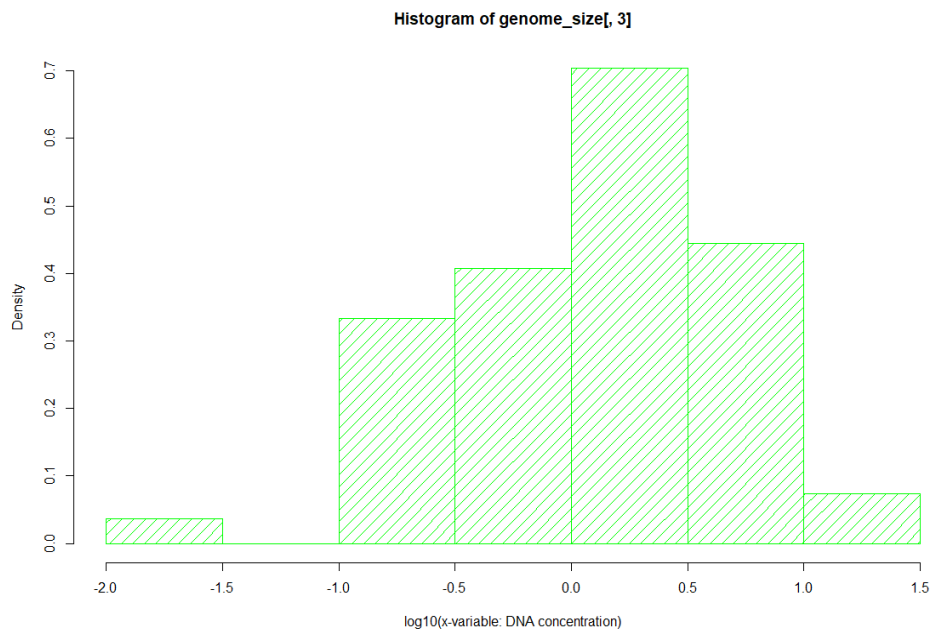
```

3. Ahora trace el histograma de los datos transformados. ¿Se ven casi normales?

```

#SOLUTION
hist(genome_size[,3], density=10, breaks=10, prob=T,
     xlab="log10(x-variable: DNA concentration)", ylim=c(0, 0.7),
     col="green")

```

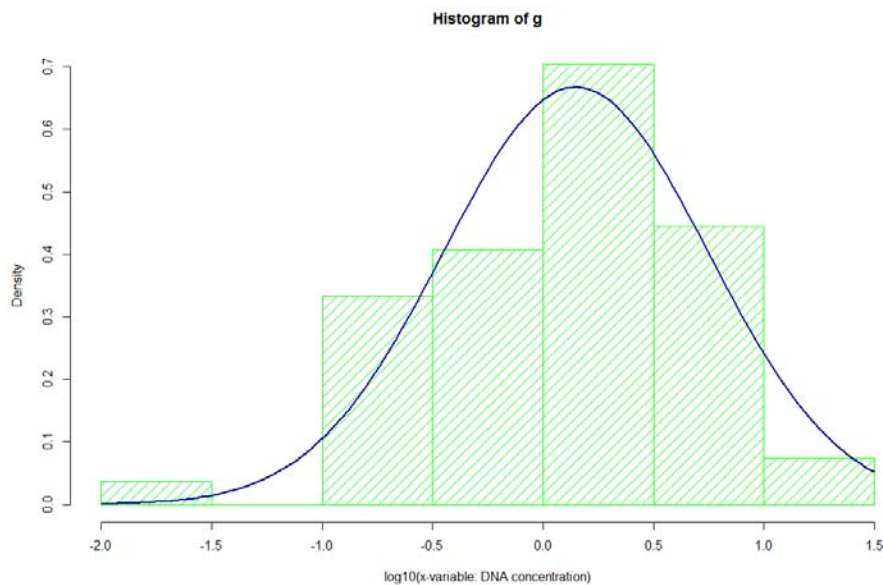


4. Verifique si los datos transformados siguen una distribución normal usando un método gráfico.

```

#SOLUTION
g = genome_size[,3]
m<-mean(g)
std<-sqrt(var(g))
hist(g, density=10, breaks=10, prob=T,
     xlab="log10(x-variable: DNA concentration)", ylim=c(0, 0.7),
     col="green")
curve(dnorm(x, mean=m, sd=std),
     col="darkblue", lwd=2, add=TRUE, yaxt="n")

```

now the variable it seems to be normal

- Después de transformar los datos ($\log_{10}()$), ahora podemos aplicar la prueba t de Student para responder a la pregunta: ¿Ambos grupos tienen el mismo tamaño medio del genoma? ¿Cuál es el valor del estadístico t? ¿Y el valor p?

```
#SOLUTION
result_StudentT <- t.test(genome_size[genome_size[,1]=="Decapods",3],
genome_size[genome_size[,1]=="Isopods",3],var.equal = TRUE)
result_StudentT

##
## Two Sample t-test
##
## data: genome_size[genome_size[, 1] == "Decapods", 3] and
genome_size[genome_size[, 1] == "Isopods", 3]
## t = 3.4308, df = 52, p-value = 0.001187
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2111807 0.8062808
## sample estimates:
## mean of x mean of y
##  0.4020461 -0.1066846

tvalue <- result_StudentT$statistic
pvalue <- result_StudentT$p.value
```

Veamos a continuación unos ejemplos de planificación de experimentos con diferentes tipos de supuestos, donde es necesario trabajar con alfa y con la potencia estadística.



PLANIFICACION DE
EXPERIMENTOS Y ESTUDIOS

2.6.2- Estudios multiobjetivo (Y), significación o potencia estadística. Un ejemplo práctico. ¿Cómo establecer la regla de decisión?

Basado parcialmente en Jordi Cortés et al. (2014).

En muchas situaciones el experimentador debe ser crítico y elegir cómo plantear el experimento y cómo analizarlo. Así imaginemos que disponemos de k variables medidas en un experimento (variables de respuesta, Y_1, Y_2, Y_3 , por ejemplo diversos biomarcadores sanguíneos o moleculares)

Imaginemos un ensayo clínico donde sólo se desea comparar un nuevo tratamiento respecto (T) a un tratamiento estándar (S) y de sea verificar que el primero tiene una eficacia superior al otro respecto a k variables de respuesta (por ejemplo $k=3$ variables de respuesta: [biomarcador 1] [biomarcador 2] [biomarcador 3]).

¿Cómo deben plantearse los análisis para las K variables?

Así antes de iniciar los análisis y el propio experimento debemos definir las reglas de decisión.

Para estudiar el efecto del tratamiento T frente al S sobre tres variables de respuesta de interés Y_1, Y_2, Y_3 , se formulan 3 hipótesis y por tanto se hacen tres contrastes, cada uno con un riesgo *alfa*. Así formularíamos una hipótesis como:

$$\begin{cases} H_o : \mu_{Y1_T} = \mu_{Y1_S} \\ H_1 : \mu_{Y1_T} > \mu_{Y1_S} \end{cases}$$
$$\begin{cases} H_o : \mu_{Y2_T} = \mu_{Y2_S} \\ H_1 : \mu_{Y2_T} > \mu_{Y2_S} \end{cases}$$
$$\begin{cases} H_o : \mu_{Y3_T} = \mu_{Y3_S} \\ H_1 : \mu_{Y3_T} > \mu_{Y3_S} \end{cases}$$

La regla de decisión podría funcionar por intersección (positivo si las 3 pruebas fueran positivas) o por unión (positivo si cualquiera fuera positiva). Es decir, en el primer caso, se autorizaría el producto sólo si las 3 pruebas resultan significativas; en el segundo, bastaría con que lo fuera una de ellas. En el primer caso, el estudio pierde

potencia (hay menos opciones de alcanzar el mercado de las que tendría una sola prueba); en el segundo, sobre-consume alfa (cada contraste “gasta” alfa).

Caso 1: Regla de decisión por intersección de las tres hipótesis. Pérdida de potencia estadística.

Se considera que el medicamento es eficaz si las tres pruebas son positivas (rechazo de hipótesis nula y aceptación de alternativa).

Vamos a calcular la potencia estadística global (1-probabilidad (error tipo II) si todas las pruebas han de ser significativas.

Supongamos que la probabilidad de error de tipo II (β) = 0.2 (algo muy habitual cuando se fija un tamaño muestral).

```
#Suponemos que Las variables de respuesta Y1, Y2 y Y3 son independientes  
  
#se puede calcular la pérdida de potencia si el proceso exige que las 3 pruebas (Y1, Y2 y Y3) sean significativas (p<alfa). Tomando una potencia del 80% para cada prueba, que equivale a un Beta=0.2, la potencia global es:  
  
# 1- beta.global = (1-beta.individual)^3  
  
potencia.global = (1-0.2)^3  
  
1-potencia.global #probabilidad(error tipo II global) = 0.488  
## [1] 0.488
```

La potencia estadística de las 3 pruebas estadísticas simultáneas y significativas ($1-\beta'$)=0.488

Es decir, que si la intervención tuviera el efecto especificado (hemos rechazado las 3 hipótesis nulas, para Y1, Y2, Y3) con una probabilidad de $1-\beta' = 0.488$ (potencia de test o potencia estadística), las probabilidades de falso negativo serían de un 50% (aprox) y las probabilidades de siendo falsas las hipótesis nulas y de haber rechazado aquellas son también del 50% (aprox)

Ningún laboratorio o empresa querría fallar en 5 de cada 10 intervenciones eficaces

Conclusión: Se pierde potencia estadística (1-beta) si se requiere que todas las pruebas para las variables de respuesta sean significativas.

Caso 2: Regla de decisión por unión de las tres hipótesis.

Imaginemos ahora que sólo una de las pruebas ha de ser positiva (significativa). Si el criterio para autorizar la intervención solo requiere que una de las 3 pruebas fuera significativa, se pierde el control del riesgo alfa y la probabilidad de autorizar una intervención no eficaz es mayor del 5% (alfa), ya que asumimos este riesgo en 3 ocasiones (Y1, Y2, Y3).

Observemos como calcular en R la probabilidad de tipo I global (alfa global) para las 3 pruebas a la vez:

```
#Suponemos que las pruebas Y1, Y2 y Y3 son idenpendientes
alfa.individual <- 0.05
n.pruebas<- 3
alfa.global = 1-(1-alfa.individual)^n.pruebas
alfa.global #probabilidad(error tipo I global) = 0.142625
## [1] 0.1426
```

Si hemos definido alfa como la probabilidad del error de tipo I: siendo cierta H0, rechazamos ésta y aceptamos H1, nos conduce que si en múltiples pruebas para aceptar la superioridad del nuevo tratamiento sólo una de las tres pruebas es significativa obtenemos un alfa global del 14.26%, no significa que un 14% de veces rechazaríamos la hipótesis nula (medicamento eficaz) por azar y no porque realmente es eficaz. Demasiado alto para una prueba clínica, y en general no aceptable, ya que como es aceptado por consenso alfa = 5%.

Conclusión: si la intervención no tuviera efecto, un 14% de estudios conducirían a administrar el nuevo medicamento T a los pacientes: ninguna agencia de regulación aceptaría que 1 de cada 7 intervenciones no eficaces terminara siendo aconsejado a los pacientes.

Imaginemos que ahora son 7 variables de respuesta Y1,...,Y7:

```
#Suponemos que las pruebas Y1,... ,Y7 son idenpendientes
alfa.individual <- 0.05
n.pruebas<- 7
alfa.global = 1-(1-alfa.individual)^n.pruebas
```

```
alfa.global #probabilidad(error tipo I global) = 0.4012631  
  
## [1] 0.4012631
```

Llegamos hasta un 40% de probabilidad de error de tipo I, es lo que se conoce como inflamiento del error de tipo I o el fenómeno de la multiplicidad de pruebas.

Otra aplicación de alfa y beta es el cálculo del tamaño muestral de los experimentos.

2.7. Potencia estadística y cálculo del tamaño muestral en diseño de experimentos

Realizar análisis de potencia estadística y estimar el tamaño de la muestra es un aspecto importante del diseño experimental, ya que, sin estos cálculos, el tamaño de la muestra puede ser demasiado alto o demasiado bajo. En diseños experimentales tales como los ensayos clínicos (muy regulados legalmente) es imprescindible la justificación del diseño experimental.

Si el tamaño de la muestra es demasiado bajo, el experimento carecerá de la precisión para proporcionar respuestas confiables a las preguntas que está investigando. Si el tamaño de la muestra es demasiado grande, el tiempo y los recursos se desperdiciarán, a menudo para una ganancia mínima.

2.7.1. ¿Cómo se calcula el tamaño muestral de un estudio? Ideas generales

En algunos programas informáticos de análisis de potencia (por ejemplo en R), hay disponibles una serie de herramientas gráficas y analíticas para permitir una evaluación precisa de los factores que afectan la potencia y el tamaño de la muestra en muchos de los análisis estadísticos más comunes. Esta información puede ser crucial para el diseño de un estudio que sea rentable y científicamente útil.

A menudo prevalecen criterios económicos, pero suelen justificarse mediante un enfoque estadístico para lo que existen dos tipos de enfoques:

- Tamaño fijo del experimento: determinado antes de iniciar el estudio, a partir de cierto criterio de optimalidad (ejemplo: comparación de dos tratamientos, uno nuevo y un estándar, en el ámbito biológico).

- Diseños secuenciales: repetidamente, un nuevo dato (o un grupo de datos) seguida de un análisis que es concluyente y conduce a una decisión estadística final (y paramos el muestreo) o bien a la necesidad de obtener un nuevo dato. Este tipo de diseños se utilizan ampliamente en el sector clínico (ej: oncología)

Cálculo del tamaño muestral en el caso de comparación de dos medias en el diseño experimental al que aplicaríamos la t de Student

Este campo es muy amplio, pero como ejemplo siempre utilizamos el de la comparación de dos tratamientos, tomando muestras seleccionadas al azar y en el caso de varianzas desconocidas, donde debemos conocer cual es el mínimo número de individuos que debemos seleccionar para detectar diferencias estadísticamente significativas entre las medias de las poblaciones del tratamiento 1 y el tratamiento 2:

$$\begin{cases} H_0 : \mu_A - \mu_B = 0 \\ H_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

Se trata de un diseño totalmente aleatorizado, estudio de un factor con sólo dos niveles. (P.e. Queremos comparar dos tratamientos respecto de cierta variable fisiológica, asignamos al azar individuos cada tratamiento, posteriormente haremos una prueba t de comparación de medias.)

Cuántos individuos debemos asignar a cada grupo?

Supongamos por un momento que es verdad que las medias son diferentes (H_1 es cierta). En este caso la decisión correcta es rechazar H_0 . Si no lo hacemos se ha producido un error de tipo II.

La probabilidad de que se produzca este error depende de los tamaños muestrales, de la verdadera diferencia de medias, de σ , ...

El cálculo del tamaño muestral se obtiene mediante la determinación de la llamada curva característica de operación del test t, que es una función específica del tamaño muestral por grupo (n):

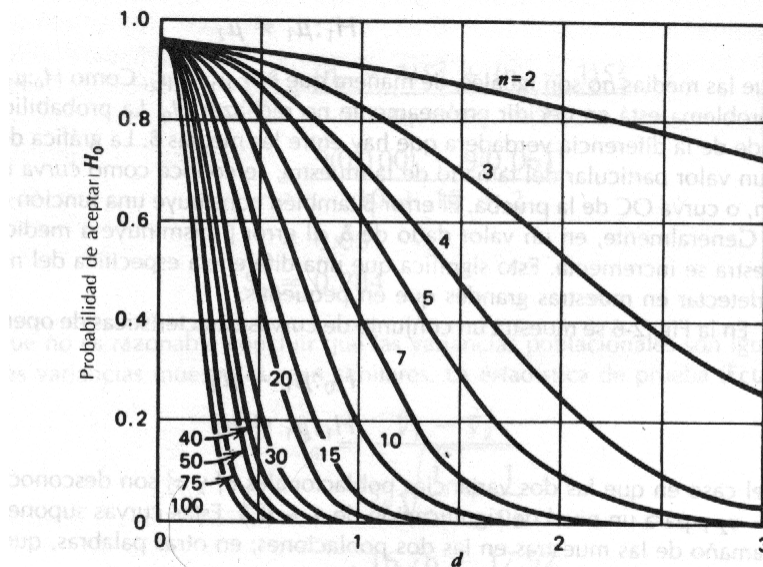


Figura 2-9. Curvas características de operación para pruebas t bilaterales con $\alpha = 0.05$. (Reproducida con permiso de "Operating characteristics for the common tests of significance", C.L. Ferris, F.E. Grubs y C.L. Weaver, *Annals of Mathematical Statistics*, junio 1946)

Existe una curva para cada $n^*=2n-1$ (suposem $n_1=n_2=n$), donde las ordenadas son la probabilidad d'error de tipus II (β) y las abscisas (X) son el valor de d (diferencia media ponderada por variabilidad, que es proporcional al valor del parámetro δ que utilizan los programas informáticos actuales).

$$d = \frac{|\mu_A - \mu_B|}{2\sigma} = \frac{|\varphi|}{2\sigma}$$

$$\varphi = |\mu_A - \mu_B|$$

Para el cálculo del tamaño muestral necesitamos valores de los parámetros conocidos, por ello o bien los obtenemos de estudios anteriores o realizamos una pequeña prueba piloto a fin de calcularlos.

Decidimos el valor de φ a partir del cual "vale la pena" detectar diferencias con "seguridad"

- p.e. suposem que és $\delta=0.5$ a l'exemple

Decidimos que significa "seguridad"

- p.e. suponemos que es 5% de prob. de error II

Como σ es desconocido, procuramos proponer un valor (p.e. estimación) p.e. de prueba piloto previa (ejemplo $\hat{\sigma} = 0.25$)

Todo ello nos lleva a:
$$d = \frac{0.5}{2 \times 0.25} = 1$$

Si observem las curvas características con $n^*=16$ permite un 5% de prob. De error de tipo II,

Por tanto $16=2n-1 \Rightarrow n=9$: habría que asignar 9 individuos en cada grupo experimental (tratamiento).

Actualmente los diseños son cada vez más complejos y a veces con grupos no balanceados, existen softwares específicos que permiten el cálculo del tamaño muestral a partir de las curvas características, supuestos los parámetros de interés conocidos.

Un paquete de R específico para el cálculo de potencia estadística y el cálculo del tamaño muestral es `library(pwr)`.

Ver más información en este link (Power analysis and sample size in r): <https://www.statmethods.net/stats/power.html>

Otro link interesante puede ser en:

<http://utstat.toronto.edu/~nathan/teaching/STA305/classnotes/week3/classnotes-week3.html>

El cálculo de valores específicos de potencia, tamaño de muestra o tamaño del efecto (delta= δ)

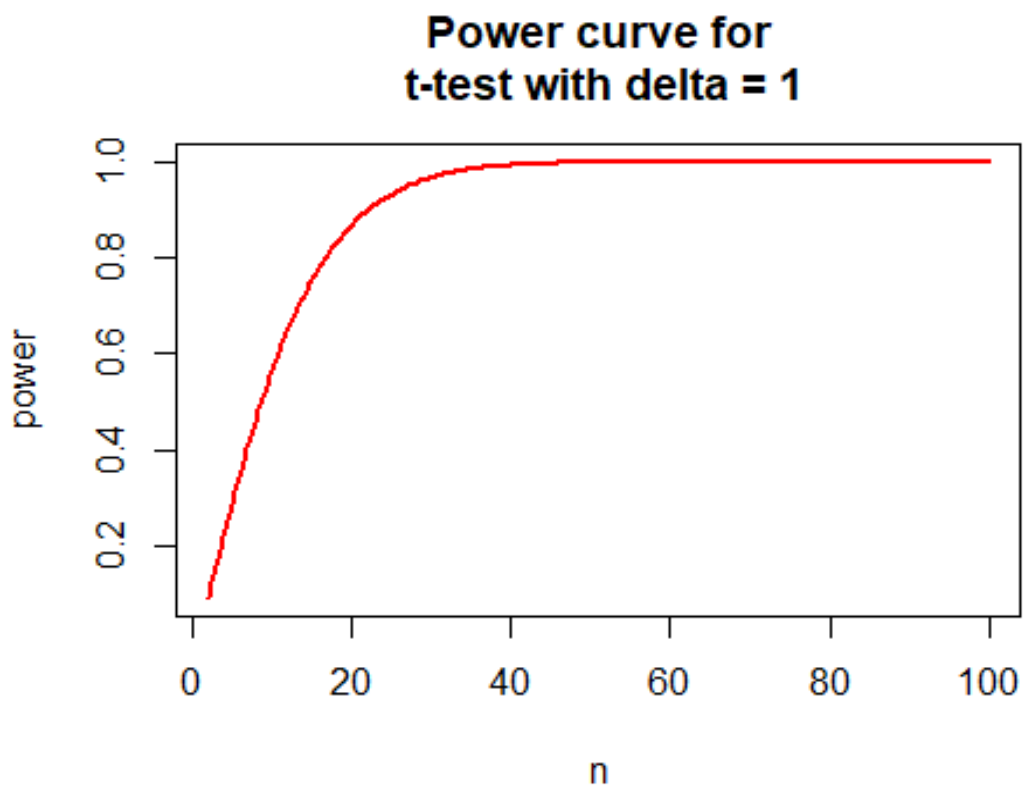
Recordemos que si tenemos dos poblaciones a comparar (ej: 2 grupos experimentales donde se toma una medida absolutamente continua) delta (δ) es una relación entre las medias de las poblaciones y la varianza conjunta de ambas, así:

$$\delta = \frac{|mean_1 - mean_2|}{\sigma}$$

Delta (δ) puede ser un parámetro esclarecedor con respecto a las restricciones estadísticas en el diseño y análisis experimental. Pero con frecuencia los gráficos presentan mejor los resultados. Aquí mostramos cómo dibujar curvas de potencia para un t-test, tamaño de muestra y tamaño de efecto usando las funciones anteriores

en R. Veamos un ejemplo donde se presenta una curva de la potencia estadística con respecto al tamaño muestral para $\delta=1$:

```
#Power curve for t-test with delta = 1 (size effect)
nvals <- seq(2, 100, length.out=200)
powvals <- sapply(nvals, function (x) power.t.test(n=x,
delta=1)$power)
plot(nvals, powvals, xlab="n", ylab="power",
main="Power curve for\n t-test with delta = 1",
lwd=2, col="red", type="l")
```



En la gráfica anterior podemos observar que se alcanza la saturación aproximadamente para un tamaño muestral de $n=30$ para un t-test.

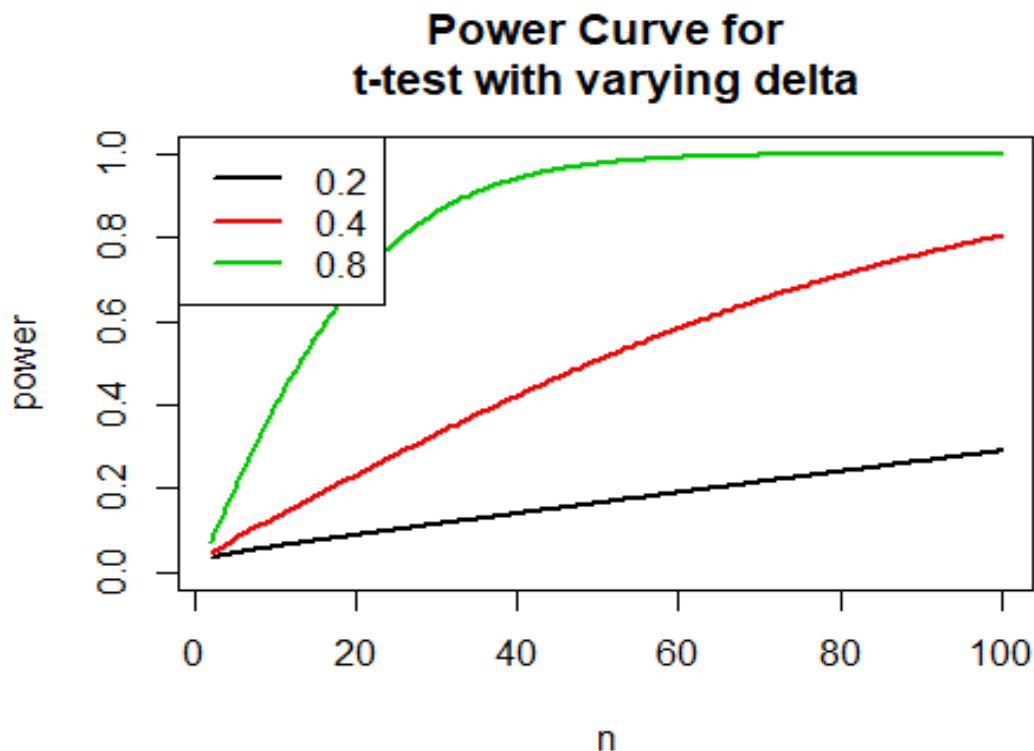
En la gráfica anterior de nuestro tamaño del efecto (con $\delta = 1$), también podemos modificar el δ para ver el efecto del tamaño del efecto y el tamaño de la muestra en la potencia estadística:

```
#Power Curve for\n t-test with varying delta
deltas <- c(0.2, 0.4, 0.8)
plot(nvals, seq(0,1, length.out=length(nvals)), xlab="n",
```

```

ylab="power",
main="Power Curve for\nt-test with varying delta", type="n")
for (i in 1:3) {
powvals <- sapply(nvals, function (x) power.t.test(n=x,
delta=deltas[i])$power)
lines(nvals, powvals, lwd=2, col=i)
}
legend("topleft", lwd=2, col=1:3, legend=c("0.2", "0.4", "0.8"))

```



2.7.2. Cálculo del tamaño muestral en pruebas T-test para comparar medias de dos poblaciones

A continuación, vamos a ver cómo calcular el tamaño muestral (n) en los test de hipótesis del tipo T test, mediante el uso de diversas librerías como `library(pwr)`, utilizando los conceptos anteriores y un ejemplo numérico concreto:

```

library(pwr)

#calculate the sample size in our "classical example T-test 2
populations". H1: "two.sided"
yA <- c(0.92, 1.29, 1, 1.5, 1.25, 1.54, 1.26, 1.71, 1.28) #first
population

```

```

yB <- c(1.16, 1.43, 1.8, 1.5, 1.57, 1.75) #second population
mean(yA); sd(yA, na.rm = FALSE)

## [1] 1.305556
## [1] 0.2511031
mean(yB); sd(yB, na.rm = FALSE)

## [1] 1.535
## [1] 0.2326156
yA_B <- c(0.92, 1.29, 1, 1.5, 1.25, 1.54, 1.26, 1.71, 1.28, 1.16, 1.43,
1.8, 1.5, 1.57, 1.75)
sd(yA_B, na.rm = FALSE) #join sdv

## [1] 0.2624736

# T test two populations in the case of H1: "two.sided"
tResult <- t.test(yA, yB, var.equal = TRUE)
tResult

#This is the t-test result-----:

##
## Two Sample t-test
##
## data: yA and yB
## t = -1.783, df = 13, p-value = 0.09793
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.50744650 0.04855761
## sample estimates:
## mean of x mean of y
## 1.305556 1.535000

mean1 <- mean(yA)
mean2 <- mean(yB)
sigma <- sd(yA_B, na.rm = FALSE)
delta= (mean1-mean2) / sigma
delta #This is the estimated delta value

## [1] -0.8741621

# Two-sample t test power calculation:

pwr.t.test(d=delta,power=0.8,sig.level=0.05,type="two.sample",alternative
="two.sided")

## In this case the use delta, power and significance level to obtain
sample size (n):

```

```

##      Two-sample t test power calculation
##
##              n = 21.54634
##              d = 0.8741621
##      sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
##
## NOTE: n is number in *each* group

# So we need n=22 samples/groups (in total 44) to detect a difference od
delta = -0.87, with a power = 0.8 and a significance level (alfa = 0.05)

# On the other hand we can calculate power in function of delta, n and
significance level

# Two-sample t test power calculation:

pwr.t.test(d=delta,n=9,sig.level=0.05,type="two.sample",alternative="two.
sided")

##      Two-sample t test power calculation
##
##              n = 9
##              d = 0.8741621
##      sig.level = 0.05
##              power = 0.41426
##      alternative = two.sided
##
## NOTE: n is number in *each* group

# In consequence if we use a n=9 (by group, in total 18) and an alfa =
0.05 we expect a low statistical power = 0.4142 to detect differences.

```

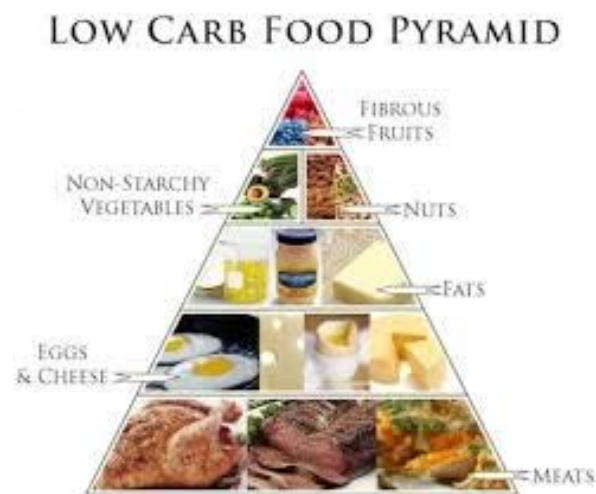
Más ejemplos de tamaño muestral y explicaciones en el siguiente link:

<https://stats.idre.ucla.edu/r/dae/power-analysis-for-two-group-independent-sample-t-test/>)

2.7.3 Ejemplo 1 cálculo del tamaño muestral. ¿Es eficaz un programa de pérdida de peso?

Una empresa comercializa un programa de pérdida de peso de ocho semanas de duración y afirma que al final del programa, en promedio, un participante habrá perdido 5 libras (1 kg =libra /2.2046). Por otro lado, se ha estudiado el programa de pérdida de peso y crees que su programa es científicamente erróneo y no debería

funcionar en absoluto. Se desea probar la hipótesis de que el programa de pérdida de peso no ayuda a las personas a perder peso. El plan es obtener una muestra aleatoria de personas y ponerlas en el programa de pérdida de peso. Se medirá su peso al comienzo del programa y luego se medirá su peso nuevamente al final del programa. Al final del programa se medirá la diferencia de peso basal y final. Según algunas investigaciones anteriores, creemos que la desviación estándar (sd) de la diferencia de peso durante ocho semanas será de 5 libras. Ahora desea saber cuántas personas debe inscribirse en el programa para probar su hipótesis.



Pirámide de los alimentos (Imagen de: http://cubasi.cu/media/k2/items/cache/fefc9e06ac64524d4de5b8eb1099062f_L.jpg)

Este ejemplo procede de las librerías utilizadas en R para el cálculo de la potencia estadística y cálculo del tamaño muestral:

- pwr – R package for power analysis.
- pwr.t.test – Sample size and power determination.

Para el cálculo del ejemplo de la pérdida de peso, podemos establecer la potencia en diferentes niveles y calcular el tamaño de la muestra para cada nivel de potencia estadística. Por ejemplo, podemos configurar la potencia para que esté en el nivel .80 al principio, y luego restablecerla para que esté en el nivel .85, y así sucesivamente. Primero, especificamos las dos medias, la media para la hipótesis nula y la media para la hipótesis alternativa. Luego especificamos la desviación estándar para la diferencia en las medias. El nivel de significación predeterminado (nivel alfa) se establece en .05, por lo que no lo especificaremos para las ejecuciones iniciales. Por último, le decimos a R que estamos realizando una prueba t de muestra apareada.

Veamos como se realiza esta prueba con un ejemplo concreto, donde se calculará delta como $d=(0-5)/5$; así si calculamos el tamaño muestral suficiente para detectar diferencias entre el peso basal final y que este sea diferente de 0, para unos niveles de potencia estadística de 0.8, 0.085 y 0.9, vemos que $n = 10, 11$ y 13 respectivamente:

```
pwr.t.test(d=(0-5)/5,power=0.8,sig.level=0.05,type="paired",alternative="two.sided")

##
##      Paired t test power calculation
##
##           n = 9.93785
##           d = 1
##      sig.level = 0.05
##           power = 0.8
##      alternative = two.sided
##
## NOTE: n is number of *pairs*

pwr.t.test(d=(0-5)/5,power=0.85,sig.level=0.05,type="paired",alternative="two.sided")

##
##      Paired t test power calculation
##
##           n = 11.06279
##           d = 1
##      sig.level = 0.05
##           power = 0.85
##      alternative = two.sided
##
## NOTE: n is number of *pairs*

pwr.t.test(d=(0-5)/5,power=0.9,sig.level=0.05,type="paired",alternative="two.sided")

##
##      Paired t test power calculation
##
##           n = 12.58546
##           d = 1
##      sig.level = 0.05
##           power = 0.9
##      alternative = two.sided
##
## NOTE: n is number of *pairs*
```

Ahora vamos a cambiar el nivel de significación a 0.01, con una potencia estadística de 0.85, obteniendo $n=17$:


```

pwr.t.test(d=(0-
5)/5,power=0.85,sig.level=0.01,type="paired",alternative="two.sided")

##
##      Paired t test power calculation
##
##              n = 16.4615
##              d = 1
##      sig.level = 0.01
##      power     = 0.85
##      alternative = two.sided
##
## NOTE: n is number of *pairs*

```

Como puede ver, el tamaño de la muestra aumenta de 11 a 17 para una potencia especificada de .85 cuando el valor alfa cae de .05 a .01. Esto significa que si queremos que nuestra prueba sea más confiable, es decir, que no rechacemos la hipótesis nula en caso de que sea cierta, necesitaremos un tamaño de muestra más grande.

Si pensamos que queremos un alfa más bajo en el nivel 0.01 y una potencia alta en .90, necesitaríamos 15 sujetos como se muestra a continuación. Recuerda que esto está bajo el supuesto de normalidad. Si la distribución no es normal, entonces 15 sujetos no son, en general, suficientes para esta prueba t.

```

pwr.t.test(d=(0-
5)/5,power=0.9,sig.level=0.01,type="paired",alternative="two.sided")

##
##      Paired t test power calculation
##
##              n = 18.30346
##              d = 1
##      sig.level = 0.01
##      power     = 0.9
##      alternative = two.sided
##
## NOTE: n is number of *pairs*

```

2.7.4. Ejemplo 2 cálculo del tamaño muestral en un diseño experimental: destreza manual.

Un investigador de del comportamiento humano quiere estudiar la diferencia entre la mano dominante y la mano no dominante en términos de destreza manual. Se diseña un experimento en el que cada sujeto colocaría 10 cuentas pequeñas sobre la mesa en un tazón, una vez con la mano dominante y otra con la mano no dominante. Se mide el número de segundos necesarios en cada ronda para completar la tarea. También se ha decidido que el orden en que se miden las dos manos debe ser contado

balanceadamente (igual tamaño muestral). Así se espera que la diferencia promedio en el tiempo sea de 5 segundos con la mano dominante siendo más eficiente con una desviación estándar de 10 (sd). Recopila sus datos en una muestra de 35 sujetos. La pregunta es, ¿cuál es el poder estadístico del diseño con $n= 35$ para detectar la diferencia en la magnitud (diferencia de medias) de 5 segundos?

Se calculará delta como $d=(0-5)/10$



Destreza manual (Imagen de:
<https://www.mumuchu.com/media/catalog/category/akros-carrusel-destreza-manual.jpg>)

En este ejemplo, el investigador ya ha recopilado datos sobre 35 sujetos. ¿Cuánto poder estadístico tiene su diseño para detectar la diferencia de 5 segundos con una desviación estándar de 10 segundos?

Nuevamente usamos la función **pwr.t.test()** para calcular la potencia.

Indicaremos en la función **pwr.t.test()** la primera media como 0 y la segunda media como 5, ya que lo único que sabemos es que la diferencia de las dos medias es de 5 segundos. En términos de hipótesis, esta es la misma manera de decir que la hipótesis nula es que la diferencia es cero, y la hipótesis alternativa es que la diferencia media es 5. Luego ingresamos la desviación estándar para la diferencia y el número de sujetos. Nuevamente, especificamos la opción `type="paired"`, ya que el diseño es una prueba t de muestras apareadas.

```
pwr.t.test(d=(0-5)/10,n=35,sig.level=0.01,type="paired",alternative="two.sided")  
##  
## Paired t test power calculation
```

```
##
##           n = 35
##           d = 0.5
##       sig.level = 0.01
##           power = 0.5939348
##       alternative = two.sided
##
## NOTE: n is number of *pairs*
```

Esto significa que el investigador detectaría la diferencia de 5 segundos aproximadamente el 59 por ciento del tiempo. Note que hicimos esto como una prueba de dos caras. Dado que se cree que nuestra mano dominante siempre es mejor que la mano no dominante, el investigador en realidad podría realizar una prueba de una cola (unilateral). Ahora, recalcularemos la potencia para la prueba t de muestra emparejada de una cola (unilateral)..

```
pwr.t.test(d=(0-5)/10,n=35,sig.level=0.01,type="paired",alternative="less")
```

```
##
##       Paired t test power calculation
##
##           n = 35
##           d = -0.5
##       sig.level = 0.01
##           power = 0.6961194
##       alternative = less
##
## NOTE: n is number of *pairs*
```

Probablemente haya notado que la forma de realizar el análisis de potencia para la prueba t de muestras apareadas es la misma que para la prueba t de una muestra (one t test). Esto se debe al hecho de que en la prueba t de muestra pareada calculamos la diferencia en las dos puntuaciones para cada sujeto y luego calculamos la media y la desviación estándar de las diferencias. Esto convierte la prueba t de muestra apareada en una prueba t de una muestra.

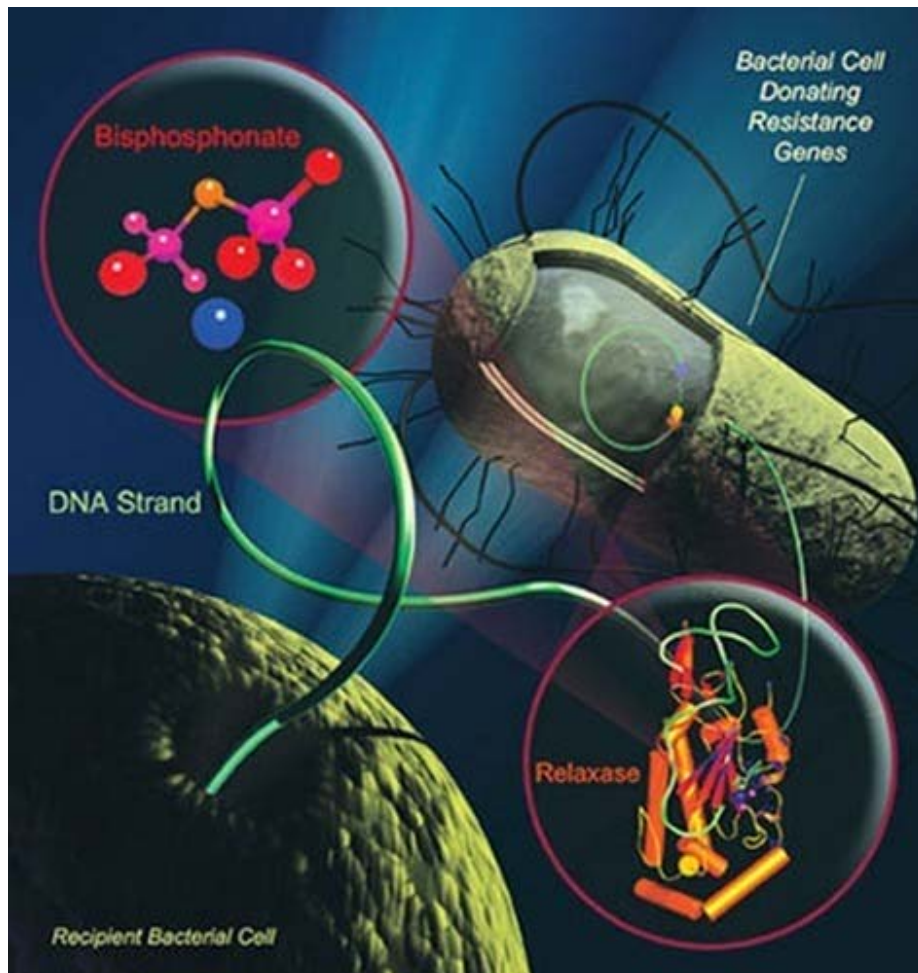
El otro supuesto técnico es el supuesto de normalidad. Si la distribución está sesgada, entonces un tamaño de muestra pequeño puede no tener la potencia mostrada en los resultados, porque el valor en los resultados se calcula utilizando el método basado en el supuesto de normalidad. Puede que ni siquiera sea una buena idea hacer una prueba t en una muestra pequeña para empezar.

Lo que realmente necesitamos saber es la diferencia entre los dos medias, no los valores individuales. De hecho, lo que realmente importa es la diferencia de los medios sobre la desviación estándar. A esto le llamamos el tamaño del efecto (delta).

Por lo general, no es una tarea fácil determinar el tamaño del efecto. Suele proceder del estudio de la literatura existente o de estudios piloto. Una buena estimación del tamaño del efecto es la clave para un análisis de potencia exitoso.

2.7.5. Cálculo del tamaño muestral en el caso de la comparación de dos proporciones

Existen muchos ejemplos de comparación de dos proporciones, por ejemplo en una determinada muestra metagenómica ambiental (suelos, aguas, heces, etc) la proporción entre dos condiciones (ej: tipos de tratamiento) de un determinado taxón. Por ello es necesario conocer este tipo de cálculo del tamaño muestral.



Metagenómica (Imagen procedente de: https://4.bp.blogspot.com/-chtGZyvf8JQ/WTtg3ZZizrI/AAAAAAAAAIQ/4BjSgmNOyi4Oo8bt7BZmXVs5Hue8YTE0QCLcB/s1600/relaxase-antibiotic-resistance_thumb.jpg)

Usando una prueba de comparación de proporciones (denominada binomial en general) de dos colas y asumiendo un nivel de significancia de 0.01 y un tamaño de muestra común de 30 para cada proporción, ¿qué tamaño de efecto puede detectarse con una potencia de .75?

```
pwr.2p.test(n=30,sig.level=0.01,power=0.75)

##
##      Difference of proportion power calculation for binomial
##      distribution (arcsine transformation)
##
##              h = 0.8392269
##              n = 30
##      sig.level = 0.01
##              power = 0.75
##      alternative = two.sided
##
## NOTE: same sample sizes
```

donde h es el tamaño del efecto y n es el tamaño de muestra en cada grupo.

Cohen sugiere que los valores de h de 0.2, 0.5 y 0.8 representan tamaños de efectos pequeños, medianos y grandes, respectivamente.

Usando test de proporciones de prueba de dos colas (bilaterales), y asumiendo un nivel de significancia de 0.05 y una potencia de 0.8, ¿cuál es la n con un tamaño de efecto de 0.35?

```
pwr.2p.test(h=0.35,sig.level=0.05,power=0.8)

##
##      Difference of proportion power calculation for binomial
##      distribution (arcsine transformation)
##
##              h = 0.35
##              n = 128.1447
##      sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
##
## NOTE: same sample sizes
```

Podemos ver que es de n=129 por grupo, en total 129*2

2.7.6.Cálculo del tamaño muestral en diseños ANOVA (“One way”) mediante el uso de la función `pwr.anova.test()`

Una buena introducción teórica a este tema puede encontrarse en:
<https://stat.ethz.ch/~meier/teaching/anova/power.html#introduction-3>

Para un ANOVA de un factor (one way ANOVA) donde se comparan 5 grupos (niveles del factor), calcular el tamaño de muestra necesario en cada grupo para obtener una potencia de 0,80, cuando el tamaño del efecto es moderado ($\delta = 0,25$) y se emplea un nivel de significación de 0,05. Se supondrá que se trata de un diseño balanceado, donde el tamaño muestral por grupo de tratamiento es igual.

Vemos como se realiza utilizando la función `pwr.anova.test()`:

```
pwr.anova.test(k=5, f=.25, sig.level=.05, power=.8)

##
##      Balanced one-way analysis of variance power calculation
##
##           k = 5
##           n = 39.1534
##           f = 0.25
##      sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

Puede observarse que son necesarios 40 réplicas por cada condición experimental (nivel del factor) para ese delta, alfa y potencia estadística.

5.6. Cálculo del tamaño muestral ANOVA de un factor mediante simulación

¿Qué sucede si nos enfrentamos a una situación en la que no hay una función dedicada como `power.anova.test`? Podemos calcular fácilmente la potencia utilizando simulaciones. Basándonos en nuestra configuración específica bajo la alternativa, estamos simulando muchas veces un nuevo conjunto de datos, ajustamos el modelo ANOVA de una vía y verificamos si la prueba F correspondiente es significativa o no.

Veamos un ejemplo de Meier, 2018):
<https://stat.ethz.ch/~meier/teaching/anova/power.html#calculating-power-for-a-certain-design>

En este ejemplo se va a simular el p-valor de un test ANOVA mediante 1000 simulaciones de un diseño de 1 factor, con 5 niveles, con los siguientes parámetros:

- $\mu(57, 63, 60, 60, 60)$

- $\text{Sigma}^2=7.5$
- N=4 réplicas (por grupo)

```

n.sim <- 1000          ## number of simulations
mu    <- c(57, 63, rep(60, 3)) ## group means
sigma2 <- 7.5         ## error variance
n     <- 4            ## number of observations per group

g     <- length(mu)
group <- factor(rep(LETTERS[1:g], each = n))

results <- numeric(n.sim) ## vector to store results in

for(i in 1:n.sim){
  ## Simulate new response, build data set
  y <- rnorm(n * g, mean = rep(mu, each = n), sd = sqrt(sigma2))
  data <- data.frame(y = y, group = group)

  ## Fit one-way ANOVA model
  fit <- aov(y ~ group, data = data)

  ## Extract result of global F-test
  results[i] <- summary(fit)[[1]][1, "Pr(>F)"] < 0.05 ## 1 = reject
}

mean(results) ## proportion of simulation runs that rejected H_0

```

Se obtiene un valor de 0.558, que correspondería a la potencia estadística (probabilidad de no cometer el error de tipo II)

```
## [1] 0.558
```

Obtenemos el resultado de que en el 56% (Algo escasa, teniendo en cuenta que trabajamos normalmente con un 80%) de los casos obtenemos un resultado significativo. Esto está muy cerca del resultado de arriba. Si quisiéramos aumentar la precisión de nuestros cálculos, podríamos hacerlo fácilmente utilizando un valor mayor para el número de ejecuciones de simulación n.sim.

Si ahora cambiamos `n <- 6`, obtendremos:

```
> mean(results) #
[1] 0.812 (Potencia estadística)
```

Un efecto colateral agradable de hacer un análisis de potencia es que realmente hace el análisis completo de los datos en datos simulados y ve inmediatamente si esto funciona según lo previsto.

Desde un punto de vista conceptual, podemos utilizar dicho procedimiento basado en simulación para cualquier diseño. Sin embargo, la cantidad de parámetros aumenta bastante rápidamente a medida que aumenta la complejidad del modelo y algunos de ellos son difíciles de especificar en la práctica (por ejemplo, la varianza del error o las varianzas de diferentes efectos aleatorios). Si tenemos suerte tenemos datos de un estudio previo. Desafortunadamente, las estimaciones de varianza son bastante imprecisas si solo tenemos una cantidad muy limitada de datos.

En ese sentido, los resultados de un análisis de poder típicamente no son muy precisos. Sin embargo, aún deben darnos una idea aproximada del tamaño de muestra requerido en el sentido de que necesitamos 6 o 60 observaciones por grupo.

2.7.7. Cálculo del tamaño muestral en un diseño experimental ANOVA de varios factores mediante simulación

(De <http://www.evolutionarystatistics.org/document.pdf>)

Análisis de potencia por simulación de frecuencias: la complejidad de nuestros diseños experimentales significa que debemos ir mucho más allá de lo que se puede lograr con el software estándar, como las funciones de potencia incorporadas y el paquete `pwr`, que hemos visto anteriormente. Afortunadamente, R se puede programar fácilmente para producir análisis de potencia para cualquier diseño experimental. El enfoque general es:

1. Simular datos bajo la hipótesis nula (para verificar las probabilidades de error de Tipo I) y también para diferentes tamaños de efectos, para estimar la potencia.
2. Ajustar el modelo a los datos simulados.
3. Registre si el análisis del conjunto de datos simulados fue significativo, utilizando las pruebas habituales.
4. Almacena el nivel de significación en un vector.
5. Repita desde el paso 1. un gran número de veces.
6. Tabule cuántas simulaciones produjeron un resultado significativo y, por lo tanto, calcule la potencia.

Vamos a ver un ejemplo sencillo: supongamos que deseamos realizar un estudio con dos factores fijos con interacción, lo que nos lleva a un análisis de varianza de dos vías (ANOVA), con dos niveles para cada factor. Podríamos simular datos bajo la hipótesis nula (suponiendo que no existen diferencias entre medias) usando el siguiente código:


```

## First set up the design
reps <- 1000
design <- expand.grid(A=c("a", "b"), B=c("c", "d"), reps=1:10)
pvals <- vector("numeric", length=reps)
## simulate data under the null hypothesis.
for (i in 1:reps) {
design$response <- rnorm(40) # random data with zero mean
## fit the model
fit <- aov(response~A*B, data=design)
## record the p-value for the interaction term.
## You could record other values.
## Save the p value
pvals[i] <- summary(fit)[[1]][[5]][3]
}
Type1 <- length(which(pvals < 0.05))/reps
print(Type1)

## [1] 0.056

```

Parece que la tasa de error Tipo I ($\alpha = 0.056$) es aceptable para el término de interacción ANOVA de 2 factores. Ahora para calcular algunas estadísticas de potencia.

Calcularemos la potencia para una diferencia entre las medias de 2 unidades, para los tamaños de muestra 5, 10, 20.

```

ssize <- c(5, 10, 20)
pvals <- matrix(NA, nrow=reps, ncol=3)
## First set up the design
for (j in 1:3) {
reps <- 1000
design <- expand.grid(reps=1:ssize[j], A=c("a", "b"), B=c("c", "d"))
## simulate data under the null hypothesis.
for (i in 1:reps) {
design$response <- c(rnorm(3*ssize[j]), rnorm(ssize[j], mean=2))
## fit the model
fit <- aov(response~A*B, data=design)
## record the p-value for the interaction term.
## You could record other values.
## Save the p value
pvals[i,j] <- summary(fit)[[1]][[5]][3]
}
}
Power <- apply(pvals, 2, function (x) length(which(x < 0.05))/reps)
names(Power) <- as.character(ssize)
print(Power)

```

| | | | |
|----|-------|-------|-------|
| ## | 5 | 10 | 20 |
| ## | 0.553 | 0.888 | 0.994 |

Vemos que la potencia estadística es demasiado baja para un tamaño de muestra de 5 (potencia = 0.553), pero aumenta a un nivel aceptable para 10 (potencia = 0.888), repeticiones por tratamiento.



ANOVA ONE WAY

2.8 Introducción al análisis de la varianza (ANOVA) unifactorial “One-way”

Una buena introducción a este tema y al ANOVA puede encontrarse en “ANOVA: A Short Intro Using R” de Lukas Meier de 2018 (Ver en: <https://stat.ethz.ch/~meier/teaching/anova/>)

Otra recomendable es: Oehlert, Gary. 2010. *A First Course in Design and Analysis of Experiments*. <http://users.stat.umn.edu/~gary/book/fcdae.pdf>.

Se utilizarán algunos de los ejemplos de estas obras con propósito docente.

2.8.1 Concepto y suposiciones

El análisis de varianza de una vía (unifactorial) o “One way” (ANOVA) se utiliza para comparar varios grupos de una variable cuantitativa (factor). Específicamente, es una forma de calcular si hay diferencias estadísticamente significativas entre las medias de tres o más grupos independientes. Prueba la siguiente hipótesis nula:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_a$$

Donde μ = media de grupo y a = número de grupos (niveles del factor).

En el caso de que el ANOVA unifactorial devuelva un resultado significativo, se acepta la hipótesis alternativa (H_A), lo que significa que al menos dos medias de grupo son estadísticamente diferentes. Sin embargo, el estadístico obtenido en el ANOVA no dirá cuáles son los grupos significativamente diferentes, para eso necesitaremos una prueba post hoc.

Para utilizar el estadístico de prueba ANOVA, se deben cumplir tres suposiciones principales:

- La variable dependiente se distribuye normalmente en cada grupo.
- Hay homogeneidad de varianzas, las varianzas de población en cada grupo son iguales.
- Independencia de observaciones.

Si estos supuestos no se cumplen puede haber problemas y errores en las conclusiones obtenidas. Se dedicarán ejemplos y análisis prácticos en estos casos.

Este apartado simplemente intenta introducir la modelización de experimentos de factores y su solución mediante ANOVA, se encuentra ampliada con más detalle en la parte teórica.

En un experimento donde quiere estudiarse el efecto de un factor experimental categórico A (ej: tratamiento, medicamento, etc) sobre una variable cuantitativa continua Y, las observaciones vienen descritas por el modelo lineal:

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

donde:

$$i = 1, 2, 3, \dots, a,$$

$$j = 1, 2, 3, \dots, r,$$

$e_{ij} \sim N(0, \sigma)$, son los errores de cada observación (diferencia entre los valores observados y los predichos por el modelo $\mu + \alpha_i + e_{ij}$) y poseen distribución normal con media 0 y varianza σ^2

Siendo a = número de niveles del factor A, r = número de réplicas de cada condición experimental, y se impone la restricción de los efectos:

$$\sum_1^a \alpha_i = 0$$

Los errores e_{ij} son variables aleatorias independientes y normalmente distribuidas con media cero y varianza constante σ^2 .

Si el modelo es apropiado para los datos, los residuos observados e_{ij} reflejarán las propiedades exigidas los errores aleatorios que se cumplen en los modelos lineales. Esta es la idea básica en el análisis de los residuos. Así, si las hipótesis relativas al modelo son ciertas, los residuos variarán aleatoriamente. Si, por el contrario, descubrimos que los residuos presentan tendencias sistemáticas inexplicadas, tendremos que sospechar de la validez del modelo y plantearnos una solución o una alternativa.

Diagnóstico y robustez en el ANOVA, alternativas y soluciones a las violaciones de los supuestos del ANOVA. Métodos de comparación múltiple (normales, no paramétricas y alternativas)

Cuando se selecciona un modelo lineal para un conjunto de datos, frecuentemente no se puede estar seguro de que ese modelo sea adecuado. En el caso del ANOVA, puede suceder que alguna o varias características del modelo tales como la normalidad de

los términos de error o la independencia en los datos investigados no se verifiquen. ¿Es válido el modelo? ¿Son válidos los resultados? ¿Qué hacer?

Por lo tanto, es importante examinar la adecuación del modelo a los datos antes de realizar un análisis de los mismos basado en dicho modelo.

En los siguientes apartados se va a tratar de los métodos de diagnóstico y la validación del modelo, que consiste en estudiar si las hipótesis básicas del modelo están o no en contradicción con los datos observados. Para ello, en este apartado, se discuten algunos métodos gráficos sencillos y procedimientos estadísticos. También se estudian los efectos que produce el incumplimiento de las suposiciones fundamentales.

Finalmente, una de las pretensiones de los siguientes apartados es la de presentar las últimas novedades en el análisis estadístico ANOVA en diferentes situaciones, muy presentes en los datos experimentales como: heterogeneidad de los datos en los diferentes niveles del factor, no normalidad de residuos, test de comparación múltiple a posteriori en estas situaciones, alternativas no paramétricas al análisis ANOVA, etc. Se presentarán también las diferentes librerías de R utilizadas.

Los métodos gráficos y los contrastes estadísticos que se describen en este capítulo para diagnosticar la adecuación del modelo de análisis de la varianza son, respectivamente, las gráficas de residuos y los contrastes de igualdad de varianzas. También se aborda el uso de transformaciones como método para mejorar la validez del modelo de análisis de la varianza y se considera la influencia que tienen las desviaciones de las hipótesis del modelo ANOVA en la inferencia. Podemos, por tanto, considerar que en el estudio de un experimento se debe seguir un proceso secuencial formado por los pasos siguientes:

El análisis ANOVA es razonablemente robustos frente a ciertos tipos de desviaciones del modelo, tales como que los términos de error no estén exactamente normalmente distribuidos. Por lo tanto, podemos decir que el principal objetivo del examen de la adecuación del modelo es detectar graves desviaciones de las condiciones supuestas en el modelo.

2.8.2 Ejemplo: Ensayo clínico de un fármaco

Veamos un ejemplo con un análisis ANOVA de 1 factor fijo, con un sencillo ejemplo.

Con la intención de evaluar la eficacia de un medicamento en el nivel de alerta de unos pacientes, tres dosis (a,b,c) de un determinado fármaco A se administraron a 18 sujetos. La eficacia del medicamento se realizó mediante un análisis de varianza ANOVA. Este ejemplo ha sido obtenido en:

<http://personality-project.org/r/datasets/R.appendix1.data>

```
#####  
##### Lectura del fichero de datos  
#####  
  
#tell where the data come from  
datafilename="http://personality-  
project.org/r/datasets/R.appendix1.data"  
data.ex1=read.table(datafilename,header=T) #read the data into a table
```

Los datos obtenidos en el experimento son:

```
data.ex1  
  
##      Dosage Alertness  
## 1      a      30  
## 2      a      38  
## 3      a      35  
## 4      a      41  
## 5      a      27  
## 6      a      24  
## 7      b      32  
## 8      b      26  
## 9      b      31  
## 10     b      29  
## 11     b      27  
## 12     b      35  
## 13     b      21  
## 14     b      25  
## 15     c      17  
## 16     c      21  
## 17     c      20  
## 18     c      19  
  
summary(data.ex1)  
  
##      Dosage      Alertness  
## a:6      Min.      :17.00  
## b:8      1st Qu.:21.75  
## c:4      Median   :27.00  
##          Mean     :27.67  
##          3rd Qu.:31.75  
##          Max.     :41.00
```

Análisis ANOVA

Para analizar si existen diferencias entre las medias de las tres dosis de medicamento, realizamos una análisis ANOVA. El comando anova utilizado en R fue aov:

```
options(show.signif.stars=T, contrasts=c("contr.sum", "contr.poly"))  
aov.ex1 = aov(Alertness~Dosage,data=data.ex1) #do the analysis of  
variance
```

Es importante tener en cuenta el orden de los argumentos. El primer argumento es siempre la variable dependiente (Alerta). Es seguido por el símbolo de tilde (~) y la variable independiente (s). El argumento final para aov es el nombre de la estructura de datos que se esta analizando. aov.ex1 es el nombre de la estructura que desea que el análisis para almacenar. Este formato general será verdad para todos los ANOVAs que llevara a cabo.



Imagen de: <https://www.clarityfoundation.org/wp-content/uploads/2017/06/dispensing-medicines-for-a-clinical-trial-jan-chlebik-for-the-icr-2014-embed-825x492.jpg>

Los resultados del ANOVA se pueden ver con el comando “summary”:

```
summary(aov.ex1)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## Dosage      2  426.3   213.13    8.789 0.00298 **
## Residuals  15   363.8    24.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como puede verse en la tabla anterior existen diferencias estadísticamente significativas entre los niveles del fármaco ($p=0.00298$). ¿Podemos estar seguros de que estas conclusiones son válidas? Para ello decidimos realizar algunas pruebas diagnósticas del ANOVA.

Las diferentes medias por nivel de los 3 medicamentos pueden verse en la siguiente tabla:

```
print(model.tables(aov.ex1,"means"),digits=3)           #report the means and
the number of #subjects/cell

## Tables of means
## Grand mean
##
## 27.66667
##
## Dosage
##      a      b      c
##  32.5  28.2  19.2
## rep  6.0  8.0  4.0
```

Las estimaciones de los parámetros del modelo lineal pueden verse en:

```
tapply(data.ex1$Alertness, data.ex1$Dosage, mean, na.rm=TRUE) # means for
Levels of A

##      a      b      c
## 32.50 28.25 19.25

tapply(data.ex1$Alertness, data.ex1$Dosage, sd, na.rm=TRUE) # std.
deviations of Levels of A

##      a      b      c
## 6.595453 4.432026 1.707825

tapply(data.ex1$Alertness, data.ex1$Dosage, function(x) sum(!is.na(x))) #
counts of Levels of A

## a b c
## 6 8 4

tapply(data.ex1$Alertness, data.ex1$Dosage, median, na.rm=TRUE)

##      a      b      c
## 32.5 28.0 19.5
```



```
dummy.coef(aov.ex1)
```

```
## Full coefficients are
```

```
##
```

```
## (Intercept):      26.66667
```

```
## Dosage:           a          b          c
```

```
##                  5.833333  1.583333 -7.416667
```

Efectos de las diferentes violaciones

Pero ¿Cuál es el efecto de las diferentes violaciones?

- No normalidad:

No es un gran problema a menos que la desviación de la normalidad sea extrema.

F-test y otros procedimientos de prueba y relacionados son bastante robustos al supuesto de normalidad, tanto en términos de nivel de significación como de poder estadístico.

- Varianza de error desigual (heterocedasticidad):

F-test y otras pruebas estadísticas relacionadas son bastante robustos contra la heterocedasticidad bajo un diseño aproximadamente equilibrado.

- Una inferencia de tipo paramétrica, como las comparaciones por pares de las medias a posteriori del ANOVA (pairwise test), podría verse sustancialmente afectada.

- No independencia:

- Puede tener efectos secundarios graves (pérdida efectiva de grados de libertad).

- También es difícil de corregir.

- Por lo tanto, es muy importante utilizar la aleatorización cuando sea necesario para asegurar que existe independencia entre las observaciones.

Diagnóstico práctico del modelo ANOVA

Una buena opción para estudiar el diagnóstico del ANOVA es el estudio de los residuos del modelo.

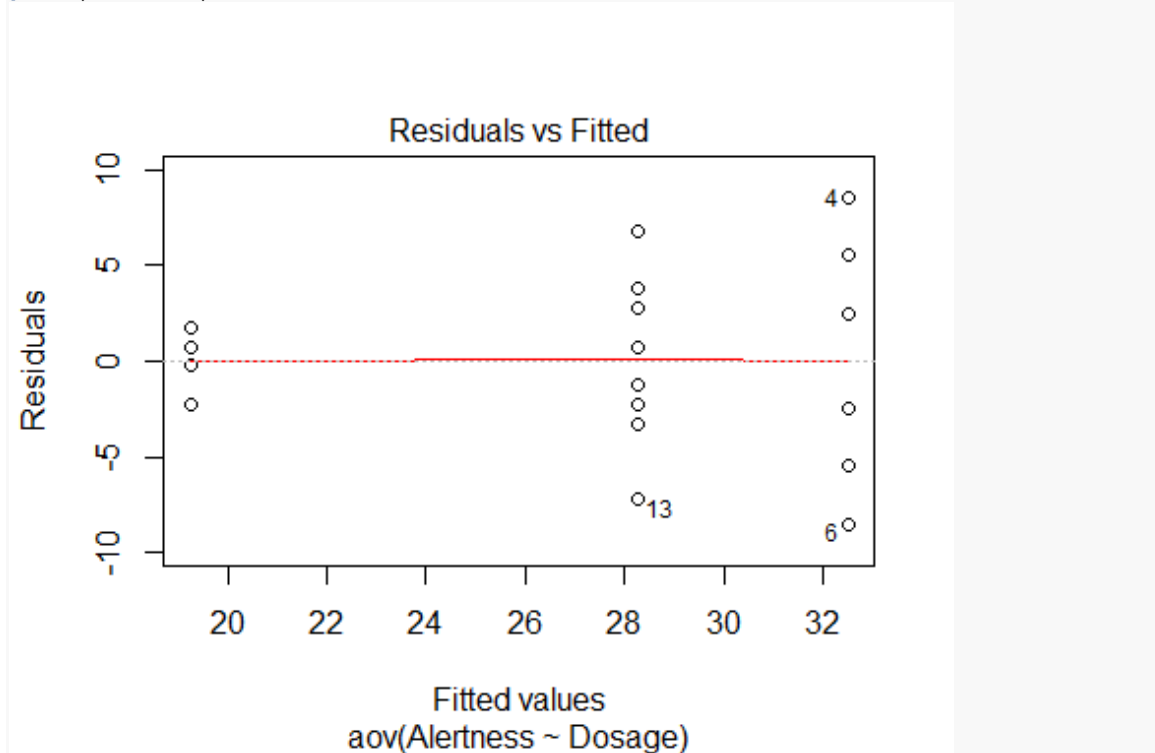
También pueden verse los residuos del modelo utilizando el comando `residuals()`

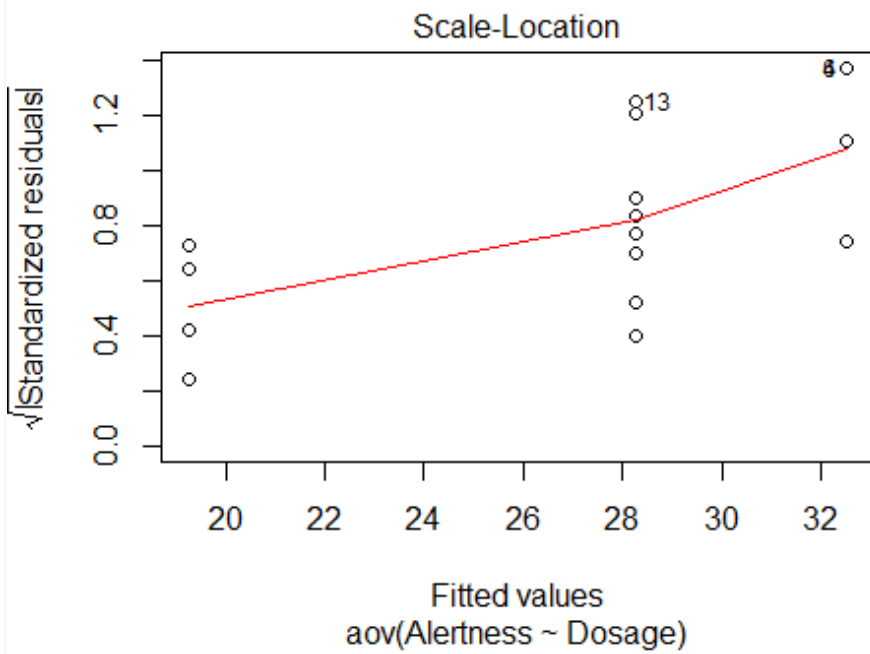
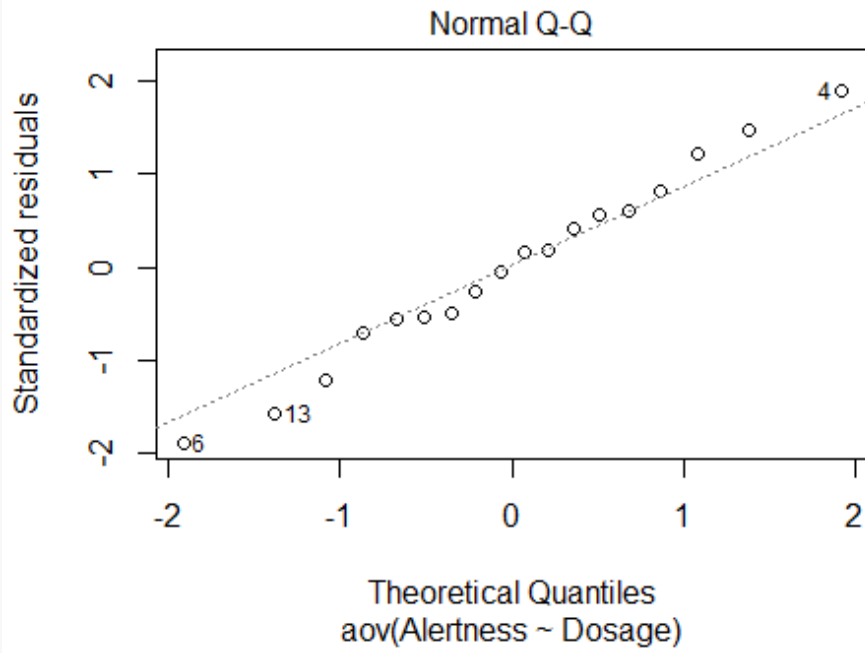
```
<>= residuals(aov.ex1) ``
```

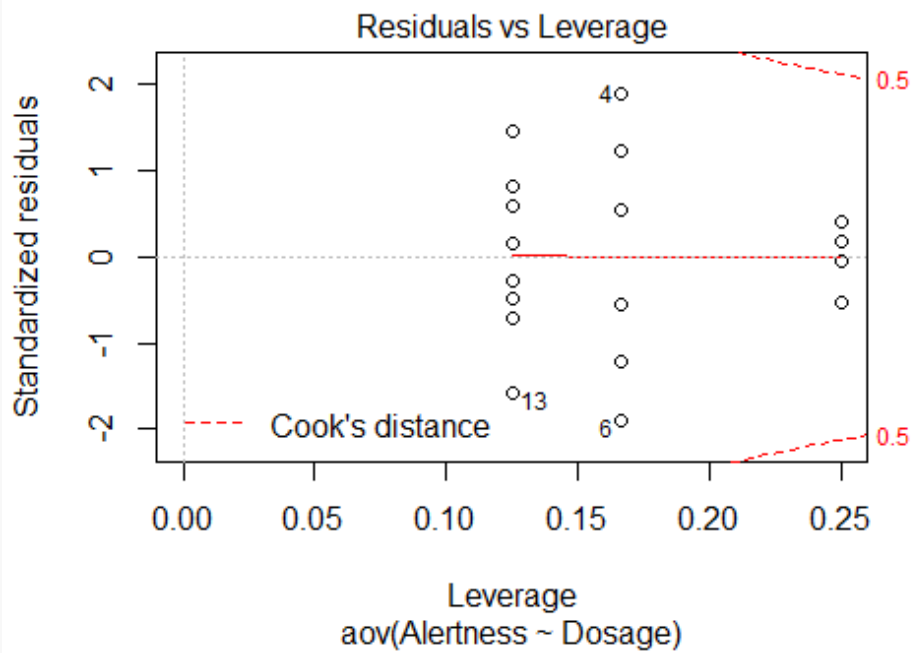
Una manera rápida de realizar un diagnóstico a posteriori y observar los posibles problemas en el cumplimiento de los supuestos necesarios para realizar un ANOVA, es utilizar el diagnóstico de residuos que ofrece `plot(aov.ex1)`:

En el gráfico de probabilidad (qqplot) los residuos aparecen razonablemente alineados, lo que sugiere normalidad

```
plot(aov.ex1)
```

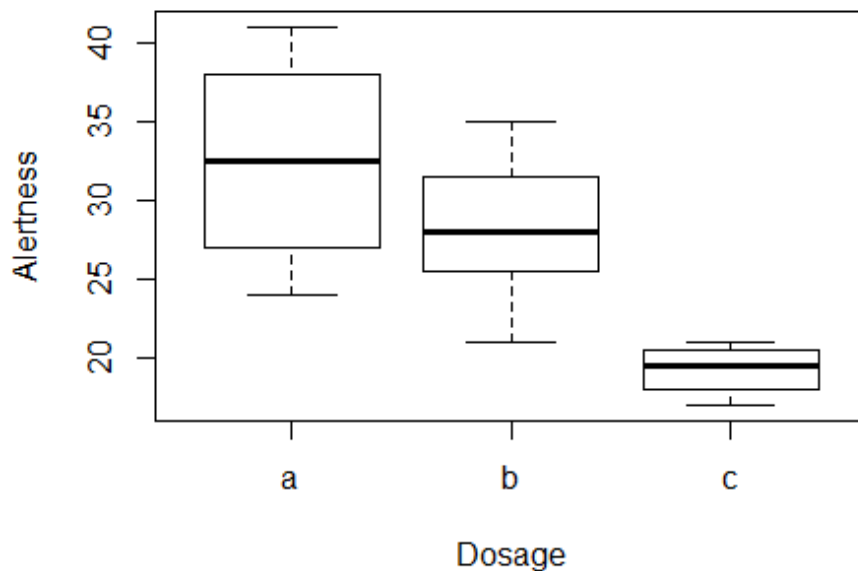






Y complementarlo con un Box-Plot para detectar observaciones extremas y heterogeneidad por niveles del factor:

```
boxplot(split(data.ex1$Alertness,data.ex1$Dosage), ylab="Alertness",
xlab="Dosage")
```



Una buena revisión y un ejemplo para el tratamiento de desviaciones en el ANOVA puede verse en:

https://stats.libretexts.org/Bookshelves/Advanced_Statistics_Computing/Regression_Analysis/Analysis_of_Variance/Effects_of_violations_of_model_assumptions

<https://www.statmethods.net/stats/anovaAssumptions.html>

<https://arc.lib.montana.edu/book/statistics-with-r-textbook/item/57>

Una posibilidad es realizar un test de Bartlett para comprobar la homocedasticidad, esto es la igualdad de varianzas entre niveles del factor analizado.

Las varianzas estimadas para los niveles del factor fármaco son:

```
tapply(data.ex1$Alertness, data.ex1$Dosage, var, na.rm=TRUE)
```

```
##          a          b          c
## 43.500000 19.642857  2.916667
```

No se detecta heterocedasticidad entre los niveles del factor estudiado. Pero quiere decir esto que esta no exista o bien debido a la poca cantidad de muestra no puede detectarse la heterocedasticidad.

La prueba de Bartlett es sensible a las desviaciones de la normalidad. Es decir, si las muestras provienen de distribuciones no normales. La Prueba de Levene, la prueba de Brown-Forsythe y especialmente la de Figner-Killeen son alternativas a la prueba de Bartlett que son menos sensibles a las desviaciones de la normalidad.

La prueba de Levene es una alternativa al test de Bartlett y se realiza mediante el siguiente comando:

```
library(car)

## Loading required package: carData

leveneTest(Alertness~Dosage,data=data.ex1)

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2  4.1667 0.03638 *
##      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O bien un test de Figner-Killeen Test of Homogeneity of Variances:

```
fligner.test(Alertness~Dosage,data=data.ex1)

##
## Fligner-Killeen test of homogeneity of variances
##
## data:  Alertness by Dosage
## Fligner-Killeen:med chi-squared = 6.3308, df = 2, p-value = 0.0422
```

La prueba de Fligner-Killeen (mediana) se ha comprobado en estudios de simulación como una de las muchas pruebas de homogeneidad de las varianzas más robusta frente a desviaciones de la normalidad (ver Conover, Johnson & Johnson, 1981).

En 2 de los test anteriores (Figner-Killeen y Levene) se detecta heterocedasticidad, lo que indica que no se cumple uno de los supuestos del ANOVA.

Se utilizan las pruebas de Shapiro, Anderson Darling, y otras son pruebas de hipótesis nula contra el supuesto de normalidad. Probablemente la prueba más utilizada para comprobar la normalidad de los residuos del modelo lineal es la prueba de Shapiro-Wilks. La función para realizar esta prueba, convenientemente llamado `shapiro.test()`, es fácil de usar. Hay que tener en cuenta el tamaño muestral: Cuando el tamaño de la muestra es pequeño, incluso grandes desviaciones de la normalidad no se detectan, y cuando el tamaño de la muestra es grande, incluso la más mínima desviación de la normalidad consiguen rechazar la hipótesis nula.

```
shapiro.test(residuals(aov.ex1))

##
## Shapiro-Wilk normality test
```

```
##  
## data: residuals(aov.ex1)  
## W = 0.98604, p-value = 0.991
```

Se recomienda utilizar el test de Shapiro-Wilks en el caso de muestras pequeñas, pero si la muestra es grande se recomienda el test de Anderson-Darling que puede encontrarse en el paquete nortest.

```
library(nortest)  
ad.test(residuals(aov.ex1))  
  
##  
## Anderson-Darling normality test  
##  
## data: residuals(aov.ex1)  
## A = 0.11805, p-value = 0.987
```

No puede rechazarse la hipótesis de que los residuos tengan distribución normal, ya que $p > 0.05$ en ambas pruebas.

Para comprobar la Independencia de los residuos puede recurrirse al test de Durbin-Watson para autocorrelación de perturbaciones. La prueba de Durbin-Watson tiene la hipótesis nula de que la autocorrelación de las perturbaciones es 0. Es posible poner a prueba frente a la alternativa que es mayor que, no es igual a, o menor que 0, respectivamente. Esto puede ser especificado por el argumento alternativo.

Bajo el supuesto de perturbaciones distribuidas normalmente, la distribución del estadístico de Durbin-Watson bajo la hipótesis nula es la distribución de una combinación lineal de variables chi-cuadrado. El p-valor se calcula utilizando la versión Fortran del algoritmo Pan o Gradsol. Para tamaños de muestra grandes el algoritmo puede fallar para calcular el valor de p; en ese caso, se imprime una advertencia y se dará un valor p aproximado; este valor p se calcula utilizando una aproximación normal con media y varianza del estadístico de Durbin-Watson.

Veamos el cálculo del test de Durbin-Watson para verificar la independencia de observaciones:

```
library(lmtest)  
  
## Loading required package: zoo  
  
##  
## Attaching package: 'zoo'  
  
## The following objects are masked from 'package:base':  
##  
## as.Date, as.Date.numeric  
  
dwtest(Alertness~Dosage, data=data.ex1)
```

```
##  
## Durbin-Watson test  
##  
## data: Alertness ~ Dosage  
## DW = 2.2765, p-value = 0.5424  
## alternative hypothesis: true autocorrelation is greater than 0
```

No se detecta autocorrelación de los residuos ya que $p > 0.05$, por lo que no puede señalarse que exista dependencia de residuos.

Alternativas y soluciones a las violaciones de los supuestos del ANOVA

La falta de normalidad no es un problema severo, pues el ANOVA es robusta a la falta de normalidad. Solo cuando se encuentran valores extremadamente alejados puede haber problemas en la significación de las pruebas. El problema fundamental que ocasiona el incumplimiento de este supuesto es que las inferencias que se hacen no son válidas.

La corrección de este problema se hace mediante el uso de: Transformaciones, pruebas no paramétricas, modelos lineales generalizados o modelos generalizados en métodos de cuasi-verosimilitud.

Hay tres maneras (al menos) de manejar esta situación no normalidad que se verán más adelante.

Consecuencia de la heterodasticidad: Puede ser un grave problema en el ANOVA. El estimador de la varianza de los errores es sesgado y por ende, las varianzas de combinaciones de los estimadores de parámetros son erradas, conllevando esto a que las pruebas de significación carezcan de validez. Especialmente en los casos más extremos. Este problema es más delicado pero también pueden encontrarse diferentes soluciones parecidas al problema de la no normalidad de los residuos.

¿Qué hacer si se detectan problemas durante el diagnóstico del modelo? Veamos diferentes alternativas.

Prueba de Kruskal-Wallis, una alternativa no paramétrica al ANOVA

Una primera solución que podemos hacer es un test no paramétrico utilizando la prueba de Kruskal-Wallis.

La prueba de Kruskal-Wallis es un método no paramétrico para probar si un grupo de datos proviene de la misma población. Intuitivamente, es idéntico al ANOVA con los datos reemplazados por categorías. Es una extensión de la prueba de la U de Mann-Whitney para 3 o más grupos.

Ya que es una prueba no paramétrica, la prueba de Kruskal-Wallis no asume normalidad en los datos, en oposición al tradicional ANOVA. Si asume, bajo la hipótesis nula, que los datos vienen de la misma distribución. Una forma común en que se viola este supuesto es con datos heterocedásticos. En el ejemplo planteado que presenta heterocedasticidad no serviría realizar este test, aunque veamos cómo se aplica y su resultado:

```
kruskal.test(Alertness~Dosage,data=data.ex1)

##
##  Kruskal-Wallis rank sum test
##
## data:  Alertness by Dosage
## Kruskal-Wallis chi-squared = 9.4272, df = 2, p-value = 0.008972
```

Al igual que en la prueba ANOVA realizada anteriormente existen efecto del medicamento sobre la variable de respuesta.

Uno de los problemas de utilizar esta prueba no paramétrica es que es menos potente que el enfoque ANOVA.

Transformaciones de la variable principal

En el caso de violaciones de los supuestos del análisis ANOVA, una solución es las transformaciones: Log, raíz, o transformaciones Box-Cox.

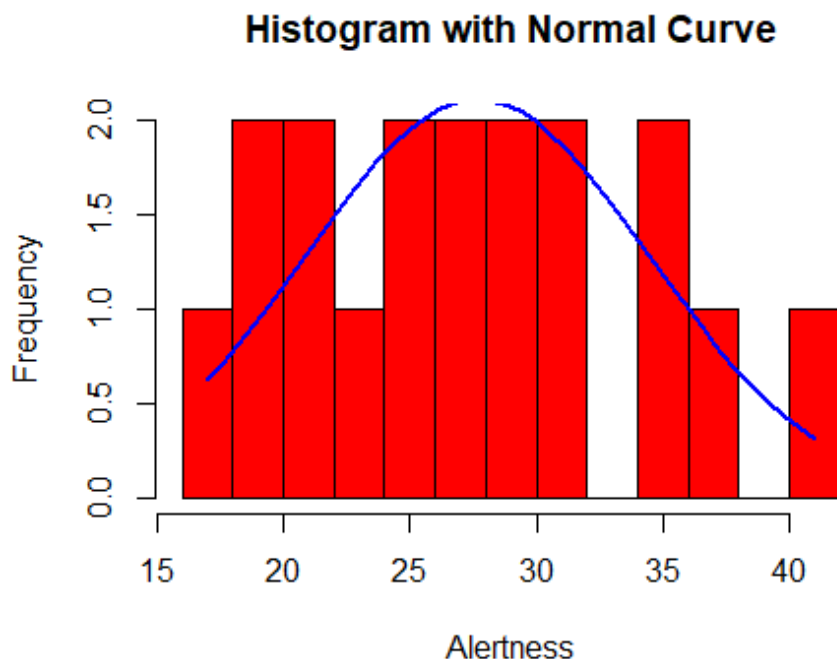
Un buen ejemplo puede encontrarse en: [link](#):

Debe cargarse la librería 'car' a la biblioteca. Algunas de las transformaciones mas comunes para la variable principal son:

Aplicar $Y_{\{0\}} = Y^{\{2\}}$ para la transformación cuadrado. Aplicar $Y_{\{0\}} = e^{\{Y\}}$ para la transformación exponencial

Es recomendable siempre hacer un histograma de lo que se quiera transformar para opinar y entender el tipo de transformación que puede hacerse. Por ejemplo un histograma de la variable original Y:

```
x <- data.ex1$Alertness
h<-hist(x, breaks=10, col="red", xlab="Alertness",
        main="Histogram with Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Transformaciones normalizadoras de Box-Cox

Las transformaciones de Box-Cox son una familia de transformaciones potenciales usadas en estadística para corregir sesgos en la distribución de errores, para corregir varianzas desiguales (para diferentes valores de la variable predictora) y principalmente para corregir la no linealidad en la relación (mejorar correlación entre las variables). Lo más usual es realizar una transformación potencial, que este definida como una función continua que varía con respecto a la potencia lambda (\$ \$). Dado el vector de observaciones $\{y_1, y_2, \dots, y_n\}$, se realiza la transformación $y_i = y_i^\lambda$, donde λ es un valor para estimar.

En R el valor de λ se puede estimar, por ejemplo utilizando el comando `powerTransform(Yi)` en `library(car)`. Veamos un ejemplo de transformación normalizante Box-Cox:

```
library(car)
summary(powerTransform(data.ex1$Alertness))

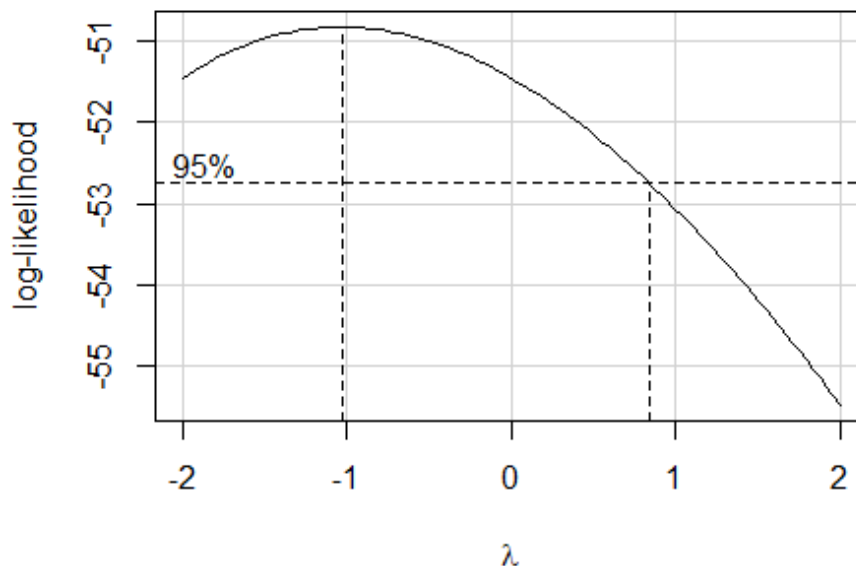
##          bcPower      Transformation      to      Normality
##          Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## data.ex1$Alertness    0.2591              1    -1.6832      2.2014
##
## Likelihood ratio test that transformation parameter is equal to 0
##                               (log transformation)
##                               LRT df      pval
## LR test, lambda = (0) 0.06839912 1 0.79368
```

```
##
## Likelihood ratio test that no transformation is needed
##                                     LRT df      pval
## LR test, lambda = (1) 0.5557829  1 0.45596
```

En este caso `powerTransform()` nos recomendaria cambiar la variable de respuesta original y_i por y_i^{λ} ($\lambda = 0.55578285$), pero en este caso $p > 0.05$ y no es necesario.

Otra libreria que puede utilizarse para hacer transformaciones normalizadoras de Box-Cox es `lmSupport` y el comando: `modelBoxCox(reaction.aov)`

```
library(lmSupport)
modelBoxCox(aov.ex1) #Lambda = -0.38 para corregir la normalidad
```



```
## [1] "Best Lambda= -1.03"
## [1] "Chi-square (df=1)= 4.51"
## [1] "p-value= 0.03362"
```

En este caso `powerTransform(Y_i)` nos recomendaria cambiar la variable de respuesta original y_i por y_i^{λ} ($\lambda = -1.03$). En este caso si podriamos hacerlo ya que $p = 0.03$. Asi podria transformar la variable principal con la transformación de Box-Cox:

```
library(lmSupport)
data.ex1$Alertness.b <- data.ex1$Alertness-1.03
```

Se repite nuevamente el análisis ANOVA con la variable transformada y se puede hacer otro diagnostico a posteriori para ver si se ha podido corregir el problema de la

heterocedasticidad. Los nuevos valores de la variable de respuesta y_i^{λ} ($\lambda = -1.03$) son ahora:

```
data.ex1$Alertness.b
## [1] 0.03009990 0.02359516 0.02568088 0.02181889 0.03355022 0.03787760
## [7] 0.02816408 0.03488008 0.02910030 0.03116952 0.03355022 0.02568088
## [13] 0.04346244 0.03631799 0.05403033 0.04346244 0.04570241 0.04818189
```

Si ahora analizamos la homogeneidad de varianzas con el test de Fligner-Killeen para la variable transformada:

```
fligner.test(data.ex1$Alertness.b ~ Dosage, data=data.ex1)
##
## Fligner-Killeen test of homogeneity of variances
##
## data: data.ex1$Alertness.b by Dosage
## Fligner-Killeen:med chi-squared = 0.64095, df = 2, p-value =
## 0.7258
```

Obtenemos un p-valor = 0.7258, que nos indicaría que ya no existe heterogenicidad. Repitiendo AONOVA para la variable transformada:

```
res.2 <- aov(data.ex1$Alertness.b ~ Dosage, data = data.ex1)
summary(res.2)
##           Df    Sum Sq  Mean Sq F value  Pr(>F)
## Dosage      2 0.0009310 0.0004655  14.74 0.000288 ***
## Residuals  15 0.0004738 0.0000316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenemos, como con la variable Y original, diferencias estadísticamente significativas entre los niveles del factor estudiado.

Otra posibilidad es utilizar transformaciones tales como el logaritmo o la raíz cuadrada para estabilizar la varianza. Eso funcionara en algunos casos, pero las transformaciones de tipo Box-Cox estabilizan la varianza juntando los datos de forma asimétrica y son eficaces en numerosas ocasiones.

Datos heterocedásticos: transformaciones y alternativas (corrección de Welch, Huber-White “sándwich”, etc)

A parte de las opciones vistas anteriormente (transformaciones, Kruskal-Wallis test), hay una serie de opciones disponibles cuando se trata de datos heterocedásticos. Por desgracia, ninguna de ellos se garantiza que funcione siempre. Estas son algunas de las opciones que se pueden utilizar:

En este tutorial no pueden estudiarse todas y solo se trataran algunas de las posibles.

Dado que este caso es un ANOVA (es decir, no hay variables continuas), una manera de lidiar con la heterogeneidad es el uso de la corrección de Welch, modificando los grados de libertad del denominador en la prueba F:

```
oneway.test(Alertness~Dosage,data=data.ex1, na.action=na.omit,
var.equal=FALSE)

##
## One-way analysis of means (not assuming equal variances)
##
## data: Alertness and Dosage
## F = 19.339, num df = 2.0000, denom df = 9.2651, p-value =
## 0.0004931
```

En econometría el Error estándar de Huber-White (“sandwich”) es muy popular. Al igual que la corrección de Welch, no se requiere saber las varianzas a priori y no requiere que se estimen los pesos de los datos. Una primera forma de aplicar esta corrección es con `library(sandwich)`:

```
library(sandwich)
mod = lm(Alertness~Dosage,data=data.ex1, na.action=na.omit)
sqrt(diag(vcovHC(mod)))

## (Intercept)      Dosage1      Dosage2
## 1.177490      2.070382      1.523764
```

La función `vcovHC()` calcula una matriz de varianza-covarianza consistente para el problema de la heterocetasticidad para los parámetros beta (sus códigos ficticios) del modelo a estimar. Para obtener los errores estándar, se extrae la diagonal principal (varianzas) y se toman las raíces cuadradas. Para obtener pruebas t para las betas, se divide sus estimaciones de los coeficientes por los errores estándar y se comparan los resultados con una distribución t apropiada (es decir, la distribución t con sus grados de libertad residuales)

Para los usuarios de R específicamente en el comando `Anova` en el paquete de `car` (`library(car)`) puede aceptar un argumento `white.adjust=TRUE` para obtener un p-valor que corrija la heterodasticidad. Así para realizar un ANOVA que corrija la heterodasticidad de los niveles del factor, se realiza de la siguiente manera:

```
Anova(mod, white.adjust=TRUE)

## Coefficient covariances computed by hccm()
## Analysis of Deviance Table (Type II tests)
##
## Response: Alertness
##           Df      F    Pr(>F)
## Dosage     2 17.064 0.0001367 ***
```

```
## Residuals 15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Existen otros métodos como el uso de bootstrap. Se puede obtener una estimación empírica de la distribución muestral real del estadístico F por bootstrapping. En primer lugar, se crea una hipótesis nula creando un grupo en que la respuesta para todos los individuos sea igual. Entonces se remuestra con reemplazo y se calcula el estadístico de prueba (F) en cada bootstrap (paso) para obtener una estimación empírica de la distribución muestral de F bajo la hipótesis nula los datos reales sean cual sea su estado con respecto a la normalidad o la homogeneidad. La proporción de que la distribución de muestreo sea extrema o más extrema que el estadístico de prueba observado es el valor de p. Un ejemplo de ANOVA-bootstrap puede encontrarse en:

<https://stats.stackexchange.com/questions/91872/alternatives-to-one-way-anova-for-heteroskedastic-data>

Comparación múltiples de medias a posteriori del ANOVA (MCP)

Si el ANOVA confirma la existencia de diferencias significativas entre los tratamientos (NIVELES DEL FACTOR), es conveniente investigar que medias son distintas. Para ello, emplearemos diversas técnicas cuyo objeto es identificar que niveles son estadísticamente diferentes y en cuanto oscila el valor de esas diferencias. Consideraremos su aplicación únicamente al modelo de efectos fijos. El uso de estas técnicas, en algunos casos, está supeditado al resultado del ANOVA; en otros casos, las técnicas pueden emplearse directamente sin haber realizado previamente dicho análisis. Este conjunto de técnicas, de las que existen numerosas variantes, se engloba bajo la denominación de contrastes para comparaciones múltiples (MCP) ya que su objetivo fundamental es comparar entre si medias de tratamientos o grupos de estos.

Las pruebas estadísticas para comparaciones múltiples más frecuentemente utilizadas se basan en la distribución t de Student. Anteriormente, en los primeros ejemplos se han visto pruebas realizadas con la función `pairwise.t.test()`, es decir un simple `t.test` pero corregido utilizando un ajuste del p-valor.

Se recomienda el anexo teórico para ver la base de este apartado de MCP.

En primer lugar, estudiamos un procedimiento basado en la representación gráfica de los datos del experimento. Una gráfica de medias por niveles con el intervalo de confianza del 95 % puede obtenerse con el comando:

```
#Library(Rcmdr)
#plotMeans(data.ex1$Alertness, data.ex1$Dosage, error.bars="conf.int",
```

```

Level=0.95)
library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

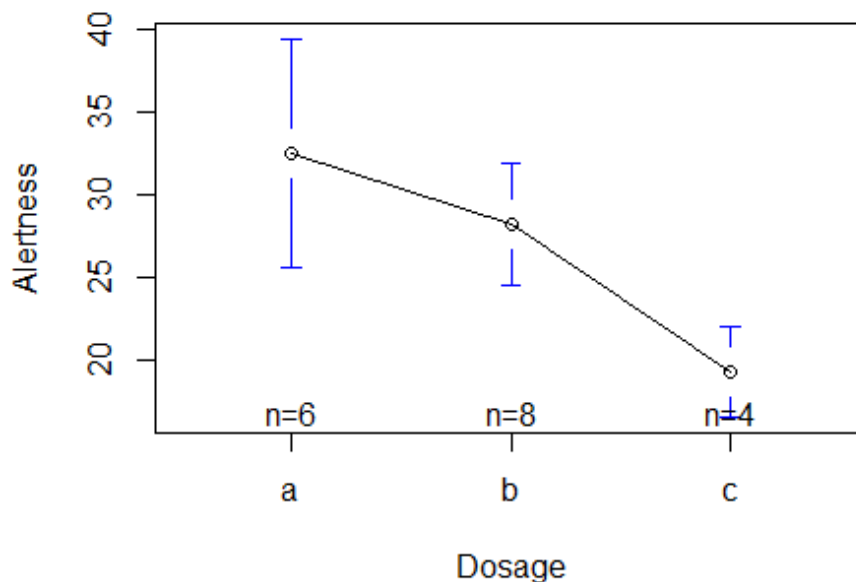
plotmeans(Alertness ~ Dosage, data = data.ex1, frame = T)

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not
a
## graphical parameter

## Warning in axis(1, at = 1:length(means), labels = legends, ...):
"frame" is
## not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not
a
## graphical parameter

```



En la gráfica anterior puede intuirse que existen diferencias entre las medias de los niveles del factor a y a, ya que los intervalos de confianza no se tocan y están perfectamente separados, pero hay que confirmarlo utilizando alguna de las técnicas MCP y ofrecer un p-valor y un intervalo de confianza correcto.

Después del método gráfico consideramos la técnica de comparación por parejas de niveles (pairwise) introducida por Fisher en 1935. Dicha técnica, denominada método de la diferencia mínima significativa o método LSD (Least Significant Difference), se basa en la construcción de tests de hipótesis para la diferencia de cualquier par de medias, que solucionan en parte el problema denominado inflamiento del error de tipo I.

Cuando el número de posibles comparaciones es elevado, la aplicación reiterada de este procedimiento, para un nivel de significación dado (alfa), puede conducir a un número grande de rechazos de la hipótesis nula aunque no existan diferencias reales (inflamiento del error de tipo I). Una solución es aplicar una corrección al p-valor (corrección de holm, "hochberg", "bonferroni", "BH", "BY", "fdr" Ver secciones anteriores), o bien utilizar una prueba diferente que tenga en cuenta este problema (Tukey, Duncan, SNK, etc).

En nuestro ejemplo de estudio, el análisis ANOVA ha señalado que el factor grupo es significativo ($p < \alpha$, rechazo de hipótesis nula), así que procedemos a realizar un prueba post-hoc MCP, como la prueba LSD. Existen muchas opciones en R, como el uso de las librerías "agricolae" o "multcompare" u otras. Para este primer análisis MCP utilizaremos la librería agricolae y la prueba MCP LSD mediante la función `LSD.test()` con diferentes correcciones si se desea:

```
library(agricolae)

LSD.test(aov.ex1, "Dosage", p.adj=c("none"), console=T, group = F)

##
## Study: aov.ex1 ~ "Dosage"
##
## LSD t Test for Alertness
##
## Mean Square Error: 24.25
##
## Dosage, means and individual ( 95 %) CI
##
## Alertness      std r      LCL      UCL Min Max
## a      32.50 6.595453 6 28.21496 36.78504 24 41
## b      28.25 4.432026 8 24.53904 31.96096 21 35
## c      19.25 1.707825 4 14.00191 24.49809 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.13145
##
## Comparison between treatments means
##
##      difference pvalue signif.      LCL      UCL
## a - b      4.25 0.1309      -1.418581  9.918581
## a - c     13.25 0.0008      ***  6.474750 20.025250
## b - c      9.00 0.0093      **   2.572434 15.427566
```



```
LSD.test(aov.ex1, "Dosage", p.adj=c("bonferroni"), console=T, group = F)
```

```
##
## Study: aov.ex1 ~ "Dosage"
##
## LSD t Test for Alertness
## P value adjustment method: bonferroni
##
## Mean Square Error: 24.25
##
## Dosage, means and individual ( 95 %) CI
##
## Alertness      std r      LCL      UCL Min Max
## a      32.50 6.595453 6 28.21496 36.78504 24 41
## b      28.25 4.432026 8 24.53904 31.96096 21 35
## c      19.25 1.707825 4 14.00191 24.49809 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.693739
##
## Comparison between treatments means
##
##      difference pvalue signif.      LCL      UCL
## a - b      4.25 0.3926      -2.9139878 11.41399
## a - c     13.25 0.0025      ** 4.6873969 21.81260
## b - c      9.00 0.0278      * 0.8768014 17.12320
```

```
LSD.test(aov.ex1, "Dosage", p.adj=c("hochberg"), group=F, console=T)
```

```
##
## Study: aov.ex1 ~ "Dosage"
##
## LSD t Test for Alertness
## P value adjustment method: hochberg
##
## Mean Square Error: 24.25
##
## Dosage, means and individual ( 95 %) CI
##
## Alertness      std r      LCL      UCL Min Max
## a      32.50 6.595453 6 28.21496 36.78504 24 41
## b      28.25 4.432026 8 24.53904 31.96096 21 35
## c      19.25 1.707825 4 14.00191 24.49809 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.13145
##
## Comparison between treatments means
##
##      difference pvalue signif.
## a - b      4.25 0.1309
```

```
## a - c      13.25 0.0025      **
## b - c       9.00 0.0185       *
```

En la función `LSD.test()` se realizó una prueba LSD para la MCP a posteriori del ANOVA significativo. Para obtener los p valores de las comparaciones se debe indicar que no se requiere grupos (`group = T`) o no (`group = F`) con la opción `group`. La prueba MCP LSD consiste en una prueba de hipótesis por parejas basada en la distribución t de Student.

Este test de comparación múltiple puede hacerse sin corrección (no recomendable) o bien con diferentes tipos de correcciones para el p-valor, como la corrección de Hochberg o de Bonferroni (muy estricta) como se ha realizado en este ejemplo.

Otras posibilidades de correcciones son: `p.adj = "holm", "hochberg", "bonferroni", "BH", "BY", "fdr"`

Los resultados nos indican que existen diferencias estadísticamente significativas entre los niveles a-c y b-c como ya se intuía en el gráfico de medias anterior

Para el caso en donde p-valor no se ha corregido:

```
a - c 13.25 0.0008239353 *** 6.474750, 20.025250
b - c 9.00 0.0092604364 ** 2.572434, 15.427566
```

Para el caso en donde p-valor corregido por el método de Bonferroni:

```
a - c 13.25 0.002472 ** 4.6873969 21.81260
b - c 9.00 0.027781 * 0.8768014 17.12320
```

Para el caso en donde p-valor corregido por el método de Bonferroni:

```
a - c 13.25 0.002472 ** 6.474750 20.025250
b - c 9.00 0.018521 * 2.572434 15.427566
```

En todos los casos los p-valores entre a-c y b-c el p-valor <0.05.

Si se desea realizar un análisis de homogeneidad de grupos debe indicarse la opción `group=T`:

```
LSD.test(aov.ex1, "Dosage", p.adj=c("none"), group=T, console=T)

##
## Study: aov.ex1 ~ "Dosage"
##
## LSD t Test for Alertness
##
## Mean Square Error: 24.25
##
## Dosage, means and individual ( 95 %) CI
##
## Alertness      std r      LCL      UCL Min Max
```

```

## a      32.50 6.595453 6 28.21496 36.78504 24 41
## b      28.25 4.432026 8 24.53904 31.96096 21 35
## c      19.25 1.707825 4 14.00191 24.49809 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.13145
##
## Groups according to probability of means differences and alpha level(
0.05 )
##
## Treatments with the same letter are not significantly different.
##
## Alertness groups
## a      32.50      a
## b      28.25      a
## c      19.25      b

LSD.test(aov.ex1, "Dosage", p.adj=c("bonferroni"), group=T, console=T)

##
## Study: aov.ex1 ~ "Dosage"
##
## LSD t Test for Alertness
## P value adjustment method: bonferroni
##
## Mean Square Error: 24.25
##
## Dosage, means and individual ( 95 %) CI
##
## Alertness      std r      LCL      UCL Min Max
## a      32.50 6.595453 6 28.21496 36.78504 24 41
## b      28.25 4.432026 8 24.53904 31.96096 21 35
## c      19.25 1.707825 4 14.00191 24.49809 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.693739
##
## Groups according to probability of means differences and alpha level(
0.05 )
##
## Treatments with the same letter are not significantly different.
##
## Alertness groups
## a      32.50      a
## b      28.25      a
## c      19.25      b

LSD.test(aov.ex1, "Dosage", p.adj=c("hochberg"), group=T, console=T)

##
## Study: aov.ex1 ~ "Dosage"

```

```

##
## LSD t Test for Alertness
## P value adjustment method: hochberg
##
## Mean Square Error: 24.25
##
## Dosage, means and individual ( 95 %) CI
##
## Alertness      std r      LCL      UCL Min Max
## a      32.50 6.595453 6 28.21496 36.78504 24 41
## b      28.25 4.432026 8 24.53904 31.96096 21 35
## c      19.25 1.707825 4 14.00191 24.49809 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of t: 2.13145
##
## Groups according to probability of means differences and alpha level(
0.05 )
##
## Treatments with the same letter are not significantly different.
##
## Alertness groups
## a      32.50      a
## b      28.25      a
## c      19.25      b

```

El análisis de homogeneidad de grupos es útil para la interpretación de los resultados. En el no aparecen los p-valores, pero si un código de grupo en donde se indican si existen diferencias al nivel 0.05:

Groups Treatments means

| | | |
|---|---|-------|
| a | a | 32.5 |
| a | b | 28.25 |
| b | c | 19.25 |

Puede verse que los niveles a y b no presentan diferencias al nivel de 0.05 pero si que a-c y b-c son diferentes.

Otro test de comparación múltiple es el test de Duncan:

```

duncan.test(aov.ex1, "Dosage", group=F, console=T)
##
## Study: aov.ex1 ~ "Dosage"
##
## Duncan's new multiple range test
## for Alertness
##
## Mean Square Error: 24.25

```

```
##
## Dosage, means
##
## Alertness      std r Min Max
## a      32.50 6.595453 6 24 41
## b      28.25 4.432026 8 21 35
## c      19.25 1.707825 4 17 21
##
## Comparison between treatments means
##
##      difference pvalue signif.      LCL      UCL
## a - b      4.25 0.1715          -2.057414 10.55741
## a - c     13.25 0.0006          ***  6.638127 19.86187
## b - c      9.00 0.0082          **   2.692586 15.30741
```

Se pueden ver las tablas de la distribución de Duncan en:

<https://www.statisticshowto.datasciencecentral.com/duncans-multiple-range-test/>

Otro test de comparación múltiple a posteriori es el test de Student-Newman-Keuls (SNK):

```
SNK.test(aov.ex1, "Dosage", group=F, console=T)

##
## Study: aov.ex1 ~ "Dosage"
##
## Student Newman Keuls Test
## for Alertness
##
## Mean Square Error: 24.25
##
## Dosage, means
##
## Alertness      std r Min Max
## a      32.50 6.595453 6 24 41
## b      28.25 4.432026 8 21 35
## c      19.25 1.707825 4 17 21
##
## Comparison between treatments means
##
##      difference pvalue signif.      LCL      UCL
## a - b      4.25 0.1715          -2.057414 10.55741
## a - c     13.25 0.0012          **   5.563531 20.93647
## b - c      9.00 0.0082          **   2.692586 15.30741
```

Al comparar las medias para que los niveles de un factor en un ANOVA, una simple comparación mediante pruebas t produce que se infle la probabilidad de declarar una diferencia significativa cuando no lo es α . Esto debido a que los intervalos de confianza se calculan con una probabilidad de cobertura determinada para cada

intervalo, pero la interpretación de la cobertura es general con respecto a todo el conjunto de intervalos.

Otra prueba MCP diferente y recomendable de utilizar es la de Tukey o HSD que se realiza mediante el comando `HSD.test`. Este procedimiento se va a desarrollar considerando en primer lugar el caso del modelo unifactorial equilibrado. En este modelo vamos a construir intervalos de confianza con coeficiente de confianza conjunto $1 - \alpha$ para todas las posibles comparaciones por parejas asociadas a los a niveles, es decir las $\binom{a}{2}$ comparaciones por parejas. El nivel de confianza conjunto $1 - \alpha$ indica que de cada 100 muestras en $(1 - \alpha)100$ de ellas, cada uno de los intervalos contiene a su correspondiente diferencia de medias. Por tanto el nivel de confianza de cada uno de los intervalos será al menos $1 - \alpha$, asegurando que no se produce el problema del inflamiento del error de tipo I.

John Tukey introdujo intervalos basados en el rango de la muestra media en lugar de las diferencias individuales. Los intervalos de esta función se basan en el rango estudentizado (Studentized range). Los intervalos construidos de esta manera solo se aplicarían exactamente a los diseños equilibrados en los que hay el mismo número de observaciones en cada nivel del factor. Esta función incorpora un ajuste de tamaño de la muestra que produce intervalos para diseños ligeramente desequilibrados.

La prueba MCP de Tuckey o HSD se realiza mediante la librería `agricolae` de la siguiente manera:

```
HSD.test(aov.ex1, "Dosage", group=F, console=T)

##
## Study: aov.ex1 ~ "Dosage"
##
## HSD Test for Alertness
##
## Mean Square Error: 24.25
##
## Dosage, means
##
## Alertness      std r Min Max
## a      32.50 6.595453 6 24 41
## b      28.25 4.432026 8 21 35
## c      19.25 1.707825 4 17 21
##
## Alpha: 0.05 ; DF Error: 15
## Critical Value of Studentized Range: 3.673378
##
## Comparison between treatments means
##
##      difference pvalue signif.      LCL      UCL
## a - b      4.25 0.2768          -2.657961 11.15796
## a - c     13.25 0.0022          * 4.993408 21.50659
## b - c      9.00 0.0237          * 1.167109 16.83289
```

una comparación de muchas metodologías MCP en R puede encontrarse en:

<http://www.jerrydallal.com/LHSP/mc.htm>

[http://eprints.aston.ac.uk/9317/1/Statnote 6.pdf](http://eprints.aston.ac.uk/9317/1/Statnote%206.pdf)

http://psych.colorado.edu/~carey/Courses/PSYC5741/handouts/Multiple_Comparison_Procedures.pdf

R también dispone en su paquete base de algunas pruebas MCP, como la de Tuckey o HSD:

```
TukeyHSD(aov.ex1)
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Alertness ~ Dosage, data = data.ex1)
##
## $Dosage
##      diff      lwr      upr    p adj
## b-a  -4.25 -11.15796  2.657961 0.2768132
## c-a -13.25 -21.50659 -4.993408 0.0022342
## c-b  -9.00 -16.83289 -1.167109 0.0237003
##?TukeyHSD
```

Otra opción también muy interesante es la de utilizar la `library(multcomp)`, por ejemplo con el test de Tukey:

```
library(multcomp) #Llibreria especial para comparacion de medianas
## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##      geyser
```

```

ex1.glht <- glht(aov.ex1, linfct = mcp(Dosage = "Tukey"))
summary(ex1.glht)

##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = Alertness ~ Dosage, data = data.ex1)
##
## Linear Hypotheses:
##      Estimate Std. Error t value Pr(>|t|)
## b - a == 0    -4.250      2.659  -1.598  0.2756
## c - a == 0   -13.250      3.179  -4.168  0.0023 **
## c - b == 0    -9.000      3.016  -2.984  0.0239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)

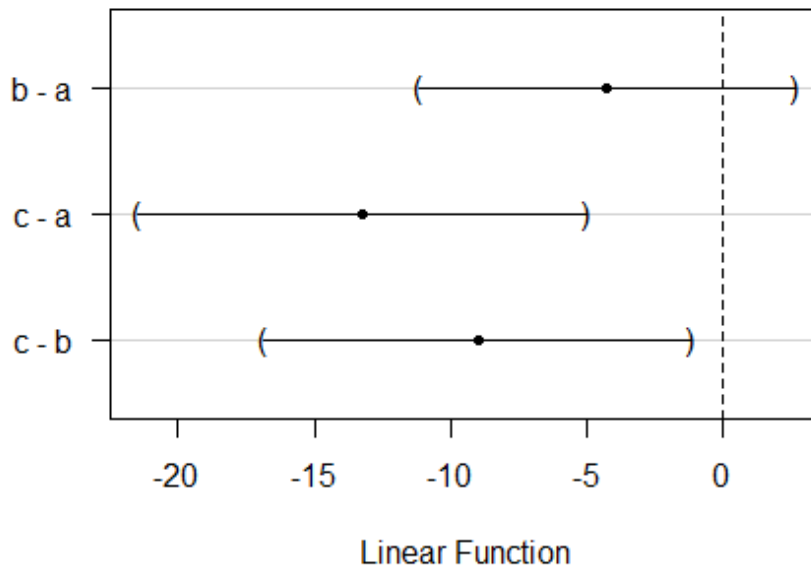
confint(ex1.glht)

##
## Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = Alertness ~ Dosage, data = data.ex1)
##
## Quantile = 2.594
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##      Estimate lwr      upr
## b - a == 0   -4.2500 -11.1487  2.6487
## c - a == 0  -13.2500 -21.4955 -5.0045
## c - b == 0   -9.0000 -16.8224 -1.1776

plot(ex1.glht)

```


95% family-wise confidence level



Un buen tutorial sobre este tema puede encontrarse en:

<http://www.jstatsoft.org/v64/i09/paper>

Pero claro, son validos los resultados de estas pruebas en el caso de incumplimiento de las pruebas? Al igual que en los apartados anteriores, tendr a que analizarse.

2.8.3 Mejora de la producci3n de calabazas con diferentes fertilizantes. C culo manual de un ANOVA

Trabajemos con el siguiente ejemplo simple, producci3n de plantas con diferentes fertilizantes:

```
fertilizer1<-c(6.27,5.36,6.39,4.85,5.99,7.14,5.08,4.07,4.35,4.95)
fertilizer2<-c(3.07,3.29,4.04,4.19,3.41,3.75,4.87,3.94,6.28,3.15)
fertilizer3<-c(4.04,3.79,4.56,4.55,4.55,4.53,3.53,3.71,7.00,4.61)
data<-data.frame(fertilizer1,fertilizer2,fertilizer3)
data
## fertilizer1 fertilizer2 fertilizer3
## 1 6.27 3.07 4.04
## 2 5.36 3.29 3.79
## 3 6.39 4.04 4.56
```

| | | | |
|-------|------|------|------|
| ## 4 | 4.85 | 4.19 | 4.55 |
| ## 5 | 5.99 | 3.41 | 4.55 |
| ## 6 | 7.14 | 3.75 | 4.53 |
| ## 7 | 5.08 | 4.87 | 3.53 |
| ## 8 | 4.07 | 3.94 | 3.71 |
| ## 9 | 4.35 | 6.28 | 7.00 |
| ## 10 | 4.95 | 3.15 | 4.61 |

Los datos representan la producción de calabazas de parcelas que recibieron tratamientos con diferentes fertilizantes. La pregunta que nos hacemos aquí es: ¿tiene el fertilizante un efecto en la producción de calabaza?



Imagen de Wikipedia(<https://en.wikipedia.org/wiki/Pumpkin>)

Para responder a esta pregunta, necesitamos probar la hipótesis de la media igual y para eso tenemos que calcular el estadístico-F.

La parte de la base matemático-estadística del ANOVA puede verse al final del manual en el apartado de teoría. Se resumen en la denominada tabla ANOVA

¿Cómo Podemos calcular manualmente el estadístico F?

Primeramente, tenemos que calcular SS_{total} , $SS_{between}$ and SS_{within} . Usando la siguiente formula para poder calcularlos:

$$SS_{total} = \sum_1^{an} (y_{ij} - y_{..})^2 = \sum_1^{an} y_{ij}^2 - any_{..}^2$$

$$SS_{between} = \sum_1^{an} (y_i - y_{..})^2 = n \sum_1^a y_i^2 - any_{..}^2$$

$$SS_{within} = SS_{total} - SS_{between}$$

¿Cuánto valen SS_{total} , $SS_{between}$ and SS_{within} en nuestro ejemplo de los fertilizantes?

```

mean_f1<-mean(fertilizer1)
mean_f2<-mean(fertilizer2)
mean_f3<-mean(fertilizer3)
mean_all<-mean(c(mean_f1,mean_f2,mean_f3))

n1<-length(fertilizer1)
n2<-length(fertilizer2)
n3<-length(fertilizer3)

SST<-sum(sum(fertilizer1^2),sum(fertilizer2^2),sum(fertilizer3^2))
- 3*n1*(mean_all^2)
SST
## [1] 36.4449

SSB<-(n1*sum((mean_f1^2),(mean_f2^2),(mean_f3^2))) -
3*n1*(mean_all^2)
SSB
## [1] 10.82275

SSW<-SST-SSB
SSW
## [1] 25.62215

```

Grados de libertad

In order to fill the next column of de la table ANOVA donde conocemos que (df = grados de libertad):

$$df_{between} = a - 1$$

$$df_{within} = an - a$$

$$df_{total} = an - 1$$

Calculamos los df y vamos completando la table ANOVA (ver al final del ejercicio).

```

df_SST<-(3*10)-1
df_SST
## [1] 29

```

```
df_SSB<-3-1
df_SSB

## [1] 2

df_SSW<-(3*10)-3
df_SSW
```

```
## [1] 27
```

Cálculo de las sumas de cuadrados medios MS_{within} and $MS_{between}$

Es muy simple obtener MS_{within} (Mean Square Within) y $MS_{between}$ (Mean Square Between), solo necesitamos dividir SS_{within} y $SS_{between}$ por sus respectivos df.

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

Calculamos $MS_{between}$ y MS_{within} y vamos completando la table ANOVA (ver al final del ejercicio).

```
MSB<-SSB/df_SSB
MSB
```

```
## [1] 5.411373
```

```
MSW<-SSW/df_SSW
MSW
```

```
## [1] 0.9489685
```

F ratio o F-statistic o estadístico F

Puede pensarse en el estadística F como un valor de lo lejos que estamos de la hipótesis nula, cuanto más grande sea F, más relevante será el efecto de un fertilizante. La estadística F está dada por la siguiente fórmula:

$$F = \frac{MS_{between}}{MS_{within}}$$

Completamos la table ANOVA calculando F.

```
F<-MSB/MSW
F
```

```
## [1] 5.702374
```

Cálculo de la significación del estadístico F

Cuando se ha calculado el estadístico F, es necesario conocer si el resultado obtenido es significativo o no. Asumiendo la igualdad de media como hipótesis nula, se puede conocer la probabilidad de obtener un valor F o ser superior a este, de esta manera podemos determinar si debemos rechazar la hipótesis nula o no.

```
#EXAMPLE:
a<-3
n<-10
#then
1-pf(F,df1=a-1,df2=a*n-a)
```

Si se elige un valor crítico de, por ejemplo, $\alpha = 0.05$ y obtiene un p valor inferior a 0.05, se rechazará la hipótesis nula de la igualdad de la media y se concluirá que no todas las medias comparadas son iguales. No sabremos cuál de las parejas de medias del grupo es diferente, solo sabemos que no todas las parejas comparados son iguales. Para saber que grupos son diferentes, utilizaremos una prueba post hoc. Esto se explicará en la siguiente sección.

¿Qué valor p se obtuvo? ¿Qué puede concluirse?

```
p_val<-1-pf(F,df1=3-1,df2=3*10-3)
p_val

## [1] 0.008594377

## p_val < 0.05, Rechazamos la hipótesis nula, las medias grupales son
diferentes.
```

La función ANOVA en R

Hasta ahora hemos intentado comprender cómo funciona el análisis de un factor del análisis ANOVA y cómo podemos calcularlo manualmente. ¡De ahora en adelante, puede usar una función R para ser más rápido cuando analices tus propios datos!

Primero, necesitamos un formato especial de los datos para trabajar con R, se llama datos de formato largo. Usaremos la librería. “tidyr” para formatear los datos:

```
library("tidyr")
fertilizer_data<-
gather(data,fertilizer,production,fertilizer1:fertilizer3)
fertilizer_data

##      fertilizer production
## 1 fertilizer1      6.27
## 2 fertilizer1      5.36
## 3 fertilizer1      6.39
## 4 fertilizer1      4.85
## 5 fertilizer1      5.99
## 6 fertilizer1      7.14
```

```
## 7 fertilizer1      5.08
## 8 fertilizer1      4.07
## 9 fertilizer1      4.35
## 10 fertilizer1     4.95
## 11 fertilizer2     3.07
## 12 fertilizer2     3.29
## 13 fertilizer2     4.04
## 14 fertilizer2     4.19
## 15 fertilizer2     3.41
## 16 fertilizer2     3.75
## 17 fertilizer2     4.87
## 18 fertilizer2     3.94
## 19 fertilizer2     6.28
## 20 fertilizer2     3.15
## 21 fertilizer3     4.04
## 22 fertilizer3     3.79
## 23 fertilizer3     4.56
## 24 fertilizer3     4.55
## 25 fertilizer3     4.55
## 26 fertilizer3     4.53
## 27 fertilizer3     3.53
## 28 fertilizer3     3.71
## 29 fertilizer3     7.00
## 30 fertilizer3     4.61
```

Una vez que tengamos el buen formato de los datos, usaremos la función R `aov()`. A ver si completamos la tabla de resumen de ANOVA correctamente:

```
anova_out<-aov(production~fertilizer,fertilizer_data)
# production represents the column with numeric variables
# fertilizer represents the factors
# fertilizer_data is the name of the data frame where you will find the
columns
# named production fertilizer
anova_out

## Call:
##   aov(formula = production ~ fertilizer, data = fertilizer_data)
##
## Terms:
##           fertilizer Residuals
## Sum of Squares    10.82275  25.62215
## Deg. of Freedom         2      27
##
## Residual standard error: 0.9741502
## Estimated effects may be unbalanced
```

El resultado obtenido por la función `aov()` es una table donde se puede encontrar en la primera fila: $SS_{between}$ (primera columna) y SS_{within} (también llamada Residuals en la segunda columna). En la segunda fila pueden encontrarse los grados de libertad

(degrees of freedom) de $SS_{between}$ (a-1) y SS_{within} (an-a). Probablemente se ha observado que no se obtiene ningún estadístico F ni p-valor en la salida. Necesitamos usar la función `summary()` para obtener ambos.

```
summary(anova_out)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## fertilizer  2  10.82   5.411   5.702 0.00859 **
## Residuals  27  25.62   0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Puede observarse que se obtiene el mismo resultado para el valor de F ($F=5.702$) que el obtenido de manera manual.

Comparación multiple “a posteriori” (t-tests PAIRWISE) de un ANOVA significativo

Una vez que se obtiene un resultado significativo con el ANOVA “One -way”, solo podemos decir que al menos una media de la población es diferente de al menos otra media de la población (pairwise). Ahora, aprenderemos cómo saber cuál es la media de la población que es diferente de las demás.

Para eso usaremos pruebas t pareadas con el ajuste p de Bonferroni⁵. Se utiliza el siguiente comando R:

```
pairwise.t.test(fertilizer_data$production, fertilizer_data$fertilizer, p.adjust.method = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data: fertilizer_data$production and fertilizer_data$fertilizer
##
##           fertilizer1 fertilizer2
## fertilizer2 0.0078      -
## fertilizer3 0.1099      0.8175
##
## P value adjustment method: bonferroni
```

Así podemos concluir que existen diferencias estadísticamente significativas entre el fertilizante 1 y el 2, pero no entre el 1 y el 3 y entre el 3 y el 2. Si se utilizará otro tipo de ajuste del p-valor podrían obtenerse otros resultados (Ver en otras secciones).

⁵ Este ajuste es el más antiguo y sencillo de explicar. Posteriormente en otros ejemplos se ofrecen otras soluciones que presentan un mejor balance entre el incremento de alfa y la pérdida de potencia estadística cuando se utiliza este tipo de ajuste.

Intervalo de confianza para todas las diferencias de medias (pairwise differences)

Siguiendo nuestro ejemplo de trabajo de producción de calabazas en relación con diferentes fertilizantes, se pueden calcular los intervalos de confianza (IC) para todas las posibles diferencias de medias. Lo que es simplemente calcular el IC de la diferencia de dos medias de población (todas las combinaciones posibles). En el caso de los fertilizantes, sería calcular el IC de:

$$\mu_{fertilizer1} - \mu_{fertilizer2}$$

$$\mu_{fertilizer1} - \mu_{fertilizer3}$$

$$\mu_{fertilizer2} - \mu_{fertilizer3}$$

A $1 - \alpha$ CI para una diferencia entre dos medias de la población está dada por la fórmula:

$$\bar{y}_i - \bar{y}_j \pm t_{\alpha/2} \times SE(\bar{y}_i - \bar{y}_j)$$

Donde \bar{y}_i es la media del grupo i , $t_{\alpha/2}$ es el t-value para un valor específico de $1 - \alpha$ (habitualmente 0.95) y SE es el error estándar de la diferencia de medias. Hay que ajustar el tipo de error I para múltiples pruebas y usar la SE apropiada de la diferencia de medios. ($0.05/3/2=0.008$) Hay que tener en cuenta que:

$$SE(\bar{y}_i - \bar{y}_j) = \sqrt{MS_{within} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Donde n_i y n_j son los tamaños muestrales de sus grupos respectivos i y j .

¡Importante! Cuando se calcule $t_{\alpha/2}$, los grados de libertad vienen dados por los de MS_{within} que son $(an - a)$.

Se puede calcular el CI para cada uno de los pares (pairwise comparisons). Primero hay que obtener el t-value utilizando una función R:

```
#qt(0.975, df=an-a)
t<-qt(1-0.05/2/3, df=10*3-3)
```

Posteriormente se obtendrá el CI de todos los pares.

```
high_CI_f1f2<-(mean_f1-mean_f2)+(t*sqrt(MSW*(1/10+1/10)))
low_CI_f1f2<-(mean_f1-mean_f2)-(t*sqrt(MSW*(1/10+1/10)))
c(low_CI_f1f2,high_CI_f1f2)

## [1] 0.3340132 2.5579868
```



```

high_CI_f1f3<-(mean_f1-mean_f3)+(t*sqrt(MSW*(1/10+1/10)))
low_CI_f1f3<-(mean_f1-mean_f3)-(t*sqrt(MSW*(1/10+1/10)))
c(low_CI_f1f3,high_CI_f1f3)

## [1] -0.1539868  2.0699868

high_CI_f2f3<-(mean_f2-mean_f3)+(2.051831*sqrt(MSW*(1/10+1/10)))
low_CI_f2f3<-(mean_f2-mean_f3)-(2.051831*sqrt(MSW*(1/10+1/10)))
c(low_CI_f2f3,high_CI_f2f3)

## [1] -1.3818867  0.4058867

```

Como puede observarse el resultado es paralelo al obtenido con el pairwise t-test “a posteriori” del ANOVA realizado anteriormente.

2.8.4 Diversidad biológica y contaminación de zinc en las Montañas Rocosas (USA)

Este ejemplo está fundado en los que se explican en las clases de estadística y que se basan en los diferentes modelos creados por Hafid Laayouni y colaboradores de la UPF (Barcelona).

Medley y Clements (1998) investigaron el impacto de la contaminación por zinc (y otros metales pesados) en la diversidad de especies de diatomeas en las Montañas Rocosas de los Estados Unidos. La diversidad de diatomeas (número de especies) y el grado de contaminación por zinc (categorizados como alto, medio, bajo o nivel de fondo natural) se registraron entre cuatro y seis estaciones de muestreo dentro de cada una de los seis ríos que se sabe están contaminadas. Estos datos se utilizaron para probar la hipótesis nula de que no había diferencias en la diversidad de diatomeas entre los diferentes niveles de zinc. Puedes encontrar los datos en el archivo contaminación.txt.

- 1) Evalúa la normalidad / homogeneidad de la varianza utilizando un “box-plot” de la diversidad de especies contra cada grupo de zinc.
- 2) Realiza un test estadístico para comprobar la hipótesis (H_0) que la diversidad de especies es igual para todos los grupos de nivel de zinc (categorizados como alto, medio, bajo o nivel de fondo natural). ¿Cuál es tu conclusión?
- 3) Si es necesario, realiza una prueba post hoc para investigar las diferencias de medias entre todos los grupos.

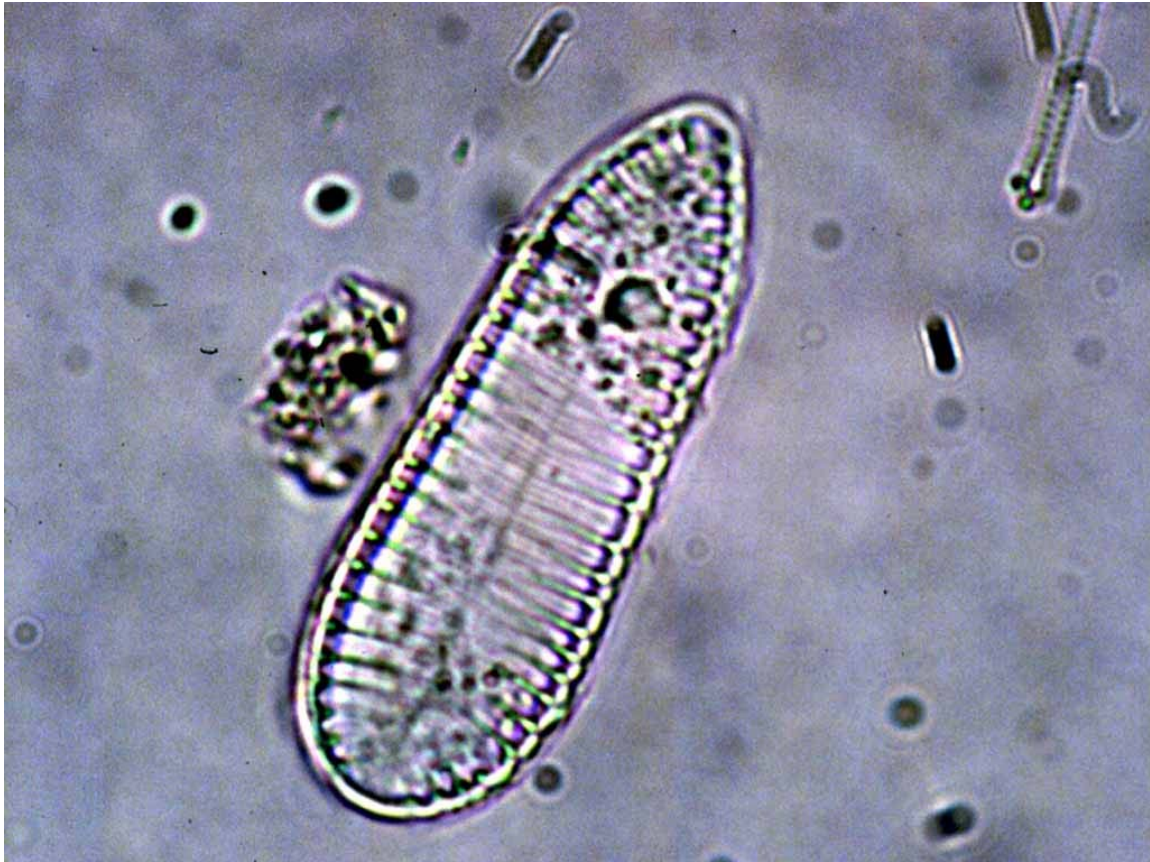
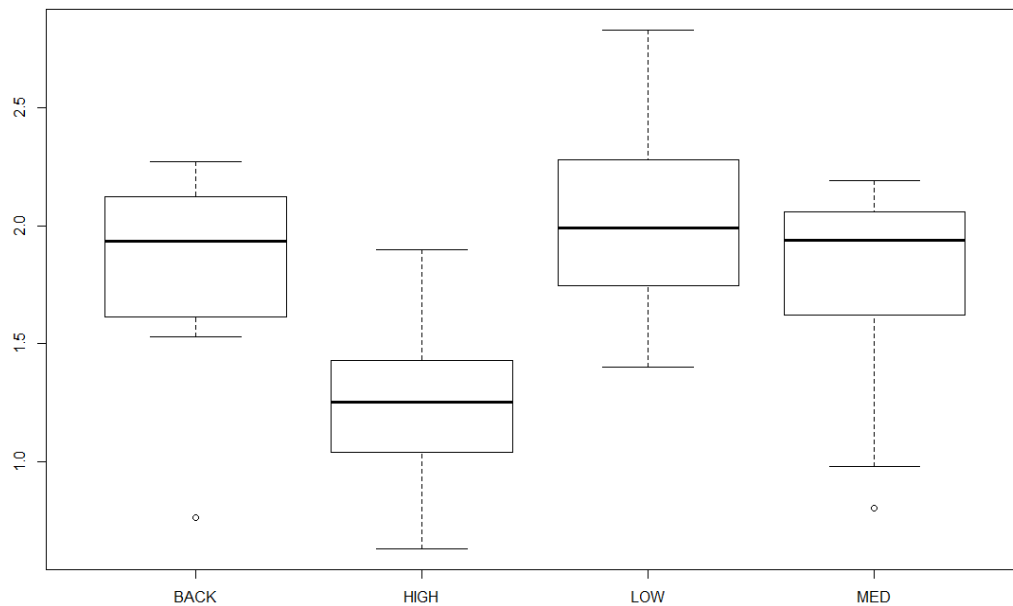


Imagen de una diatomea (Wikipedia,
<https://es.wikipedia.org/wiki/Diatomea#/media/File:Surirelasublinearis.jpg>)

Obtenemos los datos del siguiente fichero (ver dirección de los datos en Hithub)

```
data_zinc<-read.table("contamination.txt",header=T)
boxplot(data_zinc$DIVERSITY~data_zinc$ZINC,data = data_zinc)
```



No se observan dispersión asimétricas en los “box-plots” ni tampoco grandes diferencias entre los anchos de las diferentes cajas, descartando no normalidad ni heterocedasticidad, aunque se debería comprobar a posteriori del ANOVA mediante el análisis de los residuos.

Suponiendo que se cumplen las condiciones de aplicación del test ANOVA, este se puede realizar mediante:

```

anova_out_cont<-aov(data_zinc$DIVERSITY~data_zinc$ZINC,data_zinc)
summary(anova_out_cont)

##              Df Sum Sq Mean Sq F value Pr(>F)
## data_zinc$ZINC  3  2.567  0.8555   3.939 0.0176 *
## Residuals     30  6.516  0.2172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Efectivamente (suponiendo alfa = 0.05, p-valor =0.0176) puede rechazarse la hipótesis nula de que la diversidad de especies es igual para todos los grupos de nivel de zinc (categorizados como alto, medio, bajo o nivel de fondo natural).

Ahora investigamos a posteriori del ANOVA las diferencias entre pares de medias de los grupos de interés (alto, medio, bajo o nivel de fondo natural) mediante un test de medias `pairwise.t.test()`

```
pairwise.t.test(data_zinc$DIVERSITY,data_zinc$ZINC,p.adjust.method =
"bonf")
```

Obteniendo el siguiente resultado:

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: data_zinc$DIVERSITY and data_zinc$ZINC
##
##      BACK  HIGH  LOW
## HIGH 0.173 -    -
## LOW  1.000 0.014 -
## MED  1.000 0.326 1.000
##
## P value adjustment method: bonferroni
```

There are only difference in population means between HIGH and LOW Zinc groups

Como resultado, sólo se detecta una única diferencia en la media de la población entre los grupos de zinc alto y bajo($p=0.014 < \alpha=0.05$).

2.8.5 Mejora de la resistencia de los tejidos de algodón mediante diseño experimental

A continuación, se presenta un ejemplo clásico, utilizado en clase para introducir los modelos lineales y el análisis ANOVA. Utilizado durante muchos cursos y que nos ha dado siempre buen resultado en las clases de diseño de experimentos y análisis de datos de la Facultad de Biología de la UB.



Imagen obtenida en: <http://www.i4english.es/images/pic/-Algodon-n-De-Color-Rosa-Collar-De-Bordados-1891.jpg>

Un ingeniero de desarrollo de productos está interesado en maximizar la resistencia a la tensión de una nueva fibra sintética que se empleará en la manufactura de tela para camisas de hombre. El ingeniero sabe por experiencia que la resistencia está influida por el porcentaje de algodón presente en la fibra. Además, sospecha que el contenido de algodón debe estar aproximadamente entre un 10 y 40% para que la tela resultante tenga otras características de calidad que se desean (como la capacidad de recibir un tratamiento de planchado permanente). El ingeniero decide probar muestras a cinco niveles de porcentaje de algodón: 15, 20, 25, 30 y 35%. Asimismo, decide ensayar cinco muestras a cada nivel de contenido de algodón. Probar si existe relación entre la resistencia y el % de algodón.

```
#####  
# EJEMPLO 0: ALGODÓN resistencia a La tensión de una nueva fibra  
sintética  
#####  
  
# Pequeños ajustes del entorno R. Cuestión muy técnica, nos Lo creemos.  
# ejecutar  
options(show.signif.stars=FALSE, contrasts=c("contr.sum", "contr.poly"))
```

```

# Lectura "sencilla" de Los datos
Datos <- read.table("coto.txt", header = T)
Datos

##      grupo medida
## 1      15      7
## 2      15      7
## 3      15     15
## 4      15     11
## 5      15      9
## 6      20     12
## 7      20     17
## 8      20     12
## 9      20     18
## 10     20     18
## 11     25     14
## 12     25     18
## 13     25     18
## 14     25     19
## 15     25     19
## 16     30     19
## 17     30     25
## 18     30     22
## 19     30     19
## 20     30     23
## 21     35      7
## 22     35     10
## 23     35     11
## 24     35     15
## 25     35     11

#lectura con mas opciones de archivo de datos de algodón
Datos <- read.table("coto.txt",
header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)

#etiquetar los niveles del factor
Datos$grupo <- factor(Datos$grupo,
labels=c('15%', '20%', '25%', '30%', '35%'))
Datos

##      grupo medida
## 1     15%      7
## 2     15%      7
## 3     15%     15
## 4     15%     11
## 5     15%      9
## 6     20%     12
## 7     20%     17
## 8     20%     12
## 9     20%     18
## 10    20%     18

```

```

## 11 25% 14
## 12 25% 18
## 13 25% 18
## 14 25% 19
## 15 25% 19
## 16 30% 19
## 17 30% 25
## 18 30% 22
## 19 30% 19
## 20 30% 23
## 21 35% 7
## 22 35% 10
## 23 35% 11
## 24 35% 15
## 25 35% 11

# Análisis de la varianza sobre estos datos
# Explicación:
# Como primer argumento pasamos una fórmula. En R un objeto de clase
'formula'
# Siempre tiene la forma 'variable de respuesta' ~ 'expresión'
# En expresión 'intervienen diferentes variables, a partir de las cuales
# Quisiéramos explicar los valores de la variable de respuesta.
# Como segundo argumento indicamos qué 'data.frame' están las variables
que
# Corresponden a los nombres de variable que se han utilizado en la
fórmula.

# Dos maneras de realizarlo serían
anova(lm(medida ~ grupo, data=Datos)) #ANOVA

## Analysis of Variance Table
##
## Response: medida
##          Df Sum Sq Mean Sq F value    Pr(>F)
## grupo      4 475.76  118.94  14.757 9.128e-06
## Residuals 20 161.20    8.06

# más frecuentemente:
Coto_a.aov <- aov(medida ~ grupo, data = Datos)
# También sería lo mismo:
Coto_a.aov <- aov(medida ~ ., data = Datos)

# 'Coto_a.aov' es un objeto complejo que contiene la información más
esencial
# Del ajuste del modelo lineal 'MEDIDA ~ GRUPO = factor'.
# Si pedimos que nos lo muestre genera
# Un resumen bastante poco informativo:
Coto_a.aov

```

```

## Call:
## aov(formula = medida ~ ., data = Datos)
##
## Terms:
##          grupo Residuals
## Sum of Squares 475.76    161.20
## Deg. of Freedom    4         20
##
## Residual standard error: 2.839014
## Estimated effects may be unbalanced

# Hay diferentes funciones que nos permiten extraer La información sobre
# Diferentes aspectos importantes del ajuste anterior.
# Muestra La tabla ANOVA asociada al ajuste anterior:
summary(Coto_a.aov)

##          Df Sum Sq Mean Sq F value    Pr(>F)
## grupo      4 475.8  118.94   14.76 9.13e-06
## Residuals 20 161.2    8.06

# Tabla de medias:
model.tables(Coto_a.aov, type="mean")

## Tables of means
## Grand mean
##
## 15.04
##
## grupo
## grupo
## 15% 20% 25% 30% 35%
## 9.8 15.4 17.6 21.6 10.8

tapply(Datos$medida, Datos$grupo, mean, na.rm=TRUE) # means

## 15% 20% 25% 30% 35%
## 9.8 15.4 17.6 21.6 10.8

tapply(Datos$medida, Datos$grupo, sd, na.rm=TRUE) # std. deviations

##      15%      20%      25%      30%      35%
## 3.346640 3.130495 2.073644 2.607681 2.863564

tapply(Datos$medida, Datos$grupo, function(x) sum(!is.na(x))) # counts,
tamaño de Los grupos

## 15% 20% 25% 30% 35%
## 5 5 5 5 5

tapply(Datos$medida, Datos$grupo, median, na.rm=TRUE) #median

## 15% 20% 25% 30% 35%
## 9 17 18 22 11

```



```

#PLOT MEANS -> REPRESENTACION de Las medias:
library(Rcmdr)

## Warning: package 'Rcmdr' was built under R version 3.5.3
## Loading required package: splines
## Loading required package: RcmdrMisc
## Loading required package: car
## Loading required package: carData
## Loading required package: sandwich
## Loading required package: effects
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

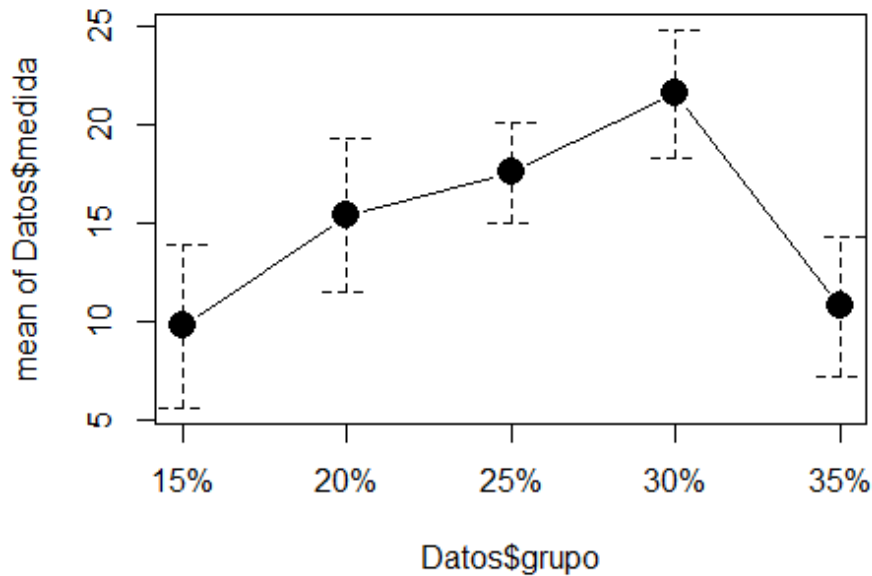
## La interfaz R-Commander sólo funciona en sesiones interactivas
##
## Attaching package: 'Rcmdr'

## The following object is masked from 'package:car':
##
##   Confint

plotMeans(Datos$medida, Datos$grupo, error.bars="conf.int",
level=0.95) #INTERVALOS DE CONFIANZA deL 95%

```

Plot of Means



```
# Coeficientes del modelo (efectes_i o alfes_i):
coef(Coto_a.aov)

## (Intercept)      grupo1      grupo2      grupo3      grupo4
##      15.04      -5.24       0.36       2.56       6.56

# Más detallado, muestra también los coeficientes "virtuales", que no
# están realmente
# Almacenados dentro 'Coto_a.aov' debido a restricciones como que deben
# sumar
# Cero:
dummy.coef(Coto_a.aov)

## Full coefficients are
##
## (Intercept):      15.04
## grupo:           15%   20%   25%   30%   35%
##                 -5.24  0.36  2.56  6.56 -4.24

# Residuos (errores: eij):
residuals(Coto_a.aov)

##   1   2   3   4   5   6   7   8   9  10  11  12  13  14
## 15
## -2.8 -2.8  5.2  1.2 -0.8 -3.4  1.6 -3.4  2.6  2.6 -3.6  0.4  0.4  1.4
## 1.4
## 16  17  18  19  20  21  22  23  24  25
## -2.6  3.4  0.4 -2.6  1.4 -3.8 -0.8  0.2  4.2  0.2
```

```

#ANOVA - GRÁFICOS DE DIAGNOSTICO (VALIDEZ DEL MODELO)
Coto_a.aov <- aov(medida ~ grupo, data = Datos)
Coto_a.aov

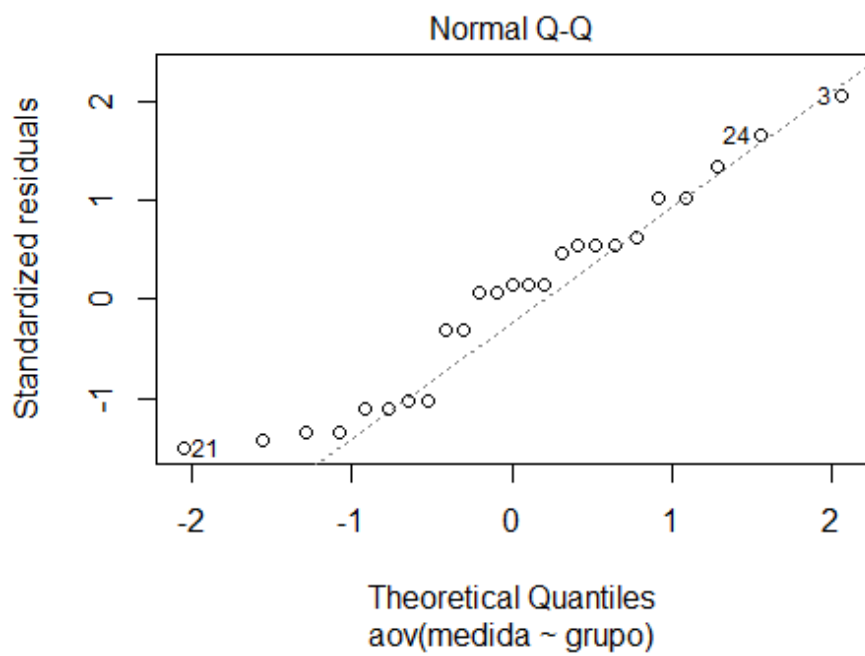
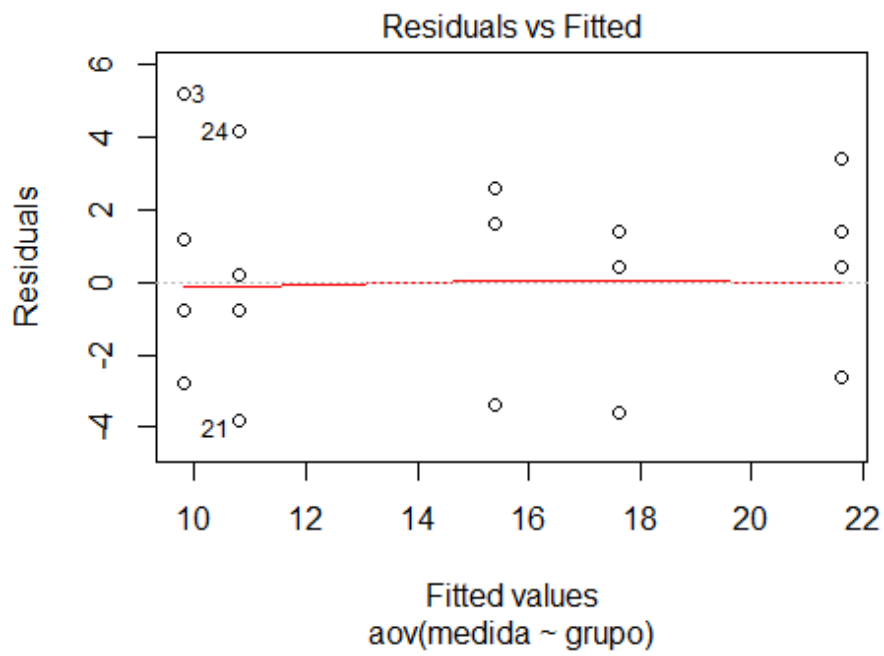
## Call:
## aov(formula = medida ~ grupo, data = Datos)
##
## Terms:
##          grupo Residuals
## Sum of Squares 475.76   161.20
## Deg. of Freedom    4      20
##
## Residual standard error: 2.839014
## Estimated effects may be unbalanced

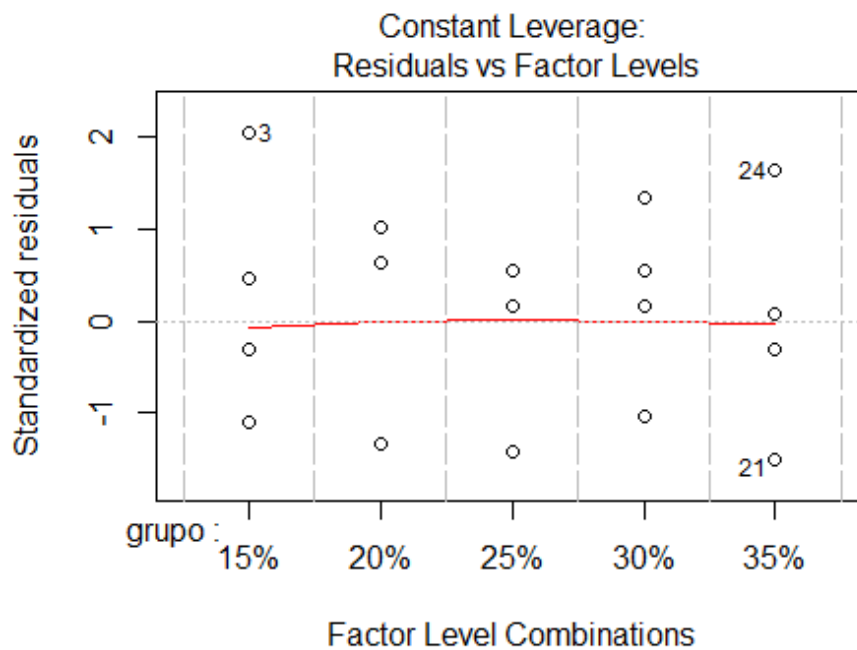
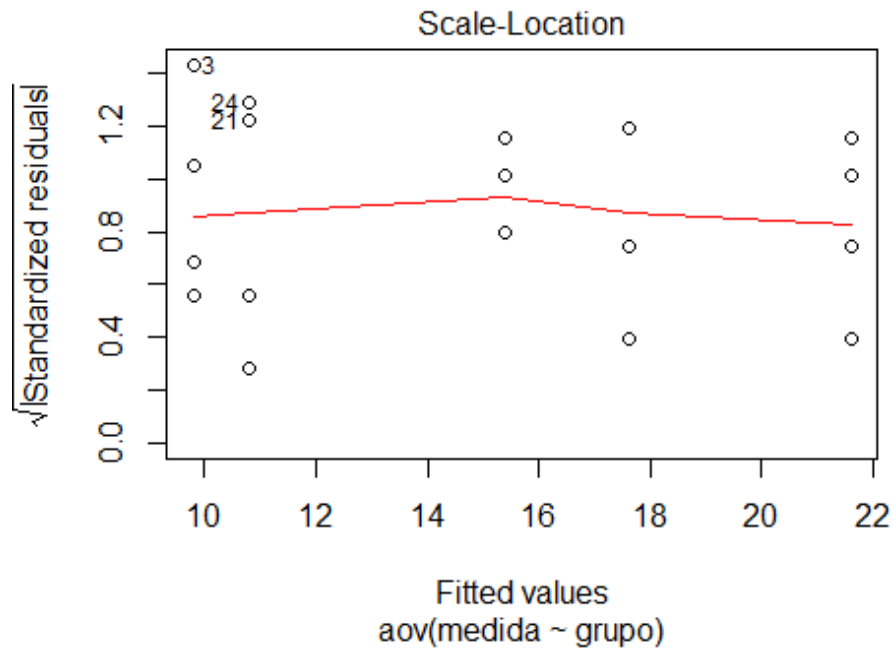
summary(Coto_a.aov)

##          Df Sum Sq Mean Sq F value    Pr(>F)
## grupo      4 475.8  118.94   14.76 9.13e-06
## Residuals 20 161.2    8.06

# Vayan pulsando 'INTRO' para ver aparecer varios gráficos de
# diagnóstico,
# en particular en el gráfico de probabilidad (qqplot) los residuos
# aparecen
# razonablemente alineados --> sugerencia de normalidad
#http://stat.ethz.ch/R-manual/R-patched/Library/stats/html/plot.Lm.html
plot(Coto_a.aov)

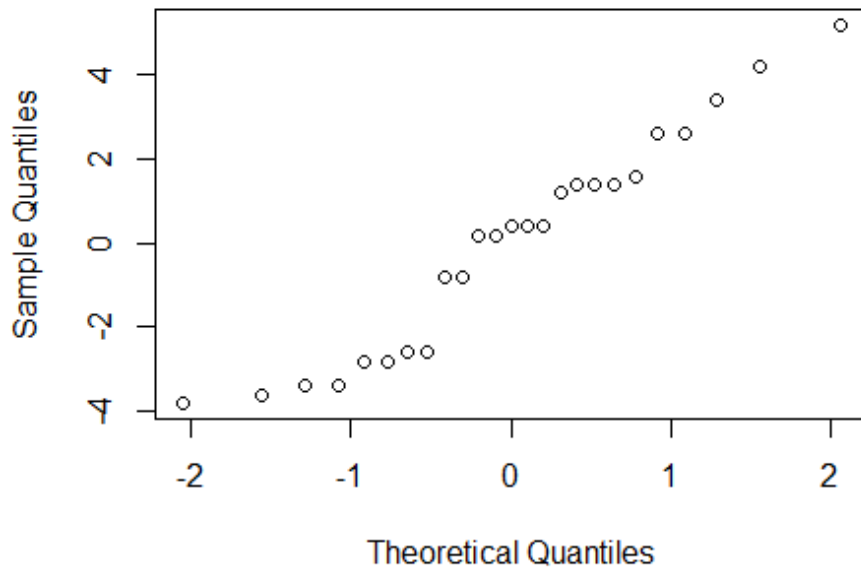
```





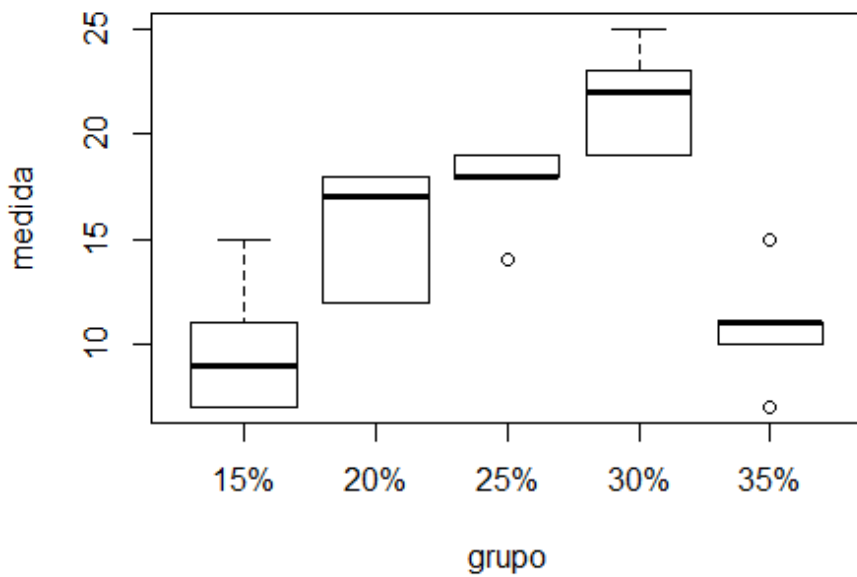
Si únicamente estuvieran interesados en un gráfico QQ de Los residuos:
`qqnorm(residuals(Coto_a.aov)) #veure.`
<http://en.wikipedia.org/wiki/Q%E2%80%93plot>

Normal Q-Q Plot



#OBSERVACIONES EXTREMAS CON UN BOX-PLOT

```
boxplot(medida~grupo, ylab="medida", xlab="grupo", data=Datos)
```



```

#homoescedasticidad: Homogeneity of Variance
tapply(Datos$medida, Datos$grupo, var, na.rm=TRUE)

## 15% 20% 25% 30% 35%
## 11.2 9.8 4.3 6.8 8.2

bartlett.test(medida ~ grupo, data=Datos)

##
## Bartlett test of homogeneity of variances
##
## data: medida by grupo
## Bartlett's K-squared = 0.93309, df = 4, p-value = 0.9198

#otro método diferente
library(lawstat)

## Warning: package 'lawstat' was built under R version 3.5.3

##
## Attaching package: 'lawstat'

## The following object is masked from 'package:car':
##
##     levene.test

tapply(Datos$medida, Datos$grupo, var, na.rm=TRUE)

## 15% 20% 25% 30% 35%
## 11.2 9.8 4.3 6.8 8.2

levene.test(Datos$medida, Datos$grupo)
#http://rss.acs.unt.edu/Rdoc/Library/Lawstat/html/Levene.test.html

##
## Modified robust Brown-Forsythe Levene-type test based on the
## absolute deviations from the median
##
## data: Datos$medida
## Test Statistic = 0.31795, p-value = 0.8626

#Si los datos no se dan las condiciones de NORMALIDAD NI
HOMOESDASTICITAT???????
#TEST DE KRUSKALL WALLIS: NO SE SUPONE NORMALIDAD NI HOMOESDASTICITAT
kruskal.test(medida ~ grupo, data=Datos)

##
## Kruskal-Wallis rank sum test
##
## data: medida by grupo
## Kruskal-Wallis chi-squared = 19.064, df = 4, p-value = 0.0007636

```

```

#AQUI SE PUEDE VER UN EJEMPLO DE DIAGNOSTICS COMPLETO:
http://itl.nist.gov/div898/handbook/pri/section2/pri24.htm

#comparacions múltiples: diferencia de medias
# Si el ANOVA indica que el factor "grupo" es significativo, procedemos a
Realizar pruebas post-hoc de comparación múltiple de tipo T-test
require(agricolae)

## Loading required package: agricolae

LSD.test(Coto_a.aov, "grupo", p.adj=c("none"))
#?LSD.test
LSD.test(Coto_a.aov, "grupo", p.adj=c("bonferroni"))

#En la función LSD.test() se realizó la comparación múltiple. Para
obtener las probabilidades de las comparaciones y se debe indicar que no
se requiere grupos,
# compara <- LSD.test(yield, virus,df, MSerror, group=F)
LSD.test(Coto_a.aov, "grupo", p.adj=c("none"), group=F)
LSD.test(Coto_a.aov, "grupo", p.adj=c("bonferroni"), group=F)

#Duncan:
duncan.test(Coto_a.aov, "grupo", group=F)
#ver tablas de Duncan en:
http://cse.niaes.affrc.go.jp/miwa/probcalc/duncan/index.html

#SNK
SNK.test(Coto_a.aov, "grupo", group=F)

#TukeyHSD
HSD.test(Coto_a.aov, "grupo", group=F)

# Una comparativa de todas las metodologías post-hoc de comparación
multiple puede verse en: http://www.jerrydallal.com/LHSP/mc.htm
# Una comparativa a: http://eprints.aston.ac.uk/9317/1/Statnote\_6.pdf
#
http://psych.colorado.edu/~carey/Courses/PSYC5741/handouts/Multiple\_Comparison\_Procedures.pdf

#Otras maneras de hacer las comparaciones multiples en R, por ejemplo con
el test de Tukey:

TukeyHSD(Coto_a.aov)

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = medida ~ grupo, data = Datos)
##

```



```

## $grupo
##          diff          lwr          upr          p adj
## 20%-15%    5.6    0.2270417 10.9729583 0.0385024
## 25%-15%    7.8    2.4270417 13.1729583 0.0025948
## 30%-15%   11.8    6.4270417 17.1729583 0.0000190
## 35%-15%    1.0   -4.3729583  6.3729583 0.9797709
## 25%-20%    2.2   -3.1729583  7.5729583 0.7372438
## 30%-20%    6.2    0.8270417 11.5729583 0.0188936
## 35%-20%   -4.6   -9.9729583  0.7729583 0.1162970
## 30%-25%    4.0   -1.3729583  9.3729583 0.2101089
## 35%-25%   -6.8  -12.1729583 -1.4270417 0.0090646
## 35%-30% -10.8 -16.1729583 -5.4270417 0.0000624

#OTRA OPCION UTILIZANDO LA LIBRERIA multcomp():
require(multcomp) #Librería especial para comparación de medias

## Loading required package: multcomp
## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##      geyser

#?gLht
Coto_a.gLht <- glht(Coto_a.aov, linfct = mcp(grupo = "Tukey"))
summary(Coto_a.gLht)

##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = medida ~ grupo, data = Datos)
##
## Linear Hypotheses:
##              Estimate Std. Error t value Pr(>|t|)
## 20% - 15% == 0    5.600      1.796   3.119 0.03857
## 25% - 15% == 0    7.800      1.796   4.344 0.00256
## 30% - 15% == 0   11.800      1.796   6.572 < 0.001
## 35% - 15% == 0    1.000      1.796   0.557 0.97977
## 25% - 20% == 0    2.200      1.796   1.225 0.73725

```

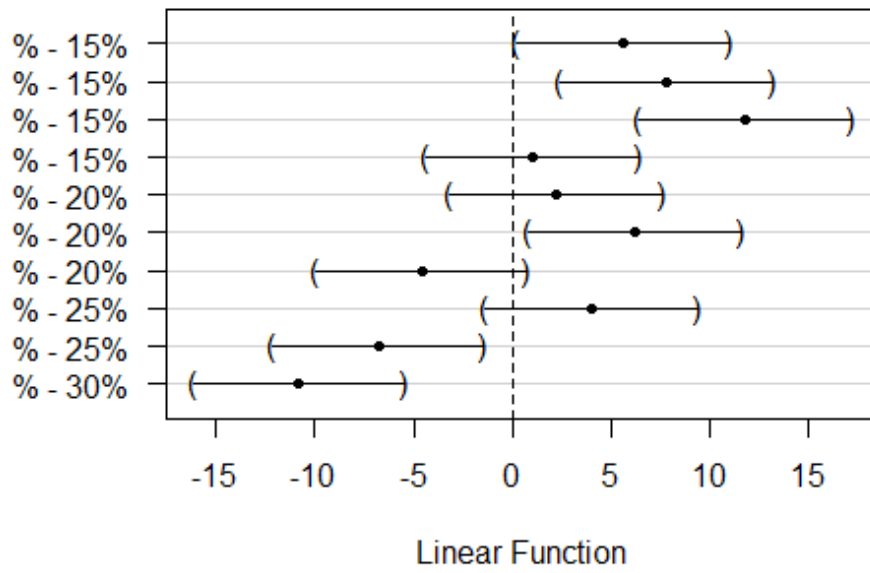
```
## 30% - 20% == 0    6.200    1.796    3.453    0.01890
## 35% - 20% == 0   -4.600    1.796   -2.562    0.11623
## 30% - 25% == 0    4.000    1.796    2.228    0.21002
## 35% - 25% == 0   -6.800    1.796   -3.787    0.00908
## 35% - 30% == 0  -10.800    1.796   -6.015 < 0.001
## (Adjusted p values reported -- single-step method)
```

```
confint(Coto_a.glht)
```

```
##
## Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = medida ~ grupo, data = Datos)
##
## Quantile = 2.9929
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##           Estimate lwr      upr
## 20% - 15% == 0    5.6000    0.2261  10.9739
## 25% - 15% == 0    7.8000    2.4261  13.1739
## 30% - 15% == 0   11.8000    6.4261  17.1739
## 35% - 15% == 0    1.0000   -4.3739   6.3739
## 25% - 20% == 0    2.2000   -3.1739   7.5739
## 30% - 20% == 0    6.2000    0.8261  11.5739
## 35% - 20% == 0   -4.6000   -9.9739   0.7739
## 30% - 25% == 0    4.0000   -1.3739   9.3739
## 35% - 25% == 0   -6.8000  -12.1739  -1.4261
## 35% - 30% == 0  -10.8000  -16.1739  -5.4261
```

```
plot(Coto_a.glht)
```

95% family-wise confidence level



2.8.6 Ejercicio de simulación de una distribución F en una prueba ANOVA

Ahora realizaremos un ejercicio mediante simulación de variables aleatorias para observar la distribución del estadístico F.



Imagen de Wikipedia(<https://cronicaglobal.lespanol.com/uploads/s1/40/36/92/4/imagen-del-ordenador-mas-grande-del-mundo-capaz-de-imitar-la-comunicacion-del-cerebro-humano.jpeg>)

Generad una distribución normal con parámetros $\mu = 168$ y $\sigma^2 = 20$ y generad 4 muestras de tamaño $n = 20$. Una vez realizada esta tarea, calculad y guardad la media y la varianza. Usad la función `aov()`, comparando $MS_{between}$ y MS_{within} entre las 4 muestras que se han generado anteriormente.

Simulación de una distribución normal, con los parámetros indicados, y 4 muestras de tamaño $n=20$:

```
population <- rnorm(n = 1e6, mean = 168, sd=sqrt(20))
sample_size=20
num_samples=4
```

```

data<-data.frame()
stats<-
data.frame(sample=numeric(num_samples),mean=numeric(num_samples),var=numeric(num_samples))

for(i in 1:num_samples){
sample_name<-paste("sample",i,sep='_')
row_names<-rep(sample_name,sample_size)
sample_values<-sample(population, size = sample_size,replace = TRUE)

var_sample<-var(sample_values)
mean_sample<-mean(sample_values)

mean_var<-c(sample_name,mean_sample,var_sample)
stats[i,]<-mean_var

new_samples<-data.frame(sample=row_names,values=sample_values)
data<-rbind(data,new_samples)
}

stats

##      sample          mean          var
## 1 sample_1 168.372419454764 21.8931969536804
## 2 sample_2 168.251942220267 30.4199526988431
## 3 sample_3 168.447926191859 18.0744457705951
## 4 sample_4 168.646638192393 15.6617306393441

```

Ahora se realiza mediante la función ANOVA en R y se realiza el cálculo de $MS_{between}$ y de MS_{within} :

```

anova_out<-aov(values~sample,data)
anova_out

## Call:
## aov(formula = values ~ sample, data = data)
##
## Terms:
##              sample Residuals
## Sum of Squares    1.6455 1634.9372
## Deg. of Freedom         3         76
##
## Residual standard error: 4.638139
## Estimated effects may be unbalanced

summary(anova_out)

##              Df Sum Sq Mean Sq F value Pr(>F)
## sample         3    1.6    0.548   0.025  0.994
## Residuals     76 1634.9   21.512

```

```

#Finalmente obtenemos el valor F
F_fertilizer<-summary(anova_out)[[1]][1,4]
#F VALUE:
F_fertilizer
## [1] 0.02549646

```

Observamos que el test ANOVA no es significativo (p-valor > 0.05).

Comparamos los dos valores de $MS_{between}$ Y MS_{within} , se puede observar que el mejor estimador de la varianza de la distribución generada ($\sigma^2 = 20$) es $MS_{within}=21.512$

Ahora vamos a replica 1000 veces este experiment (muestreo de 4 muestras de tamaño n=20 y realizar el test ANOVA), posteriormente vamos a realizar una gráfica de la distribución de los valores F calculados en el ANOVA.

Vamos a ver qué ocurre si utilizamos la distribución de los valores F para estimar el p-valor del test ANOVA con los mismos grados de libertad (df),

```

population <- rnorm(n = 1e6, mean = 168, sd=sqrt(20))
sample_size=20
num_samples=4
anova_replicates<-1000
F_values<-numeric(anova_replicates)

for(j in 1:anova_replicates){
  data<-data.frame()
  for(i in 1:num_samples){
    sample_name<-paste("sample",i,sep='_')
    row_names<-rep(sample_name,sample_size)
    sample_values<-sample(population, size = sample_size,replace = TRUE)
    new_samples<-data.frame(sample=row_names,values=sample_values)
    data<-rbind(data,new_samples)
  }

  anova_out<-aov(values~sample,data)
  F_val<-summary(anova_out)[[1]][1,4]
  F_values[j]<-F_val
}

F_values_sorted<-sort(F_values)
F_critical<-F_values_sorted[length(F_values_sorted)*0.95]
F_critical

## [1] 2.909701

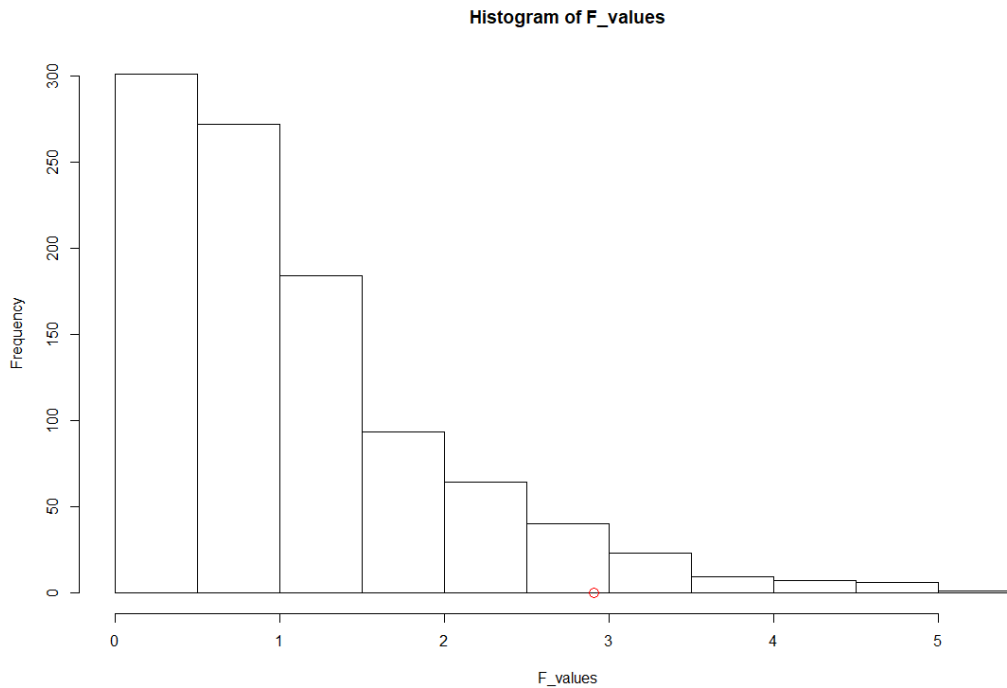
```

```
> qf(0.95, 3, 76)
```

```
[1] 2.724944
```

```
#distribución de los valores de F
```

```
hist(F_values);points(F_critical,0,col="red",cex=1.5)
```



Podemos observar que el punto crítico de F para 3 y 76 grados de libertad y un cuartil del 95% es 2.7249, mientras que el determinado con la distribución empírica del estadístico F es bastante parecido, utilizando la distribución teórica con los mismos grados de libertad y $\alpha = 0.05$, $F = 2.909701$

2.8.7. Un factor aleatorio, ejemplo del análisis del efecto lote en el rendimiento de colorante (batch).

Diseños con factores aleatorios (Uso de `library("lme4")` y de la función `lmer()`)

Existen muchos casos, donde los factores no son fijos, sino aleatorios, como por ejemplo los casos donde estudiamos lotes aleatorios, efecto batch, etc. Un buen lugar donde encontrar una buena introducción a este tema y diversos casos es <https://stat.ethz.ch/~meier/teaching/anova/random-and-mixed-effects-models.html#random-effects-models>

Ahora usamos otro punto de vista: consideramos situaciones donde los tratamientos son muestras aleatorias de una gran cantidad de tratamientos. Esto puede parecer bastante especial a primera vista, pero en realidad es muy natural en muchas situaciones. Piense, por ejemplo, en una muestra aleatoria de clases escolares que se extrajeron de todas las clases escolares de un país.

Otro ejemplo podría ser las máquinas que se muestrearon aleatoriamente de una gran cantidad de máquinas. Por lo general, nos interesa hacer una declaración acerca de algunas propiedades de toda la población y no de los individuos observados (aquí: clases escolares o máquinas).

Veamos algunos ejemplos típicos y cómo tratarlos.

El efecto batch o efecto lote aparece frecuentemente en biología y en las biociencias, en general, ya que se trabajan por lotes o partidas en lo que nos interesa es ver si existe una influencia de este factor pero en forma de si aporta varianza a la variable de respuesta.

Veamos los ejemplos que aparecen en http://www.stat.purdue.edu/~ovitek/STAT526-Spring11_files/2-mixed-part2.R

Una librería que permite el análisis de factores de tipo aleatorio de manera conveniente es `library("lme4")`, dentro de esta librería hay diferentes data-frames en R de ejemplo:

```
Data sets in package 'lme4':
```

| | |
|-------------|--|
| Arabidopsis | Arabidopsis clipping/fertilization data |
| Dyestuff | Yield of dyestuff by batch |
| Dyestuff2 | Yield of dyestuff by batch |
| InstEval | University Lecture/Instructor Evaluations by Students at ETH |
| Pastes | Paste strength by batch and cask |
| Penicillin | Variation in penicillin testing |
| VerbAgg | Verbal Aggression item responses |
| cake | Breakage Angle of Chocolate Cakes |


```

cbpp                Contagious bovine pleuropneumonia
grouseticks        Data on red grouse ticks from Elston et al.
2001
grouseticks_agg (grouseticks)
                   Data on red grouse ticks from Elston et al.
2001
sleepstudy         Reaction times in a sleep deprivation study

```

El data-frame de R Dyestuff proporciona el rendimiento del colorante (Naftaleno Negro 12B) a partir de 5 preparaciones diferentes de cada uno de 6 lotes (batches) diferentes de un producto intermedio (ácido H). Los datos de Dyestuff2 se generaron en la misma estructura pero con una gran variación residual en relación con la variación del factor lote.



Colorantes químicos (Imagen de https://sc02.alicdn.com/kf/HTB14lvEKFXXXbdXVXXq6xXFXXN/Amido-Black-10B.jpg_220x220.jpg)

Ejemplo sobre efecto batch en http://www.stat.purdue.edu/~ovitek/STAT526-Spring11_files/2-mixed-part2.R

Más información sobre factores aleatorios y este ejemplo en http://www.stat.purdue.edu/~ovitek/STAT526-Spring11_files/pdfs/2-mixed.pdf

Veamos cómo se realiza el análisis de un factor de tipo aleatorio (efecto batch) utilizando la librería `library("lme4")`:

```

#####
#           1-way random effects ANOVA (from lme4)
#####
# list the datasets available in lme4
library("lme4")

```

```

## Loading required package: Matrix

data(package = "lme4")

#The Dyestuff data frame provides the yield of dyestuff (Naphthalene
Black 12B) from 5 #different preparations from each of 6 different batches
of an intermediate product #(H-acid). The Dyestuff2 data were generated
data in the same structure but with a #large residual variance relative
to the batch variance.

#Data frames, each with 30 observations on the following 2 variables.

#Batch
#a factor indicating the batch of the intermediate product from which the
preparation                                #was                                created.

#Yield
#the yield of dyestuff from the preparation (grams of standard color).

#The Dyestuff data are described in Davies and Goldsmith (1972) as coming
from "an #investigation to find out how much the variation from batch to
batch in the quality #of an intermediate product (H-acid) contributes to
the variation in the yield of the #dyestuff (Naphthalene Black 12B) made
from it. In the experiment six samples of the #intermediate, representing
different batches of works manufacture, were obtained, and #five
preparations of the dyestuff were made in the laboratory from each
sample. The #equivalent yield of each preparation as grams of standard
colour                                was                                determined                                by                                #dye-trial."

#The Dyestuff2 data are described in Box and Tiao (1973) as illustrating
" the case #where between-batches mean square is less than the within-
batches mean square. These #data had to be constructed for although
examples of this sort undoubtedly occur in #practice, they seem to be
rarely                                published."

#O.L. Davies and P.L. Goldsmith (eds), Statistical Methods in Research
and Production, #4th ed., Oliver and Boyd, (1972), section 6.4

#G.E.P. Box and G.C. Tiao, Bayesian Inference in Statistical Analysis,
Addison-Wesley,                                #(1973),                                section                                5.1.2

#
# a balanced one-way classification of Dyestuff
# from samples produced from six colorant Batches
#?Dyestuff
str(Dyestuff)

## 'data.frame': 30 obs. of 2 variables:
## $ Batch: Factor w/ 6 levels "A","B","C","D",...: 1 1 1 1 1 2 2 2 2 2

```

```

...
## $ Yield: num 1545 1440 1440 1520 1580 ...

summary(Dyestuff)

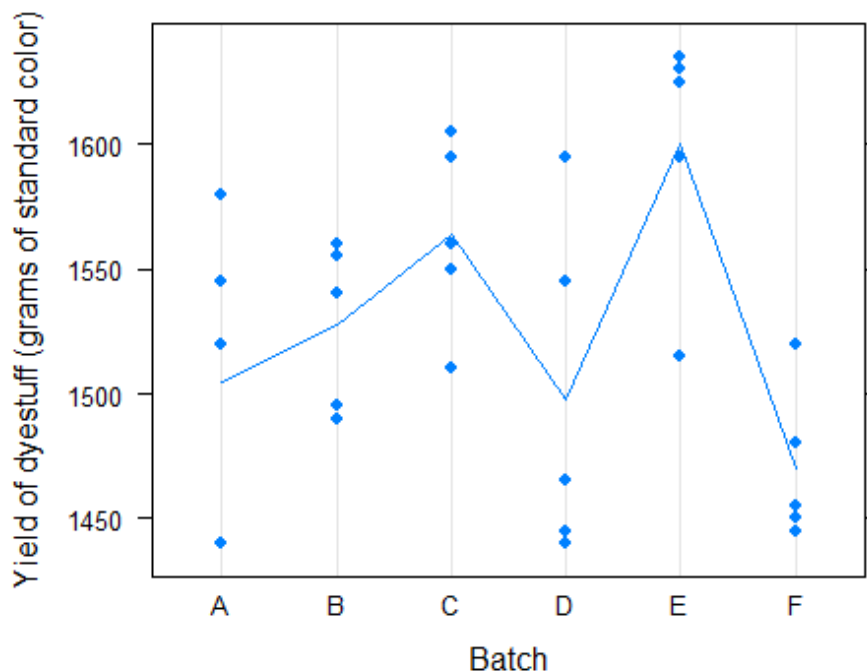
## Batch      Yield
## A:5   Min.    :1440
## B:5   1st Qu.:1469
## C:5   Median  :1530
## D:5   Mean    :1528
## E:5   3rd Qu.:1575
## F:5   Max.    :1635

xtabs(~Batch, Dyestuff)

## Batch
## A B C D E F
## 5 5 5 5 5 5

# data visualization: original order
library(lattice)
dotplot(Yield ~ Batch, Dyestuff,
        type = c("p", "a"), xlab = "Batch",
        ylab = "Yield of dyestuff (grams of standard color)")

```

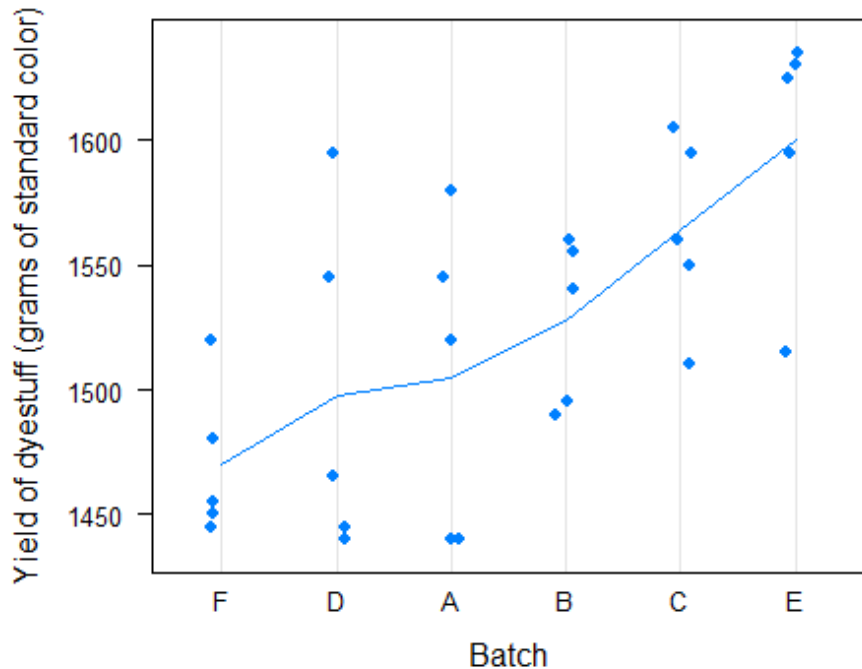


```

# data visualization: order Batch by Yield
# can help roughly assess Normality of the random effect
dotplot(Yield ~ reorder(Batch, Yield), Dyestuff,

```

```
type = c("p", "a"), jitter.x = TRUE, xlab = "Batch",
ylab = "Yield of dyestuff (grams of standard color)")
```



```
# Batch is a classical example of a random effect
# Fit 1-way random effects linear model
fit1 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff)
summary(fit1)

## Linear mixed model fit by REML ['lmerMod']
## Formula: Yield ~ 1 + (1 | Batch)
## Data: Dyestuff
##
## REML criterion at convergence: 319.7
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.4117 -0.7634  0.1418  0.7792  1.8296
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
##   Batch    (Intercept) 1764     42.00
##   Residual                2451     49.51
## Number of obs: 30, groups: Batch, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 1527.50     19.38    78.8
```

```

# access model-based quantities
class(fit1)

## [1] "lmerMod"
## attr("package")
## [1] "lme4"

fixef(fit1)

## (Intercept)
##      1527.5

ranef(fit1, drop = TRUE) # note no constraint

## $Batch
##      A      B      C      D      E
F
## -17.6068514  0.3912634  28.5622256 -23.0845385  56.7331877 -
44.9952868
## attr("postVar")
## [1] 383.6337 383.6337 383.6337 383.6337 383.6337 383.6337
##
## with conditional variances for "Batch"

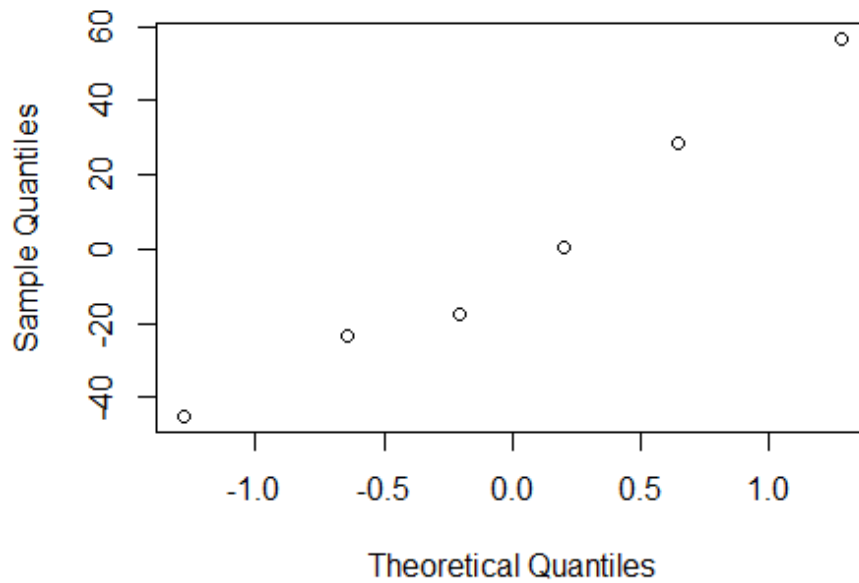
# 'drop' changes the output format from matrix to vector
fitted(fit1)

##      1      2      3      4      5      6      7
8
## 1509.893 1509.893 1509.893 1509.893 1509.893 1527.891 1527.891
1527.891
##      9     10     11     12     13     14     15
16
## 1527.891 1527.891 1556.062 1556.062 1556.062 1556.062 1556.062
1504.415
##     17     18     19     20     21     22     23
24
## 1504.415 1504.415 1504.415 1504.415 1584.233 1584.233 1584.233
1584.233
##     25     26     27     28     29     30
## 1584.233 1482.505 1482.505 1482.505 1482.505 1482.505

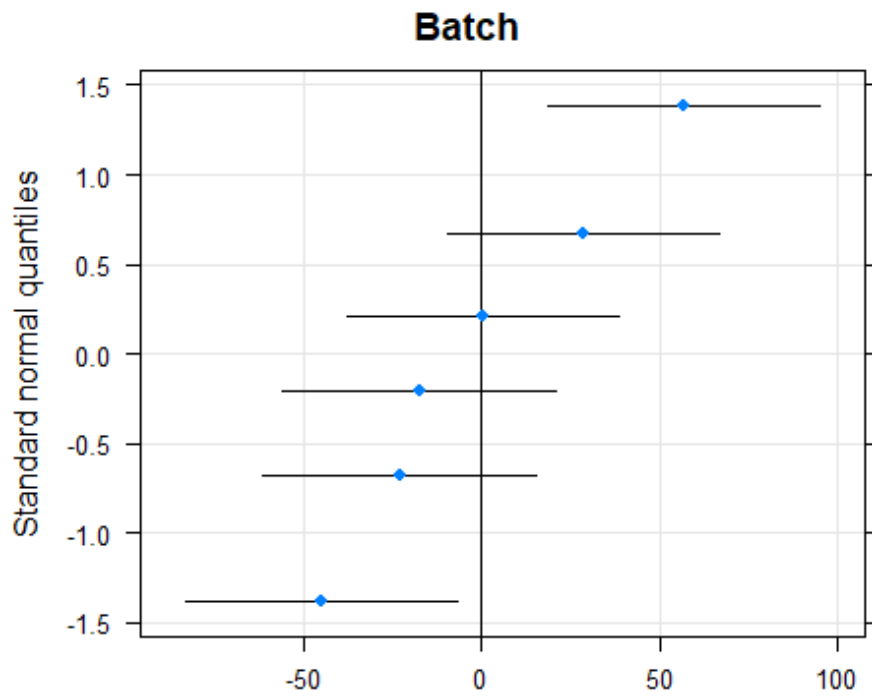
# check the assumption of Normality of random effects
# qqplot of the random effects with their variances
qqnorm(unlist(ranef(fit1)))

```

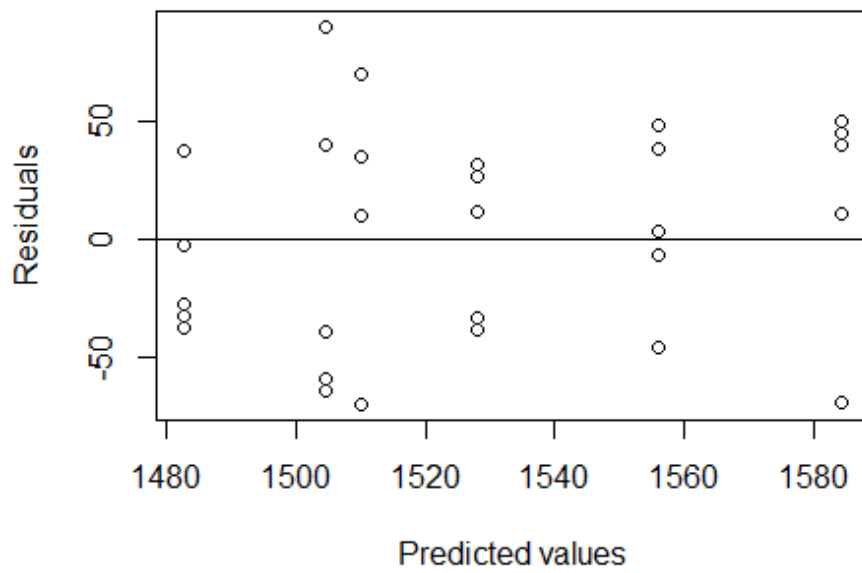
Normal Q-Q Plot



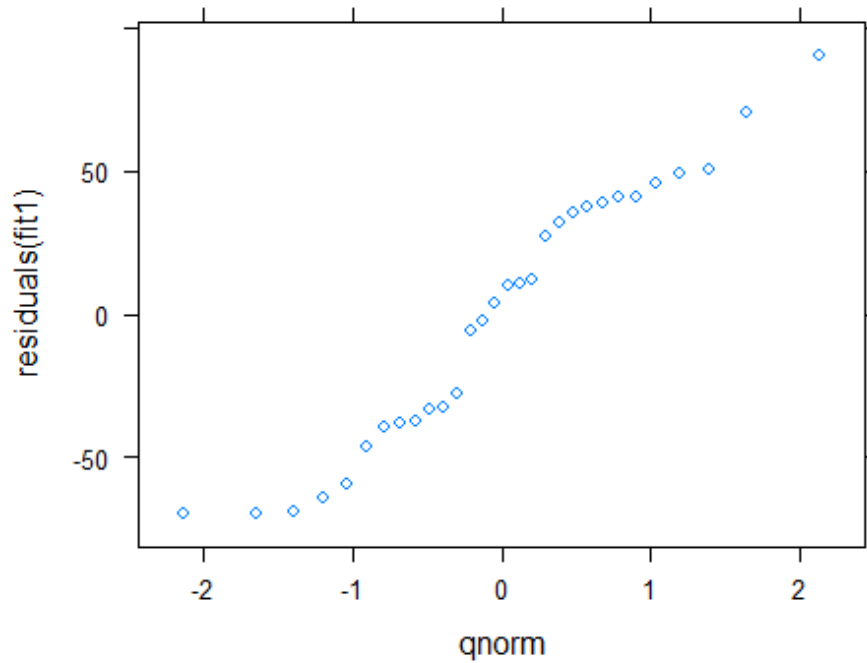
```
qqmath(ranef(fit1, postVar = TRUE), strip = FALSE)$Batch # fancy
## Warning in ranef.merMod(fit1, postVar = TRUE): 'postVar' is
deprecated:
## please use 'condVar' instead
```



```
# check the assumption of Normality of residuals  
plot(fitted(fit1), residuals(fit1), xlab="Predicted values",  
ylab="Residuals")  
abline(h=0)
```



```
qqmath(residuals(fit1))
```



```
# compare estimates of random effects with
# estimates of fixed effects
```



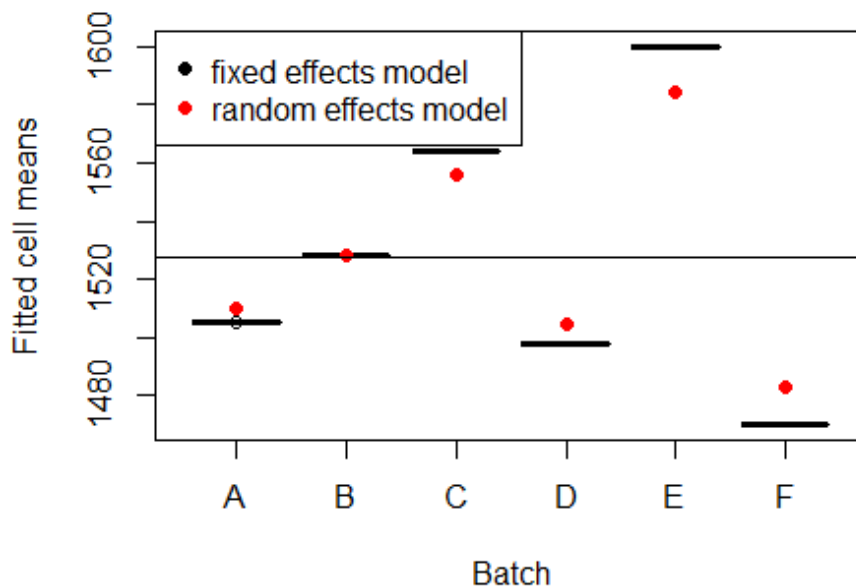
```

fit1.fixed <- lm(Yield ~ Batch, Dyestuff)
fitted(fit1.fixed)

##    1    2    3    4    5    6    7    8    9   10   11   12   13   14
15
## 1505 1505 1505 1505 1505 1528 1528 1528 1528 1528 1564 1564 1564 1564
1564
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29
30
## 1498 1498 1498 1498 1498 1600 1600 1600 1600 1600 1470 1470 1470 1470
1470

# estimates of random effects are closer to the mean
plot(Dyestuff$Batch, fitted(fit1.fixed),
     xlab = "Batch", ylab = "Fitted cell means")
points(Dyestuff$Batch, fitted(fit1), pch=16, col="red")
abline(h=mean(Dyestuff$Yield))
legend("topleft", pch=16, col=c("black", "red"),
      c("fixed effects model", "random effects model"))

```



2.8.8. Ejemplo del uso de ANOVA no paramétrico

Ahora vamos a utilizar otro ejemplo donde el ANOVA ha resultado significativo y existen diferencias estadísticamente significativas entre los niveles del factor, pero no se cumplen los supuestos del ANOVA. El ejemplo se ha obtenido en `library(nparcomp)`.

Hacemos el análisis ANOVA y detectamos heterocedasticitat.

```
library(nparcomp)
#Nonparametric Multiple Comparisons for relative effects
data("reaction")
reaction.aov <- lm(Time ~ Group, data = reaction)
summary(reaction.aov)

##
## Call:
## lm(formula = Time ~ Group, data = reaction)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.142 -1.146 -0.046  1.308  5.054
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7480     0.6019   4.566 5.10e-05 ***
## Group         1.5980     0.3217   4.967 1.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.275 on 38 degrees of freedom
## Multiple R-squared:  0.3937, Adjusted R-squared:  0.3777
## F-statistic: 24.67 on 1 and 38 DF,  p-value: 1.47e-05
```

Violación de los supuestos paramétricos del modelo

Si realizamos un diagnóstico del modelo lineal y de los supuestos del ANOVA, detectamos heterocedasticidad (diferencias de varianzas significativas entre los niveles del factor Group):

```
shapiro.test(residuals(reaction.aov))

##
## Shapiro-Wilk normality test
##
## data:  residuals(reaction.aov)
## W = 0.98166, p-value = 0.751
```

```

bartlett.test(Time ~ Group, data = reaction)

##
## Bartlett test of homogeneity of variances
##
## data: Time by Group
## Bartlett's K-squared = 23.02, df = 3, p-value = 4e-05

library(car)
leveneTest(reaction$Time, reaction$Group)

## Warning in leveneTest.default(reaction$Time, reaction$Group):
## reaction$Group coerced to factor.

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 3  2.7079 0.05954 .
##      36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

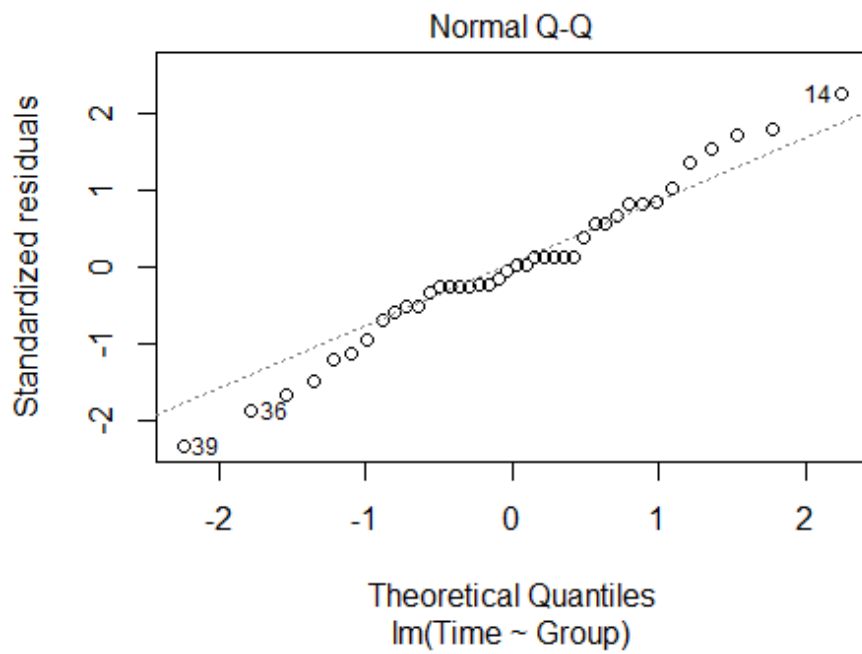
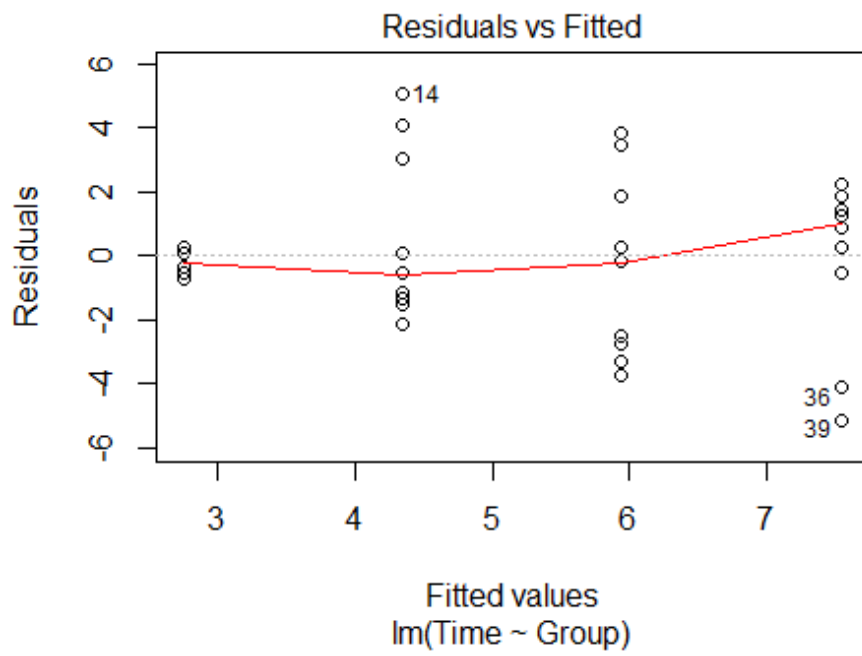
fligner.test(Time ~ Group, data = reaction) # Fligner-Killeen Test of
Homogeneity of Variances

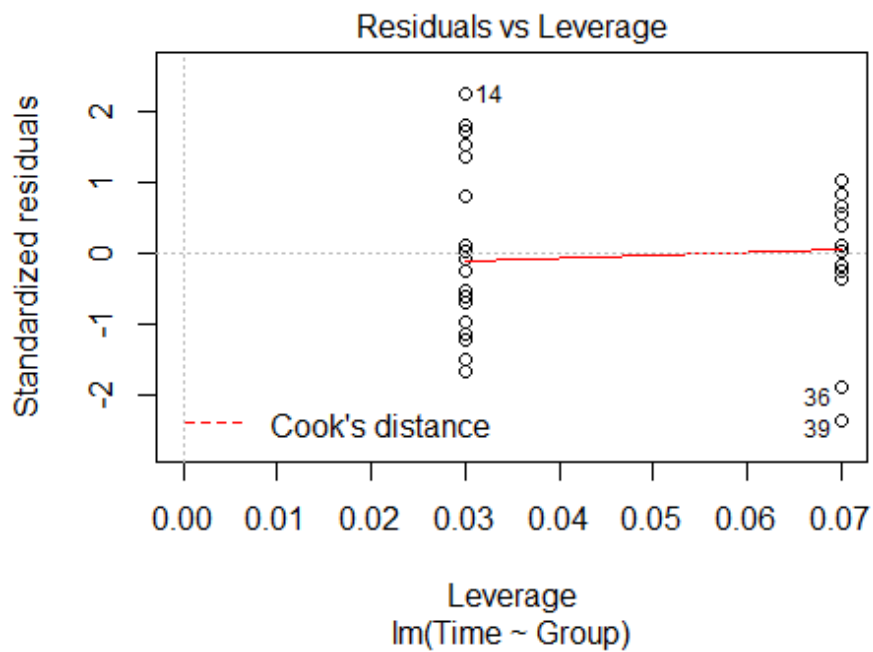
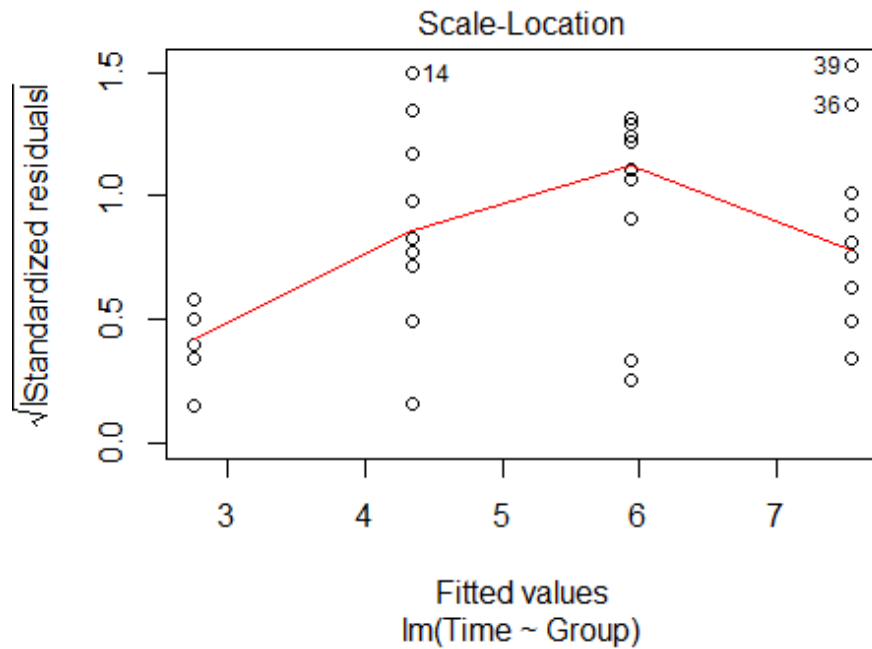
##
## Fligner-Killeen test of homogeneity of variances
##
## data: Time by Group
## Fligner-Killeen:med chi-squared = 9.0543, df = 3, p-value =
## 0.02858

# Análisis grafico del modelo anova, donde no se aprecia bien la
heterocedastiidad

plot(reaction.aov)

```





Una posible solución a esta heterocedasticidad es hacer correcciones de Box-Cox u otras técnicas de las vistas anteriormente y re-diagnosticar el resultado del modelo, para ver si se ha corregido o bien si deseamos hacer MCP, podemos utilizar unas MCP donde se tenga en cuenta el problema de la heterocedasticidad.

Uso de Nonparametric multiple test: library(nparcomp)

El paquete nparcomp proporciona MCP no paramétricas. (Nota: Este paquete ha sido retirada, pero todavía esta disponible en los archivos CRAN.)

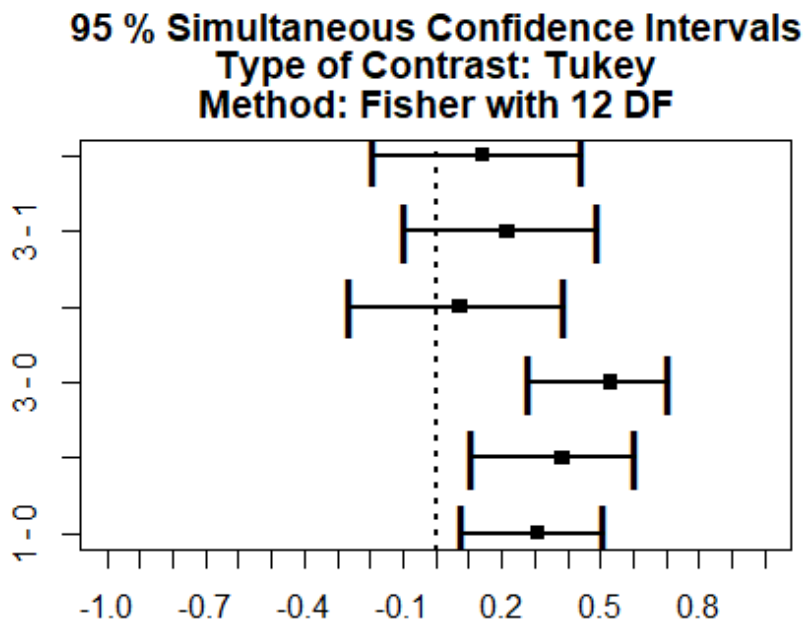
```
library(nparcomp)
#Nonparametric Multiple Comparisons for relative effects
data("reaction")
aaa<-mctp(Time ~ Group, data = reaction, type = "Tukey",
          conf.level = 0.95, info = FALSE)
summary(aaa)

##
## #-----Nonparametric Multiple Comparisons for
relative effects-----#
##
## - Alternative Hypothesis: True differences of relative effects
are less or equal than 0
## - Estimation Method: Global Pseudo ranks
## - Type of Contrast : Tukey
## - Confidence Level: 95 %
## - Method = Fisher with 12 DF
##
## #-----#
-----#
##
## #----Data Info-----#
-----#
##   Sample Size  Effect      Lower      Upper
## 0      0      10 0.19375 0.1439117 0.2556929
## 1      1      10 0.50500 0.4026620 0.6069208
## 2      2      10 0.57875 0.4608841 0.6882761
## 3      3      10 0.72250 0.6121914 0.8111133
##
## #----Contrast-----#
-----#
##           0  1  2  3
## 1 - 0 -1  1  0  0
## 2 - 0 -1  0  1  0
## 3 - 0 -1  0  0  1
## 2 - 1  0 -1  1  0
## 3 - 1  0 -1  0  1
## 3 - 2  0  0 -1  1
##
## #----Analysis-----#
-----#
##           Estimator  Lower Upper  Statistic      p.Value
## 1 - 0      0.311  0.079 0.512      3.866 0.0094618877
## 2 - 0      0.385  0.109 0.606      4.000 0.0075554996
## 3 - 0      0.529  0.286 0.708      5.844 0.0002768359
```

```
## 2 - 1      0.074 -0.259 0.391      0.636 0.9102751489
## 3 - 1      0.217 -0.095 0.491      2.040 0.2123179739
## 3 - 2      0.144 -0.190 0.447      1.256 0.5841393654
##
## #----Overall-----#
-----#
##   Quantile      p.Value
## 1 2.920003 0.0002768359
##
## #-----#
-----#
```

Se pueden calcular también unos intervalos de confianza simultáneos con el resultado obtenido con la función `mctp` y la prueba no paramétrica MCP tipo Tukey:

```
plot(aaa) #intervalos de confianza simultaneos
```



Nonparametric multiple test: función `kruskalmc()`

Otra posibilidad es hacer un MCP no paramétrico, utilizando el test de Kruskal-Wallis a posteriori, para ello deben instalarse las siguientes librerías y la función `kruskalmc()`:

```
library(rgdal)
## Loading required package: sp
```

```

## rgdal: version: 1.3-6, (SVN revision 773)
## Geospatial Data Abstraction Library extensions to R successfully
loaded
## Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
## Path to GDAL shared files: C:/Users/bootstrap3/Documents/R/win-
library/3.5/rgdal/gdal
## GDAL binary built with GEOS: TRUE
## Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
## Path to PROJ.4 shared files: C:/Users/bootstrap3/Documents/R/win-
library/3.5/rgdal/proj
## Linking to sp version: 1.3-1

library(rgeos)

## rgeos version: 0.4-2, (SVN revision 581)
## GEOS runtime version: 3.6.1-CAPI-1.10.1
## Linking to sp version: 1.3-1
## Polygon checking: TRUE

library(splancs)

##
## Spatial Point Pattern Analysis Code in S-Plus
##
## Version 2 - Spatial and Space-Time analysis

library(maptools)

## Checking rgeos availability: TRUE

library(pgirmess)
kruskalmc(Time ~ Group, data = reaction ) #alternativa al test de
Kruskaall-Wallis multiple

## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
##   obs.dif critical.dif difference
## 0-1  12.45    13.79315      FALSE
## 0-2  15.40    13.79315       TRUE
## 0-3  21.15    13.79315       TRUE
## 1-2   2.95    13.79315      FALSE
## 1-3   8.70    13.79315      FALSE
## 2-3   5.75    13.79315      FALSE

```


Nonparametric multiple test: función `pairwise.wilcox.test ()`:

Otra posibilidad es utilizar el test de Mann-Whitney/Wilcoxon de rangos por parejas a todas las parejas durante la MCP, con ajuste del p-valor, utilizando el comando `pairwise.wilcox.test`:

```
pairwise.wilcox.test(reaction$Time, reaction$Group, p.adj="holm")
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
##
## Pairwise comparisons using Wilcoxon rank sum test
##
## data: reaction$Time and reaction$Group
##
##    0      1      2
## 1 0.0165 -      -
## 2 0.0165 0.4715 -
## 3 0.0037 0.1606 0.3948
##
## P value adjustment method: holm
```

Un paquete para realizar pruebas MCP no paramétricas es `nparcomp`, que ha sido desarrollado en 2014 por Konietschke et al. (2014).

Nonparametric multiple test: Método de GAO

Una posibilidad muy reciente es utilizar el test ANOVA propuesto por Gao et al. (2008). Este autor ha desarrollado un ANOVA robusto a las desviaciones y también un

test MCP robusto a estas desviaciones. El inflamiento del error de tipo I se controla mediante el ajuste de Hochberg. Estas pruebas pueden encontrarse en `library(nparcomp)` y utilizarse mediante la función `gao-cs()` o `gao()`, como en el siguiente ejemplo:

```
gao_cs(Time ~ Group, data = reaction, alpha = 0.05 )

##
## #----Gao et al's (2008) modification of Campbell and Skillings (1985)
## (CS) stepwise multiple comparison procedure
## #---- This function uses joint ranks of the data. Attention: In the
## CS algorithm, the samples are jointly reranked!
## #----Reference: Gao, X. et al. (2008). Nonparametric Multiple
## Comparison Procedures for Unbalanced One-Way Factorial Designs. JSPI 138,
## 2574 - 2591.

## $Info
##   Order Sample Size  Effect  Variance
## 1     1           0    10 0.19375 0.01643229
## 2     2           1    10 0.50500 0.05757639
## 3     3           2    10 0.57875 0.07524479
## 4     4           3    10 0.72250 0.05256250
##
## $Single.Analysis
##   Comp Effect Statistic      DF  P.RAW p.BONF p.HOLM
## 1  3-0 0.5288      6.3656 14.1262 0.0000 0.0001 0.0001
## 2  2-0 0.3850      4.0210 12.7520 0.0015 0.0090 0.0075
## 3  3-1 0.2175      2.0725 17.9628 0.0529 0.3173 0.1587
## 4  1-0 0.3113      3.6180 13.7503 0.0029 0.0172 0.0115
## 5  2-1 0.0737      0.6399 17.6870 0.5304 1.0000 0.5304
## 6  3-2 0.1438      1.2715 17.4504 0.2202 1.0000 0.4404
##
## $CS.Analysis
##   Comp Effect Statistic      DF Quantiles Adj.P Alpha Rejected Layer
## 1  3-0 0.5288      9.0024 14.1262      4.1059 1e-04 0.0500      TRUE      1
## 2  2-0 0.4200      5.9397 14.1745      3.6962 0.0023 0.0500      TRUE      2
## 3  3-1 0.2650      3.1379 17.9953      3.6094 0.0949 0.0500      FALSE     2
## 4  1-0 0.3800      5.3725 16.8977      3.4699 0.0014 0.0253      TRUE      3
## 5  2-1 0.1000      1.0594 17.8029      3.453  0.4636 0.0253      FALSE     3
## 6  3-2 0.1750      1.9174 17.7942      3.453  0.1921 0.0253      FALSE     3

gao(Time ~ Group, data = reaction, alpha = 0.05, control = NULL)

##
## #----Xin Gao et al's (2008) Non-Parametric Multiple Test Procedure
## #----Type of Adjustment: Hochberg
## #----Level of significance = 0.05
## #----The procedure compares if the distribution functions F() are
## equal. The FWER is strongly controlled
## #---- This function uses pseudo ranks of the data!
## #----Reference: Gao, X. et al. (2008). Nonparametric Multiple
```

Comparison Procedures for Unbalanced One-Way Factorial Designs. JSPI 138, 2574 - 2591.

```
## $Info
##   Sample Size Effect Variance
## 1      0     10 0.1938  0.0164
## 2      1     10 0.5050  0.0576
## 3      2     10 0.5788  0.0752
## 4      3     10 0.7225  0.0526
##
## $Analysis
##   Comparison Estimator      df Statistic  P.Raw P.Hochberg Rejected
P.Bonf
## 1  F(1)-F(0)      0.3113 13.7503    3.6180 0.0029    0.0029    TRUE
0.0086
## 2  F(2)-F(0)      0.3850 12.7520    4.0210 0.0015    0.0029    TRUE
0.0045
## 3  F(3)-F(0)      0.5288 14.1262    6.3656 0.0000    0.0001    TRUE
0.0001
##   P.Holm
## 1  3e-03
## 2  3e-03
## 3  1e-04
```

% #tots amb tots % #sembla que sols per quan hi ha un grup control

Veamos otro ejemplo de análisis de un ANOVA con violaciones de los supuestos, utilizando el ANOVA de Gao y sus MCP:

```
data(liver)
gao(weight ~dosage, data=liver, alpha=0.05)

##
## #----Xin Gao et al's (2008) Non-Parametric Multiple Test Procedure
## #----Type of Adjustment: Hochberg
## #----Level of significance = 0.05
## #----The procedure compares if the distribution functions F() are
equal. The FWER is strongly controlled
## #---- This function uses pseudo ranks of the data!
## #----Reference: Gao, X. et al. (2008). Nonparametric Multiple
Comparison Procedures for Unbalanced One-Way Factorial Designs. JSPI 138,
2574 - 2591.

## $Info
##   Sample Size Effect Variance
## 1      1      8 0.2739  0.0406
## 2      2      7 0.3168  0.0653
## 3      3      8 0.3618  0.0279
## 4      4      7 0.6939  0.0273
## 5      5      8 0.8536  0.0121
##
```

```

## $Analysis
## Comparison Estimator      df Statistic  P.Raw P.Hochberg Rejected
P.Bonf
## 1 F(2)-F(1)    0.0430 11.4072    0.3580 0.7269    0.7269    FALSE
1.0000
## 2 F(3)-F(1)    0.0879 13.5348    0.9508 0.3584    0.7167    FALSE
1.0000
## 3 F(4)-F(1)    0.4200 12.9622    4.4348 0.0007    0.0020    TRUE
0.0027
## 4 F(5)-F(1)    0.5797 10.8470    7.1413 0.0000    0.0001    TRUE
0.0001
## P.Holm
## 1 0.7269
## 2 0.7167
## 3 0.0020
## 4 0.0001

gao(weight ~dosage, data=liver,alpha=0.05,control="3")

##
## #----Xin Gao et al's (2008) Non-Parametric Multiple Test Procedure
## #----Type of Adjustment: Hochberg
## #----Level of significance = 0.05
## #----The procedure compares if the distribution functions F() are
equal. The FWER is strongly controlled
## #---- This function uses pseudo ranks of the data!
## #----Reference: Gao, X. et al. (2008). Nonparametric Multiple
Comparison Procedures for Unbalanced One-Way Factorial Designs. JSPI 138,
2574 - 2591.

```

```

## $Info
## Sample Size Effect Variance
## 1 1 8 0.2739 0.0406
## 2 2 7 0.3168 0.0653
## 3 3 8 0.3618 0.0279
## 4 4 7 0.6939 0.0273
## 5 5 8 0.8536 0.0121
##
## $Analysis
## Comparison Estimator      df Statistic  P.Raw P.Hochberg Rejected
P.Bonf
## 1 F(1)-F(3)   -0.0879 13.5348   -0.9508 0.3584    0.6993    FALSE
1.0000
## 2 F(2)-F(3)   -0.0450 10.1109   -0.3975 0.6993    0.6993    FALSE
1.0000
## 3 F(4)-F(3)    0.3320 12.7730    3.8645 0.0020    0.0060    TRUE
0.0081
## 4 F(5)-F(3)    0.4917 12.1267    6.9522 0.0000    0.0001    TRUE
0.0001
## P.Holm
## 1 0.7167

```

2 0.7167
3 0.0060
4 0.0001



ANOVA n-WAY

2.9. ANOVA de 2 factores: introducción y asunciones

El análisis de la varianza de 2 factores o “Two way” es una extensión del análisis unidireccional (“One way”) de la varianza. Hay dos variables independientes (de ahí el nombre de dos vías).

Anteriormente se estudió la relación entre una variable dependiente continua y una variable independiente categórica. Sin embargo, cuando tenemos más de una variable categórica independiente, tenemos que usar ANOVA de N vías (N-way). Veremos el caso cuando tengamos dos variables categóricas independientes, por lo que usaremos un ANOVA de dos vías (2 factores o “two-way”).

Las principales asunciones para poder realizar el ANOVA son:

- Las poblaciones de las cuales se obtuvieron las muestras deben estar distribuidas normalmente o aproximadamente normalmente.
- Las muestras deben ser independientes.
- Las varianzas de las poblaciones deben ser iguales.
- Los grupos deben tener preferiblemente el mismo tamaño de muestra.

Hay tres conjuntos de hipótesis con el ANOVA de dos vías. Las hipótesis nulas para cada uno de los conjuntos son:

- Las medias poblacionales del primer factor son iguales. Esto es como el ANOVA de una vía para el factor fila.
- Las medias poblacionales del segundo factor son iguales. Esto es como el ANOVA de una vía para el factor de columna.
- No hay interacción entre los dos factores. Esto es similar a realizar una prueba de independencia con tablas de contingencia.

Las dos variables independientes en un ANOVA de dos vías se llaman factores. La idea es que hay dos variables, factores, que afectan a la variable dependiente. Cada factor tendrá dos o más niveles dentro de él, y los grados de libertad para cada factor son uno menos que el número de niveles (a, b, c, etc).

2.9.1. Presión ecológica sobre *Patella vulgata* (lapa de roca)

A continuación, se presentan algunos de los ejemplos basados en los que se explican en las clases de estadística y que se basan en los creados por Hafid Laayouni y colaboradores de la UPF (Barcelona).

La presión evolutiva se define como cualquier causa que reduzca el éxito reproductivo de una población en una proporción significativa ejerce una potencial presión evolutiva o presión selectiva, esto puede suceder cuando muchos individuos colonizan el mismo espacio. Veamos un ejemplo.

Un investigador quería estudiar los efectos de la densidad de adultos (presión ecológica) y la estación en la producción en masa de huevos de lapas intermareales (*Patella vulgata*). Se manipuló la densidad de las lapas (DENSITY) de los adultos en los recintos (8, 15, 30 y 45 individuos por recinto) durante dos temporadas (SEASON : invierno-primavera y verano-otoño) . Se utilizaron tres recintos replicados por combinación densidad / estación.



Imagen de lapas de las rocas (Wikipedia, https://en.wikipedia.org/wiki/Patella_vulgata)

Importar el conjunto de datos para su análisis (limpet.csv).

Observad los datos: ¿Cuántos factores tenemos? ¿Cuántas categorías (niveles) hay en cada factor? ¿Cuál es el número de réplicas por condición experimental?


```

limpet <- read.csv("limpet.csv")
head(limpet)

##      DENSITY SEASON  EGGS
## 1         8  spring 2.875
## 2         8  spring 2.625
## 3         8  spring 1.750
## 4         8  summer 2.125
## 5         8  summer 1.500
## 6         8  summer 1.875

# Two factors: SEASON and DENSITY
# 2 categories in SEASON
unique(limpet$SEASON)

## [1] spring summer
## Levels: spring summer

# 4 categories in DENSITY
unique(limpet$DENSITY)

## [1] 8 15 30 45

# if we combine both, we have 8 groups with 3 replicates
table(limpet$DENSITY,limpet$SEASON)

##
##      spring summer
## 8         3       3
## 15        3       3
## 30        3       3
## 45        3       3

```

IMPORTANTE: La variable DENSITY contiene representaciones numéricas de las densidades de la lapa del adulto (8, 15, 30 y 45), y R considerará que se trata de un vector de tipo entero en lugar de un factor categórico. Tenemos que cambiar la clase de la variable a un factor utilizando la función **factor()**.

```

class(limpet$DENSITY)

## [1] "integer"

limpet$DENSITY <- factor(limpet$DENSITY)
class(limpet$DENSITY)

## [1] "factor"

levels(limpet$DENSITY) # Levels (categories) of the factor

## [1] "8" "15" "30" "45"

```

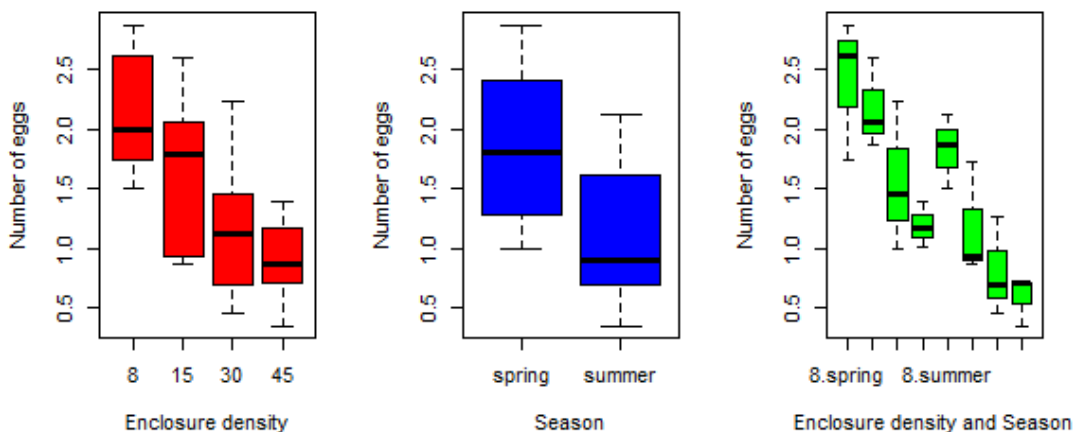
En este experimento, tanto DENSITY como SEASON se consideraron factores fijos y, por lo tanto, los datos representan un diseño ANOVA de 2 factores fijos con interacción.

Vamos a ver la dispersión de la variable de interés (número de huevos) dependiendo de cada factor por separado y conjuntamente a fin de comprobar la validez de la aplicación del ANOVA y para explorar los resultados del experimento:

```
op <- par(mfrow=c(1,3)) # modify plotting options 1row,3columns
boxplot(EGGS ~ DENSITY, data = limpet,
        xlab = "Enclosure density", ylab="Number of eggs", col="red")

boxplot(EGGS ~ SEASON, data = limpet,
        xlab = "Season", ylab="Number of eggs", col="blue")

boxplot(EGGS ~ DENSITY + SEASON + DENSITY:SEASON, data = limpet,
        xlab = "Enclosure density and Season", ylab="Number of eggs",
        col="green")
```



```
par(op) # go back to default plotting options
```

En principio los gráficos no indican evidencia de no normalidad (los diagramas de caja no son muy asimétricos).

Vamos ahora a utilizar la librería **ggplot2** de R, a fin de mejorar la representación gráfica exploratoria anterior.

```
library(ggplot2) # Load Library
# data to plot; axis; color by factor season
gb <- ggplot(limpet, aes(x = DENSITY, y = EGGS, colour = SEASON))

# add x and y Labs
gb <- gb + xlab("Enclosure density") + ylab("Number of eggs")
```

```

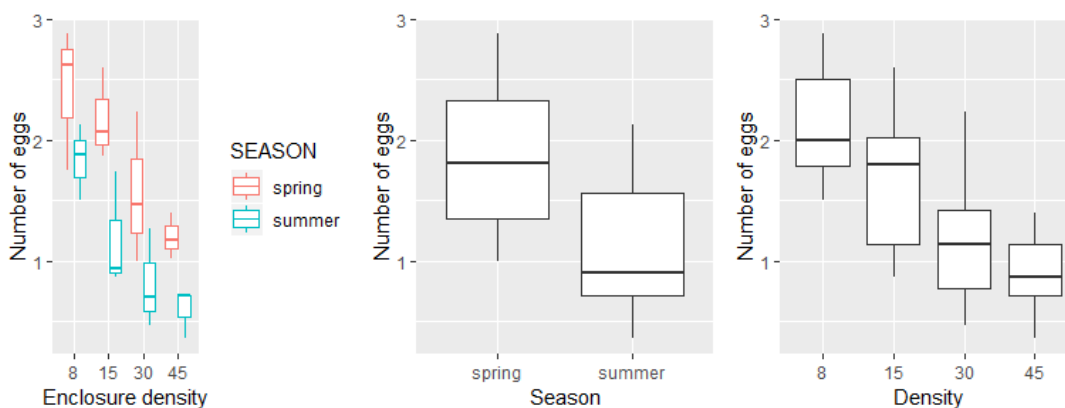
# add type of plot
gb <- gb + geom_boxplot()

# plot
plot(gb)

# eggs by season
gb <- ggplot(limpet, aes(x = SEASON, y = EGGS))
gb <- gb + xlab("Season") + ylab("Number of eggs")
gb <- gb + geom_boxplot()
plot(gb)

# eggs by density
gb <- ggplot(limpet, aes(x = DENSITY, y = EGGS))
gb <- gb + xlab("Density") + ylab("Number of eggs")
gb <- gb + geom_boxplot()
plot(gb)

```



Para realizar el análisis **two-way ANOVA**, tenemos que escribir la variable de interés en función de los dos factores y su interacción, obteniendo la tabla ANOVA con el SS, MS, p-value de manera correcta utilizando la función **aov()**:

```

limpet.aov <- aov(EGGS ~ DENSITY + SEASON + DENSITY:SEASON, data =
limpet)
limpet.aov

## Call:
## aov(formula = EGGS ~ DENSITY + SEASON + DENSITY:SEASON, data =
limpet)
##
## Terms:
##          DENSITY SEASON DENSITY:SEASON Residuals
## Sum of Squares   5.284  3.250           0.165    2.915
## Deg. of Freedom    3      1             3       16
##

```

```

## Residual standard error: 0.4268
## Estimated effects may be unbalanced

## another way of writing the formula is:
limpet.aov <- aov(EGGS ~ DENSITY * SEASON, data = limpet)

# obtain the ANOVA table with the SS, MS, p-value and so on
summary(limpet.aov)

##              Df Sum Sq Mean Sq F value Pr(>F)
## DENSITY      3   5.28   1.76   9.67 0.00070 ***
## SEASON       1   3.25   3.25  17.84 0.00065 ***
## DENSITY:SEASON 3   0.16   0.05   0.30 0.82395
## Residuals   16   2.91   0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

La conclusión es que tanto la densidad de lapas (DENSITY) como la estación (SEASON) afectan el número de huevos, pero no hay interacción entre los dos factores. Como hemos visto que las medias de cada categoría de densidad y estación son diferentes, aplicaremos una prueba t “pairwise” a posteriori del **ANOVA** en cada factor:

```

# What significant differences are present amongst the densities?
pairwise.t.test(limpet$EGGS, limpet$DENSITY, p.adj = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  limpet$EGGS and limpet$DENSITY
##
##      8      15      30
## 15 1.000 -      -
## 30 0.055 0.886 -
## 45 0.007 0.156 1.000
##
## P value adjustment method: bonferroni

# What significant differences are present amongst the seasons?
pairwise.t.test(limpet$EGGS, limpet$SEASON, p.adj = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  limpet$EGGS and limpet$SEASON
##
##      spring
## summer 0.008
##
## P value adjustment method: bonferroni

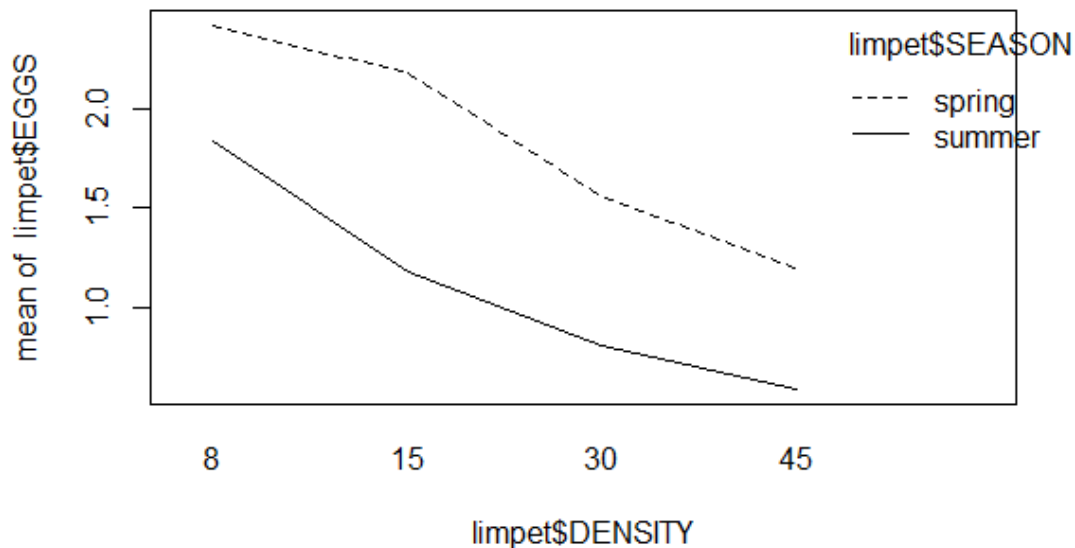
```

Una vez observamos los resultados con `pairwise.t.test()` ¿Qué diferencias significativas están presentes entre las estaciones o entre densidades?

Solo encontramos diferencias entre la media de la densidad más alta (45) y la más baja (8), y la diferencia en el número de huevos en verano y primavera es significativa.

Para observar el efecto combinado de ambos factores, dibujamos la gráfica de interacción mediante la función:

```
interaction.plot(limpet$DENSITY, limpet$SEASON, limpet$EGGS)
```



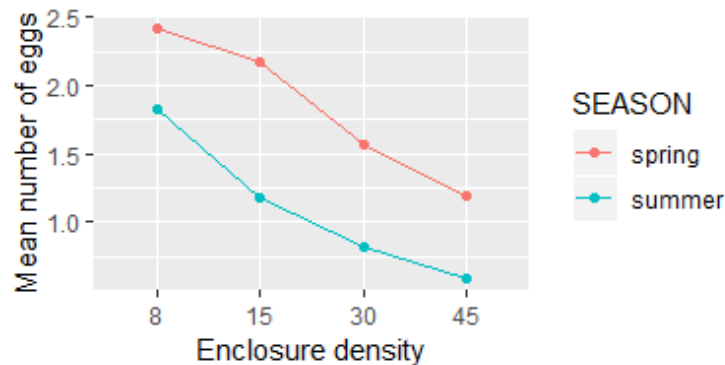
Podemos hacer lo mismo, pero mejorado utilizando la librería **ggplot2**:

```
# data to plot; axis; color and group by factor season
gi <- ggplot(limpet, aes(x = DENSITY, y = EGGS, colour = SEASON,
group=SEASON))

# add x and y labs
gi <- gi + xlab("Enclosure density") + ylab("Mean number of eggs")

# add the summary statistic to plot: mean. Draw a point and join it by a
line
gi <- gi + stat_summary(fun.y=mean, geom="point")+
  stat_summary(fun.y=mean, geom="line")

# plot
plot(gi)
```



Aquí podemos ver que no hay interacción entre los factores, solo un efecto aditivo de la densidad y la estación: las lapas ponen más huevos durante la primavera si la densidad es baja y menos huevos durante el verano si la densidad es alta.

¿Cómo calcular los cuadrados medios (MS) utilizando estadística descriptiva?

Queremos calcular MS del modelo, MS Between (MSB) de cada factor, y MS Within (MSW) utilizando sólo estadística descriptiva y no a partir de SS, esto puede permitir reducir el número de cálculos y los errores, especialmente si se hacen manualmente.

MSB es una estimación de la varianza total basada en la variabilidad entre las medias de cada grupo, por lo tanto:

$$MSB_{factor1} = bn\sigma_{y_{i..}}^2$$

$$MSB_{factor2} = an\sigma_{y_{.j.}}^2$$

$$MSB_{Model} = n\sigma_{y_{ij.}}^2$$

Donde $\sigma_{y_{i..}}^2$ es la varianza de las medias de cada categoría de factor 1, $\sigma_{y_{.j.}}^2$ es la varianza de las medias de cada categoría del factor 2, $\sigma_{y_{ij.}}^2$ es la varianza de las medias de cada categoría del factor 1 y 2, a es el número de categorías en el factor 1, b es el número de categorías en el factor 2, y n es el número de replicas por grupo de tratamiento.

Necesitamos calcular la media y la varianza para cada una de las combinaciones de temporada y densidad (8 combinaciones). Para ayudarte a hacerlo usaremos una librería llamada `data.table`. Es una librería diseñada para trabajar con grandes conjuntos de datos, por lo que tiene funciones útiles para realizar cálculos rápidos (puede realizar las mismas operaciones al subconjuntar el `data.frame` utilizado).

También introduciremos dos nuevas funciones básicas de R: `colMeans()` y `rowMeans()`, calculan los promedios de fila y columna.

```

library(data.table)
# calculate the mean value of eggs for each combination of density and
season
means_all <- dcast(setDT(limpet), DENSITY~SEASON, value.var = "EGGS",
fun=mean)
means_all

##      DENSITY spring summer
## 1:         8  2.417 1.8333
## 2:        15  2.177 1.1777
## 3:        30  1.565 0.8113
## 4:        45  1.200 0.5927

#MSB (DENSITY)
means_density <- rowMeans(means_all[,2:3])
means_density

## [1] 2.1250 1.6775 1.1883 0.8962

MSB_density <- 2*3*var(means_density)
MSB_density

## [1] 1.761

#MSB (SEASON)
means_season <- colMeans(means_all[,2:3]) # calculate the means of cols 2
and 3
means_season

## spring summer
## 1.840 1.104

MSB_season <- 4*3*var(means_season)
MSB_season

## [1] 3.25

#MSB (Model) This value is not in the summary of the anova results
MSB_model <- 3 * var(c(means_all$spring,means_all$summer))
MSB_model

## [1] 1.243

```

MSW también es una estimación de la varianza total, basada en la varianza dentro de cada grupo. Entonces, escribimos:

$$MSW = \overline{\sigma_{ij}^2}$$

Podemos calcular MSW utilizando:

```

# calculate the variance for each combination of density and season
var_all <- dcast(setDT(limpet), DENSITY~SEASON, value.var = "EGGS",
fun=var)
var_all

##   DENSITY spring summer
## 1:      8 0.34896 0.09896
## 2:     15 0.14399 0.23239
## 3:     30 0.38563 0.16930
## 4:     45 0.03611 0.04201

## MSW (Residuals)
MSW <- mean(c(var_all$spring, var_all$summer) )
MSW

## [1] 0.1822

```

F-statistic y p-value

El estadístico F se puede calcular utilizando:

$$Fstat = \frac{MSB}{MSW}$$

Y el p-valor que indica si este es un valor extremo y por tanto debe rechazarse la hipótesis nula:

$$p - value = 1 - pf(Fstat, df1, df2)$$

Ahora tenemos que aplicar la fórmula para cada factor:

```

# 4 categories in DENSITY, 2 categories in SEASON, 3 replicates
F_stat_density<-MSB_density/MSW
F_stat_density

## [1] 9.669

1-pf(F_stat_density,df1=4-1,df2=4*2*(3-1)) # Pvalue density

## [1] 0.0007041

F_stat_season<-MSB_season/MSW
F_stat_season

## [1] 17.84

1-pf(F_stat_season,df1=2-1,df2=4*2*(3-1)) # Pvalue season

```


Puede observarse tras haber calculado F y p-valor que obtenemos el mismo resultado que con la función `aov()`.

2.9.2. Ejemplo 2: Influencia de la dieta y del género en la pérdida de peso

Este ejemplo está fundado en los que se explican en las clases de estadística y que se basan en los diferentes modelos creados por Hafid Laayouni y colaboradores de la UPF (Barcelona).

El conjunto de datos `diet.csv` contiene información sobre 78 personas que realizaron una de tres dietas posibles. También se incluyen otras metadatos como información de antecedentes como la edad, el género (Mujer = 0, Hombre = 1) y la altura. El objetivo del estudio fue ver qué dieta era la mejor para perder peso, pero también se pensó que las mejores dietas para hombres y mujeres pueden ser diferentes, por lo que las variables independientes son la dieta y el género.



Imagen obtenida de: <https://xandooworld.com/how-to-make-your-diet-fit-with-your-lifestyle/>

Para abrir el archivo use el comando **read.csv()**. Deberá cambiar el comando según dónde haya guardado el archivo.

Calcule el peso perdido (weight.loss) por persona (diferencia de peso antes y después de la dieta) y agregue la variable al conjunto de datos.

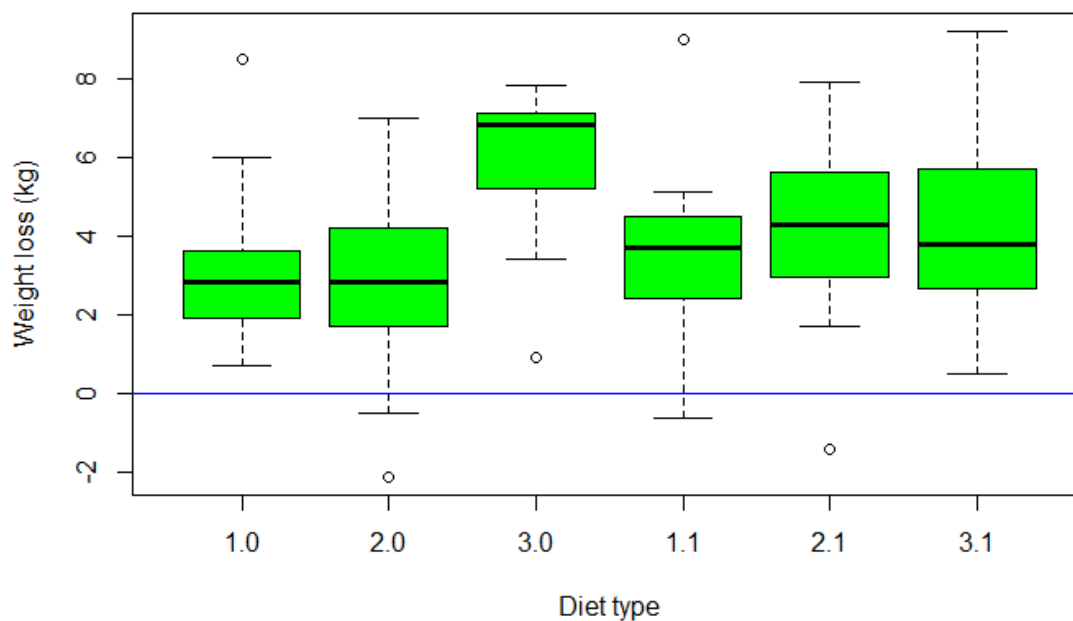
Representar en un "box plot": `weight.loss~diet.type*gender`

```
#open the file  
diet = read.csv("diet.csv", row.names=1)  
#Calculate the weight lost by person (difference in weight before and  
after the diet) and add the variable to the dataset.  
diet$weight.loss = diet$initial.weight - diet$final.weight
```

Grafica la dispersión de la variable de interés dependiendo de cada factor por separado y juntos (puede usar ggplot2 o el sistema de trazado R básico).

Por ejemplo se puede utilizar un "box plot":

```
boxplot(weight.loss~diet.type*gender, data=diet, col="light gray",  
        ylab = "Weight loss (kg)", xlab = "Diet type")  
abline(h=0, col="blue")
```



Realiza un anova de dos factores (two way) e interpreta el resultado. ¿Tiene alguno de los factores un efecto sobre la pérdida de peso (Kg)? ¿Y su interacción? Haz una gráfica de interacción que relacione ambos factores y la variable principal?

```
# We have 2 factors: gender and diet.type
# diet.type
table(diet$diet.type)

##
##  1  2  3
## 24 27 27

# 3 types of lure (scent, sugar and chemical). 20 measures per group
table(diet$gender)

##
##  0  1
## 43 33

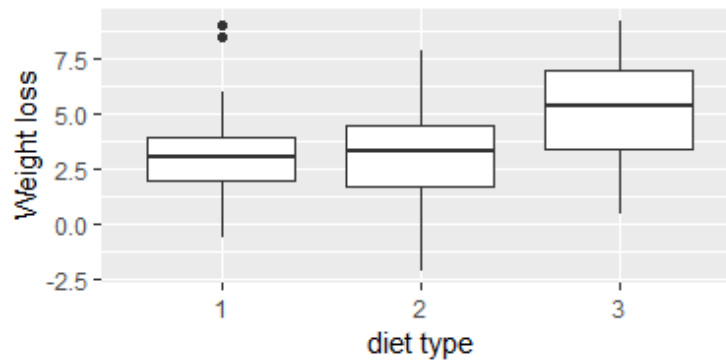
# combination
table(diet$diet.type, diet$gender)

##
##      0  1
##  1 14 10
##  2 14 11
##  3 15 12

#factor
diet$diet.type<- as.factor(diet$diet.type)
diet$gender<- as.factor(diet$gender)

gb <- ggplot(diet, aes(x = diet.type, y = weight.loss))
gb <- gb + xlab("diet type") + ylab("Weight loss")
gb <- gb + geom_boxplot()
plot(gb)
```

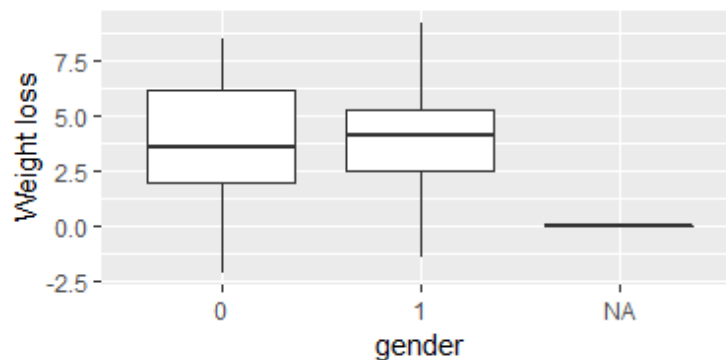
Así el resultado del “box plot” para dietas:



Con la librería ggplot2:

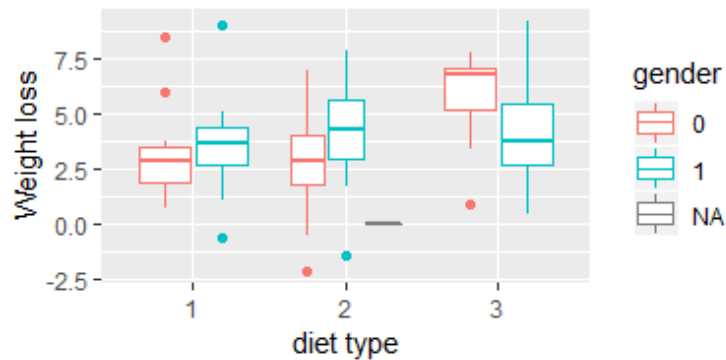
```
gb <- ggplot(diet, aes(x = gender, y = weight.loss))
gb <- gb + xlab("gender") + ylab("Weight loss")
gb <- gb + geom_boxplot()
plot(gb)
```

por sexo, donde se puede observar que hay un dato desconocido:



Conjuntamente: dieta y sexo:

```
gb <- ggplot(diet, aes(x = diet.type, y = weight.loss, color = gender))
gb <- gb + xlab("diet type") + ylab("Weight loss")
gb <- gb + geom_boxplot()
plot(gb)
```



Ahora se realiza un análisis de la varianza ANOVA, que corresponde a 2 factores con interacción, y tras éste, un análisis de medias “a posteriori” utilizando `pairwise.t.test()`:

```
#ANOVA two way
diet.anova = aov(weight.loss~diet.type*gender,data=diet)
summary(diet.anova)

##              Df Sum Sq Mean Sq F value Pr(>F)
## diet.type     2     61   30.26   5.63 0.0054 **
## gender        1      0    0.17    0.03 0.8599
## diet.type:gender 2     34   16.95   3.15 0.0488 *
## Residuals    70    376    5.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness

pairwise.t.test(diet$weight.loss, diet$diet.type, p.adj = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  diet$weight.loss and diet$diet.type
##
##   1      2
## 2 1.000 -
## 3 0.022 0.005
##
## P value adjustment method: bonferroni

# diet has an effect on weight.loss: 3 and 2 are different from 1
pairwise.t.test(diet$weight.loss, diet$gender, p.adj = "bonf")

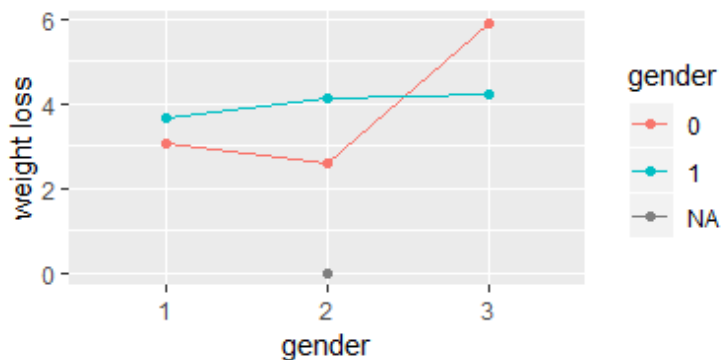
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  diet$weight.loss and diet$gender
##
##   0
```

```
## 1 0.8
##
## P value adjustment method: bonferroni

# Type of gender does not have an effect on weight.Loss
# There is combined effect of gender and type on weight.Loss
```

Ahora podemos realizar un gráfico de interacción, donde se representan dlos 2 factores y la variable principal:

```
gi <- ggplot(diet, aes(x = diet.type, y = weight.loss, colour = gender,
group=gender))
gi <- gi + xlab("gender") + ylab("weight loss")
gi <- gi + stat_summary(fun.y=mean, geom="point")+
  stat_summary(fun.y=mean, geom="line")
plot(gi)
```



```
# The diet is 3
# The best combination is diet 3 in a gender=0
dcast(setDT(diet), diet.type~gender, value.var = "weight.loss", fun=mean)

##   diet.type    0    1  NA
## 1:         1 3.050 3.650 NaN
## 2:         2 2.607 4.109  0
## 3:         3 5.880 4.233 NaN

#see the results at:

#see at <https://bioinformatics-core-shared-training.github.io/linear-models-r/ANOVA.html>
```

```
#<https://www.sheffield.ac.uk/polopoly\_fs/1.536444!/file/MASH\_2way\_ANOVA\_in\_R.pdf>
```

2.9.3. Cebos para polillas

Este ejemplo también está basado en los que se explican en las clases de estadística y que se basan en los creados por Hafid Laayouni y colaboradores de la UPF (Barcelona).

En este ejercicio, utilizaremos datos de un experimento con polillas en el que los investigadores querían saber cuál es el mejor lugar para colocar una trampa y qué cebo (lure) es más eficaz para capturar polillas. Mira los datos: ¿Cuántos factores tenemos? ¿Cuántas categorías hay en cada factor? ¿Cuál es el tamaño de cada grupo?



Imagen de: https://www.videoblocks.com/video/moths-around-light-in-a-dark-place-hb54adx_iqynqstq

Grafica la dispersión de la variable de interés dependiendo de cada factor por separado y juntos (puede usar ggplot2 o el sistema de trazado R básico). Realiza un anova de dos vías e interpreta el resultado. ¿Tiene alguno de los factores un efecto en

el número de polillas capturadas? ¿Y su interacción? Haz una gráfica de interacción indicando el factor en el eje x.

Primeramente vamos a leer los datos a partir del fichero "moth_experiment.csv"

Y vamos a dibujar un "box plot":

```
moth <- read.csv("moth_experiment.csv", header = TRUE)
# We have 2 factors: Location and Lure
# 4 Locations (top, middle, Lower and ground). 15 measures per group
table(moth$location)

##
## Ground Lower Middle Top
## 15 15 15 15

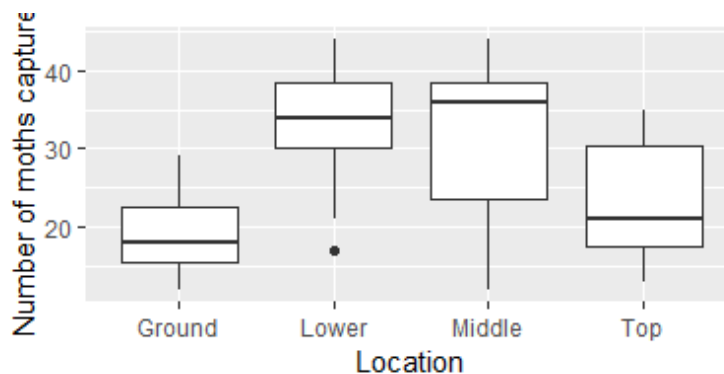
# 3 types of Lure (scent, sugar and chemical). 20 measures per group
table(moth$lure)

##
## Chemical Scent Sugar
## 20 20 20

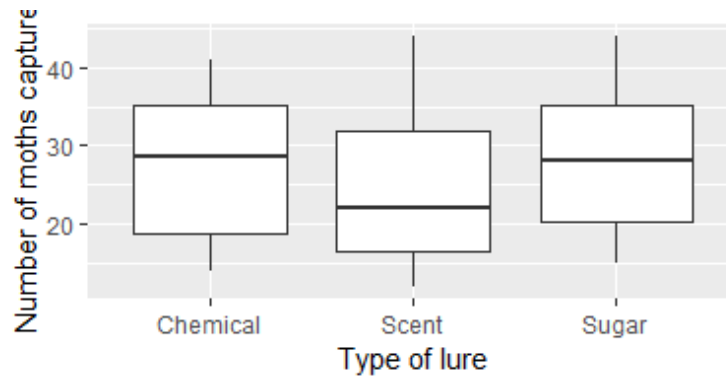
# 5 measures per combination
table(moth$location,moth$lure)

##
## Chemical Scent Sugar
## Ground 5 5 5
## Lower 5 5 5
## Middle 5 5 5
## Top 5 5 5

gb <- ggplot(moth, aes(x = location, y = numberMoths))
gb <- gb + xlab("Location") + ylab("Number of moths captured")
gb <- gb + geom_boxplot()
plot(gb)
```

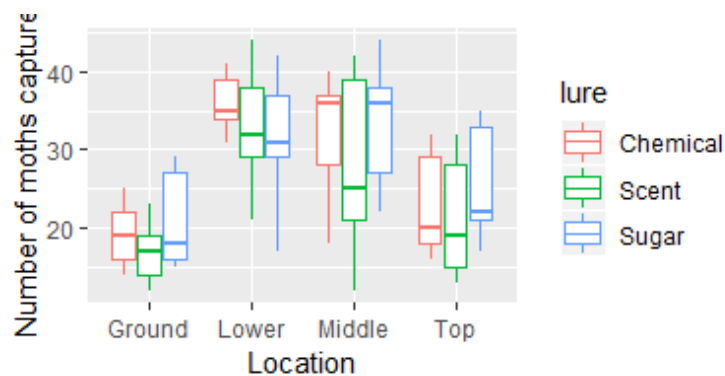



```
gb <- ggplot(moth, aes(x = lure, y = numberMoths))
gb <- gb + xlab("Type of lure") + ylab("Number of moths captured")
gb <- gb + geom_boxplot()
plot(gb)
```



Ahora utilizando la librería **ggplot2()**:

```
gb <- ggplot(moth, aes(x = location, y = numberMoths, color = lure))
gb <- gb + xlab("Location") + ylab("Number of moths captured")
gb <- gb + geom_boxplot()
plot(gb)
```



El análisis de la varianza ANOVA, poniendo 2 factores fijos con interacción:

```
moth.anova <- aov(moth$numberMoths~moth$location*moth$lure)
moth.anova

## Call:
## aov(formula = moth$numberMoths ~ moth$location * moth$lure)
##
## Terms:
##          moth$location  moth$lure  moth$location:moth$lure
Residuals
## Sum of Squares          1981          113          115
3034
## Deg. of Freedom          3          2          6
48
```

```
##
## Residual standard error: 7.95
## Estimated effects may be unbalanced

summary(moth.anova)

##              Df Sum Sq Mean Sq F value    Pr(>F)
## moth$location    3   1981     660   10.45 0.000021 ***
## moth$lure        2    113      57    0.89   0.42
## moth$location:moth$lure 6    115      19    0.30   0.93
## Residuals       48   3034      63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora vamos a realizar un test de medias “a posteriori” del ANOVA con la función:

`pairwise.t.test()`, a fin de identificar que pares de categorías son diferentes:

```
pairwise.t.test(moth$numberMoths, moth$location, p.adj = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  moth$numberMoths and moth$location
##
##      Ground  Lower Middle
## Lower 0.00002 -      -
## Middle 0.00044 1.000 -
## Top    0.788  0.004 0.048
##
## P value adjustment method: bonferroni

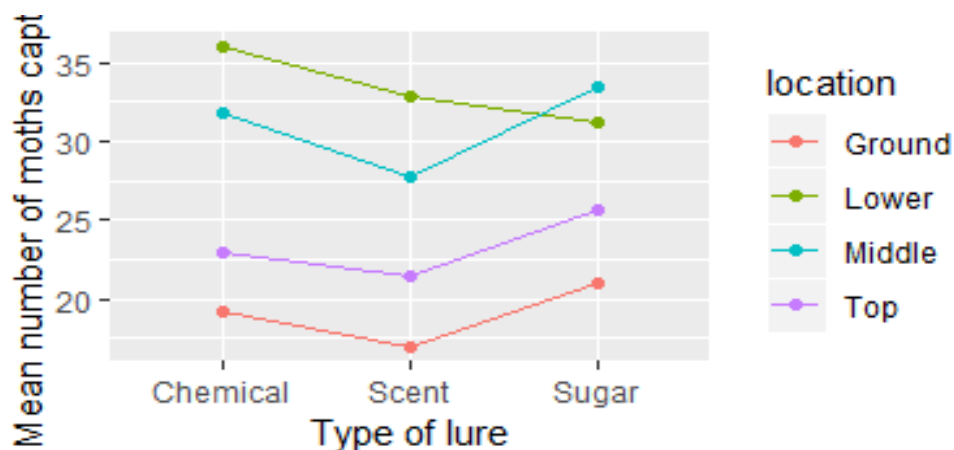
# Location has an effect on number of moths: Lower and middle are
different from ground and top
pairwise.t.test(moth$numberMoths, moth$lure, p.adj = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  moth$numberMoths and moth$lure
##
##      Chemical Scent
## Scent 1.0      -
## Sugar 1.0      0.9
##
## P value adjustment method: bonferroni
```

```
# Type of lure does not have an effect on number of moths
# There is no combined effect of location and type of lure on number of moths.
```

Una Buena forma de interpretar la interacción es realizar un gráfico de tipo interacción:

```
gi <- ggplot(moth, aes(x = lure, y = numberMoths, colour = location,
group=location))
gi <- gi + xlab("Type of lure") + ylab("Mean number of moths captured")
gi <- gi + stat_summary(fun.y=mean, geom="point")+
  stat_summary(fun.y=mean, geom="line")
plot(gi)
```



Finalmnte la interpretación y algunas conclusiones sobre los resultados obtenidos:

```
# The best location is in a Lower position
# The best lure is sugar
# The best combination is a Lower position with chemical
dcast(setDT(moth), location~lure, value.var = "numberMoths", fun=mean)
```

```
## location Chemical Scent Sugar
## 1: Ground 19.2 17.0 21.0
## 2: Lower 36.0 32.8 31.2
## 3: Middle 31.8 27.8 33.4
```

2.9.5. Dos factores aleatorios: ejemplo de la eficacia de la penicilina como inhibidor del crecimiento de *Bacillus subtilis*.

Veamos otro ejemplo clásico con 2 factores aleatorios de tipo microbiológico donde hay dos factores aleatorios, muestras de penicilina y placas, los dos de tipo aleatorio cogidos al azar de una muestra más grande.



Penicilina inhibiendo el crecimiento de bacterias (Imagen de <https://www.caracteristicas.co/wp-content/uploads/2019/02/penicilina-hongo-e1550273946338.jpg>)

Veamos la realización y análisis realizado:

```
#####  
# 2-way additive random effects ANOVA, no replicates (from lme4)  
#####  
  
#Six samples of penicillin were tested using the B. subtilis plate method  
on each of 24 plates. The response is the diameter (mm) of the zone of  
inhibition of growth of #the organism.
```

```

#A data frame with 144 observations on the following 3 variables.
#diameter (mm) of the zone of inhibition of the growth of the organism.
#plate: assay plate. A factor with levels 'a' to 'x'.
#sample: penicillin sample. A factor with levels 'A' to 'F'.

#?Penicillin
head(Penicillin)

##   diameter plate sample
## 1       27     a      A
## 2       23     a      B
## 3       26     a      C
## 4       23     a      D
## 5       23     a      E
## 6       21     a      F

str(Penicillin)

## 'data.frame':   144 obs. of  3 variables:
## $ diameter: num  27 23 26 23 23 21 27 23 26 23 ...
## $ plate   : Factor w/ 24 levels "a","b","c","d",...: 1 1 1 1 1 1 2 2 2
## $ sample  : Factor w/ 6 levels "A","B","C","D",...: 1 2 3 4 5 6 1 2 3
##           4 ...

# no replicates - therefore can only fit an additive model
xtabs(~ sample + plate, Penicillin)

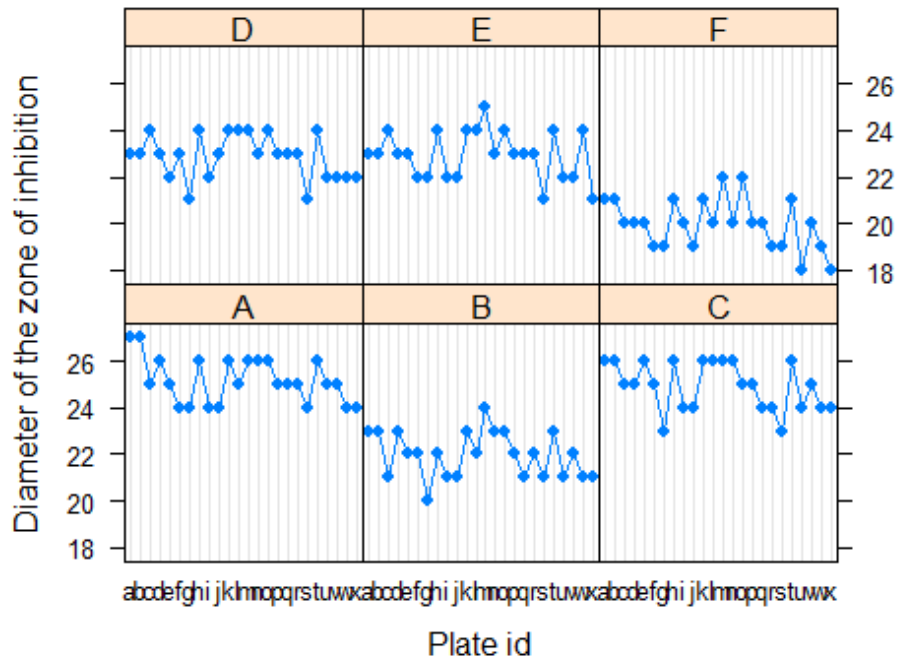
##      plate
## sample a b c d e f g h i j k l m n o p q r s t u v w x
##      A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      D 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      E 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

table(Penicillin$sample, Penicillin$plate) # equivalent

##
##      a b c d e f g h i j k l m n o p q r s t u v w x
##      A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      D 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      E 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```

```
# data visualization: original order
dotplot(diameter ~ plate | sample, Penicillin,
        type = c("p", "a"), xlab = "Plate id",
        ylab = "Diameter of the zone of inhibition")
```



```
# 2-way additive model
fit2 <- lmer(diameter ~ 1 + (1|plate) + (1|sample), Penicillin)
summary(fit2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
## Data: Penicillin
##
## REML criterion at convergence: 330.9
##
## Scaled residuals:
##   Min       1Q   Median       3Q      Max
## -2.07922 -0.67139  0.06292  0.58377  2.97957
##
## Random effects:
##   Groups      Name          Variance Std.Dev.
##   plate      (Intercept)  0.7169   0.8467
##   sample     (Intercept)  3.7311   1.9316
##   Residual                    0.3024   0.5499
## Number of obs: 144, groups:  plate, 24; sample, 6
##
## Fixed effects:
```

```

##           Estimate Std. Error t value
## (Intercept) 22.9722      0.8086  28.41

# access model-based quantities
fixef(fit2)

## (Intercept)
##      22.97222

ranef(fit2, drop = TRUE) # note no constraint

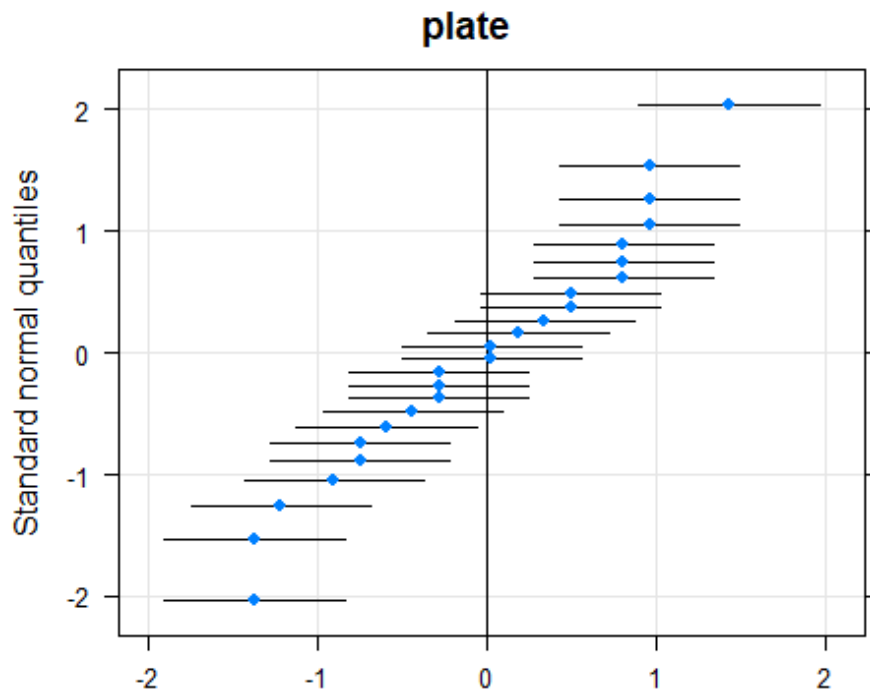
## $plate
##           a           b           c           d           e
f
##  0.80454389  0.80454389  0.18167120  0.33738937  0.02595303 -
0.44120149
##           g           h           i           j           k
l
## -1.37551052  0.80454389 -0.75263783 -0.75263783  0.96026206
0.49310755
##           m           n           o           p           q
r
##  1.42741658  0.49310755  0.96026206  0.02595303 -0.28548332 -
0.28548332
##           s           t           u           v           w
x
## -1.37551052  0.96026206 -0.90835601 -0.28548332 -0.59691966 -
1.21979235
## attr(,"postVar")
## [1] 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482
## [7] 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482
## [13] 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482
## [19] 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482 0.07363482
##
## $sample
##           A           B           C           D           E
F
##  2.18705819 -1.01047625  1.93789966 -0.09689498 -0.01384214 -
3.00374447
## attr(,"postVar")
## [1] 0.04087179 0.04087179 0.04087179 0.04087179 0.04087179 0.04087179
##
## with conditional variances for "plate" "sample"

# 'drop' changes the output format from matrix to vector

# check assumption of normality
qqmath(ranef(fit2, postVar = TRUE), strip=FALSE)$plate

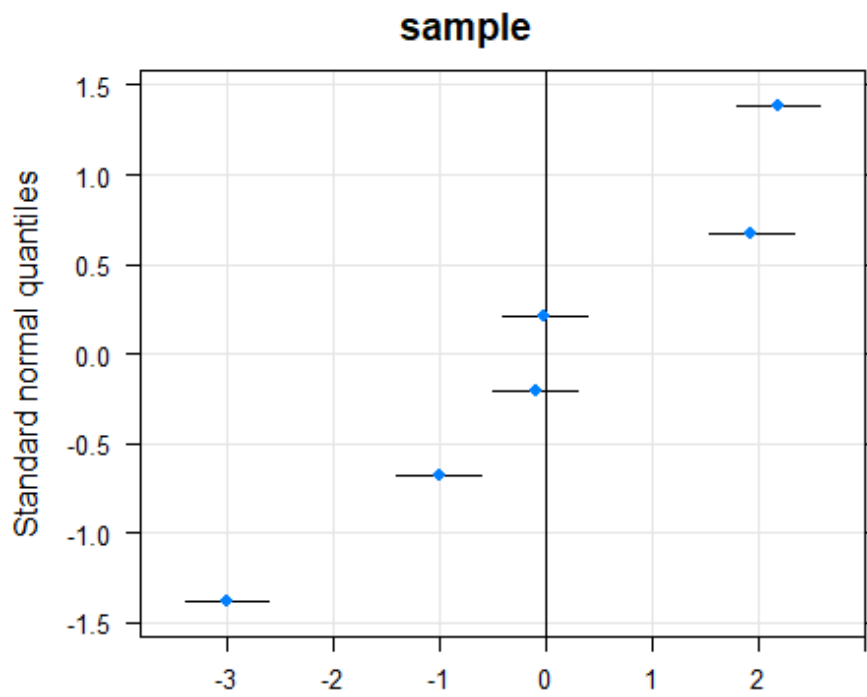
```

```
## Warning in ranef.merMod(fit2, postVar = TRUE): 'postVar' is
depreciated:
## please use 'condVar' instead
```

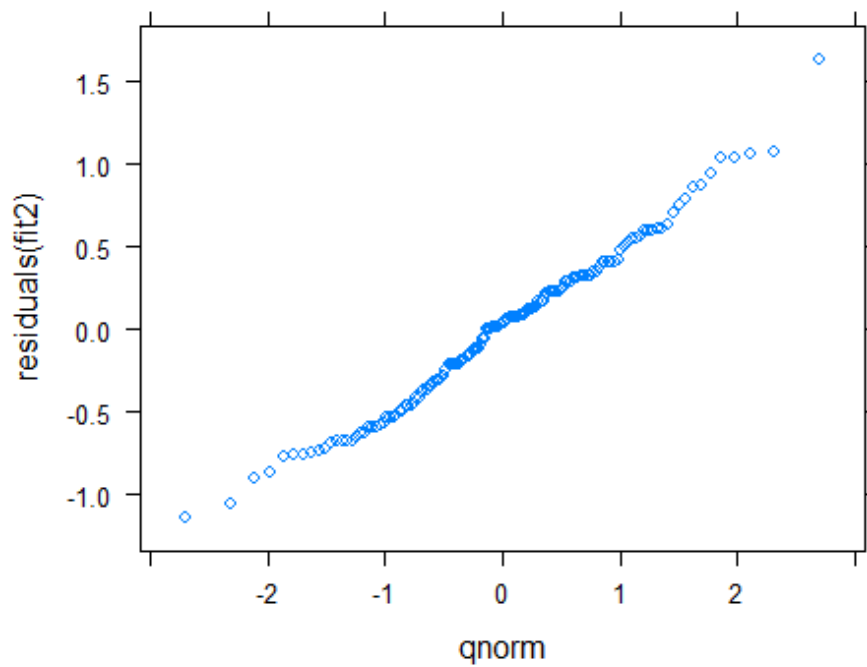


```
qqmath(ranef(fit2, postVar = TRUE), strip=FALSE)$sample
```

```
## Warning in ranef.merMod(fit2, postVar = TRUE): 'postVar' is
depreciated:
## please use 'condVar' instead
```

```
qqmath(residuals(fit2))
```



```
# Testing H0: var(sample) = 0 vs Ha: var(sample) != 0
# Note the use of ML instead of REML
```

```

fit3.Ha <- lmer(diameter ~ 1 + (1|plate) + (1|sample), Penicillin,
REML=FALSE)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| =
0.00287857
## (tol = 0.002, component 1)

summary(fit3.Ha)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
## Data: Penicillin
##
##      AIC      BIC   logLik deviance df.resid
##  340.2    352.1  -166.1   332.2     140
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.08017 -0.67221  0.06411  0.58356  2.97909
##
## Random effects:
## Groups   Name              Variance Std.Dev.
## plate    (Intercept)  0.7146   0.8453
## sample   (Intercept)  3.1344   1.7704
## Residual                    0.3025   0.5500
## Number of obs: 144, groups:  plate, 24; sample, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  22.9722    0.7445   30.86
## convergence code: 0
## Model failed to converge with max|grad| = 0.00287857 (tol = 0.002,
component 1)

fit3.H0 <- lmer(diameter ~ 1 + (1|plate), Penicillin, REML=FALSE)
summary(fit3.H0)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: diameter ~ 1 + (1 | plate)
## Data: Penicillin
##
##      AIC      BIC   logLik deviance df.resid
##  617.6    626.5  -305.8   611.6     141
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.43426 -0.53337  0.01264  0.56756  1.96875
##
## Random effects:
## Groups   Name              Variance Std.Dev.

```

```

## plate (Intercept) 0.06312 0.2512
## Residual 4.03333 2.0083
## Number of obs: 144, groups: plate, 24
##
## Fixed effects:
## Estimate Std. Error t value
## (Intercept) 22.972 0.175 131.2

# Likelihood ratio test
testStat <- as.numeric(2*( logLik(fit3.Ha) - logLik(fit3.H0)))
testStat

## [1] 279.4412

pchisq(testStat, 1, lower=FALSE)

## [1] 9.939597e-63

anova(fit3.Ha,fit3.H0) # equivalent

## Data: Penicillin
## Models:
## fit3.H0: diameter ~ 1 + (1 | plate)
## fit3.Ha: diameter ~ 1 + (1 | plate) + (1 | sample)
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## fit3.H0 3 617.63 626.54 -305.81 611.63
## fit3.Ha 4 340.19 352.07 -166.09 332.19 279.44 1 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# lmer does not provide confidence intervals on variance components on
# purpose
# (Developer Doug Bates: the asymptotic distribution of MLE is
# not appropriate for variance)

```

2.9.6 ANOVA multifactorial robusto: test de permutaciones librería RVAideMemoire, uso en múltiples factores.

Si hemos de analizar unos datos en que no existe cumplimiento de alguno de los supuestos del ANOVA: heterocedasticidad y no normalidad de los residuos, y es un modelo con varios factores, existe una alternativa: los test de permutaciones, como el que puede encontrarse en la librería RVAideMemoire, muy reciente, desarrollada en 2015:

<https://cran.r-project.org/web/packages/RVAideMemoire/index.html>

“Contains miscellaneous functions useful in biostatistics, mostly univariate and multivariate testing procedures with a special emphasis on permutation tests. Many functions intend to simplify user's life by shortening existing procedures or by implementing plotting functions that can be used with as many methods from different packages as possible.”

La idea es calcular el p-valor usando permutaciones. Si no hubiera diferencias entre los valores se podría permutar los valores (todos iguales). Calcular la cantidad de permutaciones que pueden resultar con valores mas extremos.

Vamos a realizar un análisis de permutaciones ANOVA para un experimento con 1 a 3 factores. Para 2 y 3 factores el diseño experimento debe ser equilibrado (balanceado). Para 2 factores, los factores pueden ser cruzados con o sin interacción, o anidados. El segundo factor puede ser un factor de bloqueo (al azar). Para 3 factores, el diseño se limita a 2 factores fijos cruzados (con o sin interacción) dentro de los bloques.

Veamos un ejemplo con datos simulados, no normales y sin homogeneidad de varianzas entre los niveles, y con 4 factores:

```
library(RVAideMemoire)

## *** Package RVAideMemoire v 0.9-72 ***

set.seed(1203)
response <- c(rpois(12,4),rpois(12,0.5),rnorm(12,3,1))
fact1 <- gl(3,12,labels=LETTERS[1:3])
fact2 <- gl(3,1,36,labels=letters[1:3])
fact3 <- gl(6,6,labels=letters[1:6])
block <- gl(2,6,36,labels=letters[1:2])
fact4 <- gl(3,1,36,labels=letters[1:3])
data.ej <- data.frame(response, fact1, fact2, fact3, block, fact4 )
```

Aquí utilizaremos 49 permutaciones. Generalmente cuantas mas permutaciones se realicen mas exacto es el resultado, pero también la velocidad de obtención del mismo es menor.

Veamos como se realiza el análisis ANOVA con el test de permutaciones:

```
res.1<- (aov(response~fact1*fact2))
```

Si comprobamos el cumplimiento de la homogeneidad de varianzas con el test de Bartlett o Levene:

```
bartlett.test(response~fact1)

##
## Bartlett test of homogeneity of variances
##
## data: response by fact1
## Bartlett's K-squared = 12.592, df = 2, p-value = 0.001844
```

```

bartlett.test(response~fact2)

##
## Bartlett test of homogeneity of variances
##
## data: response by fact2
## Bartlett's K-squared = 0.10692, df = 2, p-value = 0.9479

leveneTest(response~fact1)

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2  2.5236 0.09554 .
##      33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

leveneTest(response~fact2)

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2  0.0335 0.9671
##      33

```

Se puede observar que para algunos de los factores de estudio (factor 1) $p < 0.05$ con lo que se rechaza la hipótesis de homocedasticidad.

Si comprobamos la normalidad de los residuos con el test de Shapiro:

```

shapiro.test(residuals(res.1))

##
## Shapiro-Wilk normality test
##
## data: residuals(res.1)
## W = 0.92205, p-value = 0.01447

```

Se puede observar que $p < 0.05$ con lo que se rechaza la hipótesis de normalidad.

Veamos ahora como debe realizarse el ANOVA con permutaciones para las diferentes combinaciones de factores y modelos lineales:

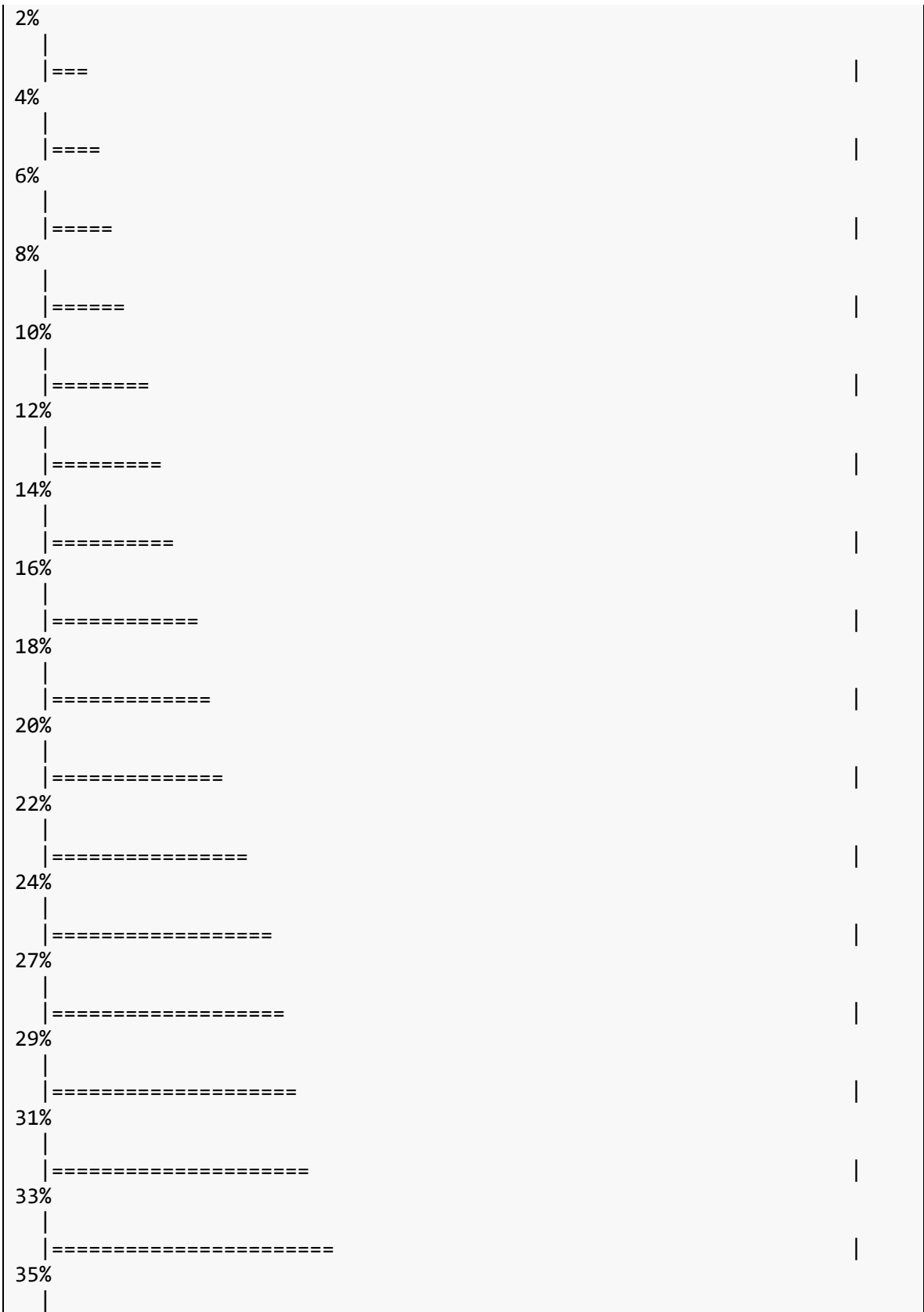
1 fixed factors

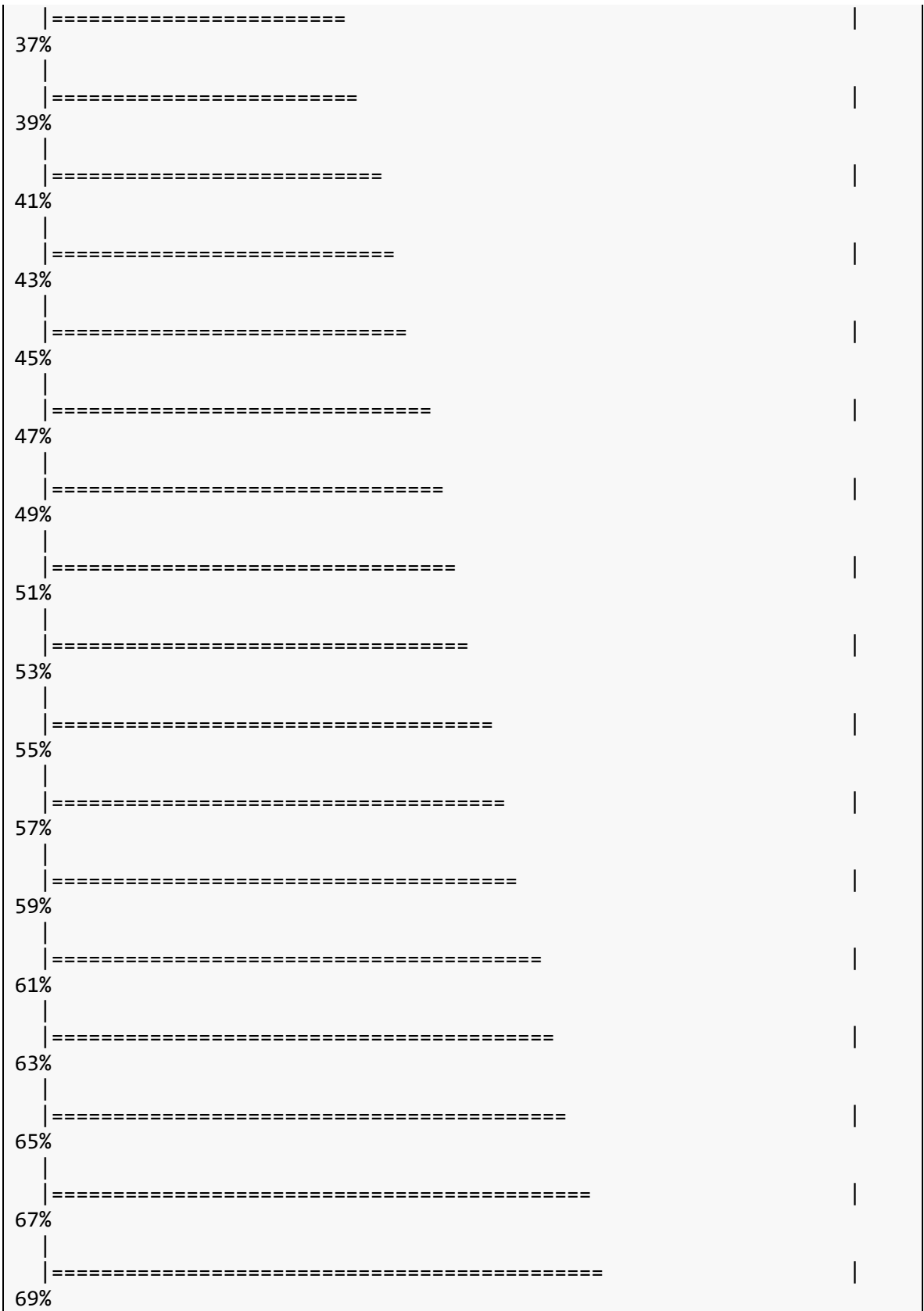
```

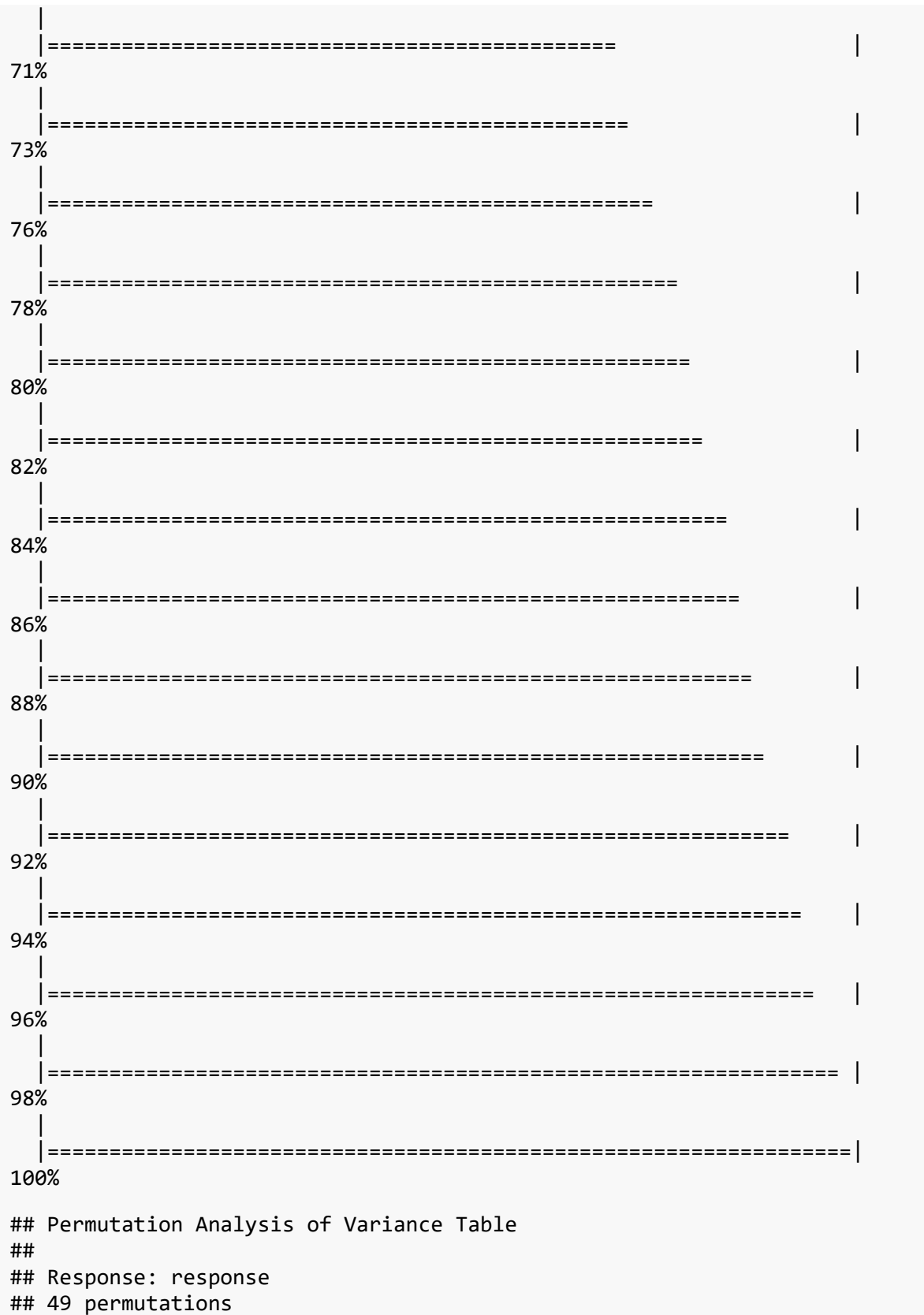
perm.anova(response~fact1, nperm=49)

##
|
|
0% |
|
|= |

```

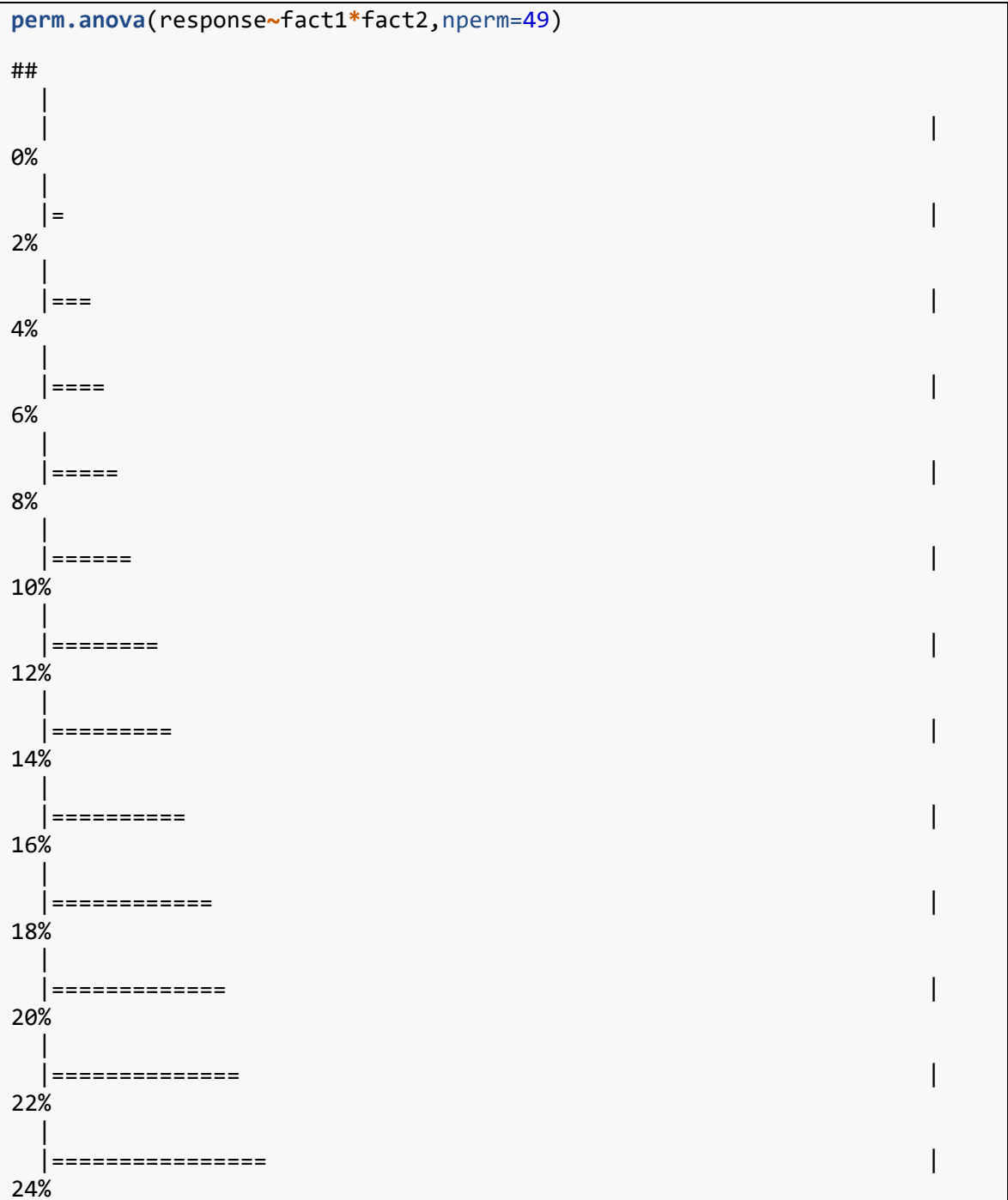


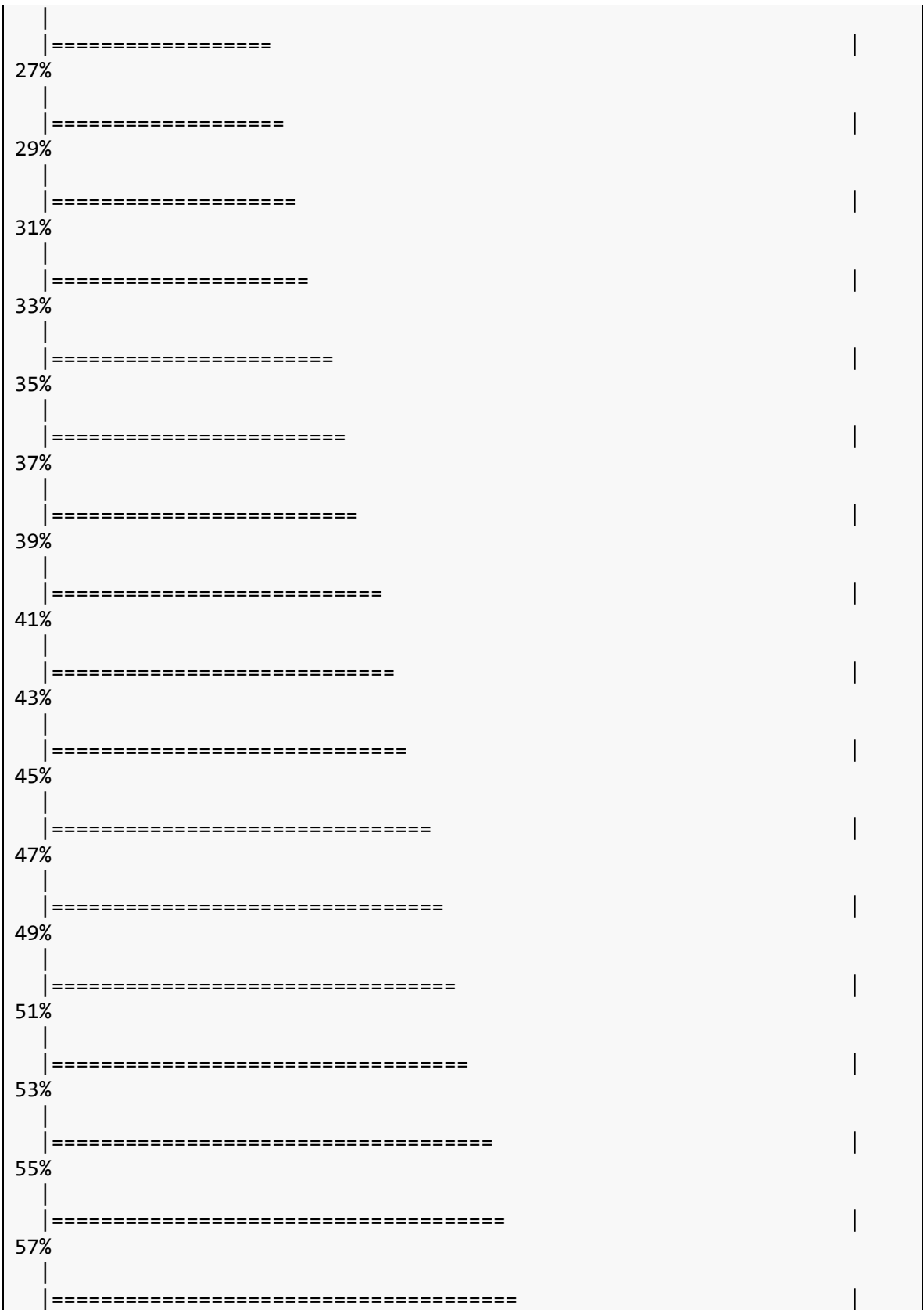


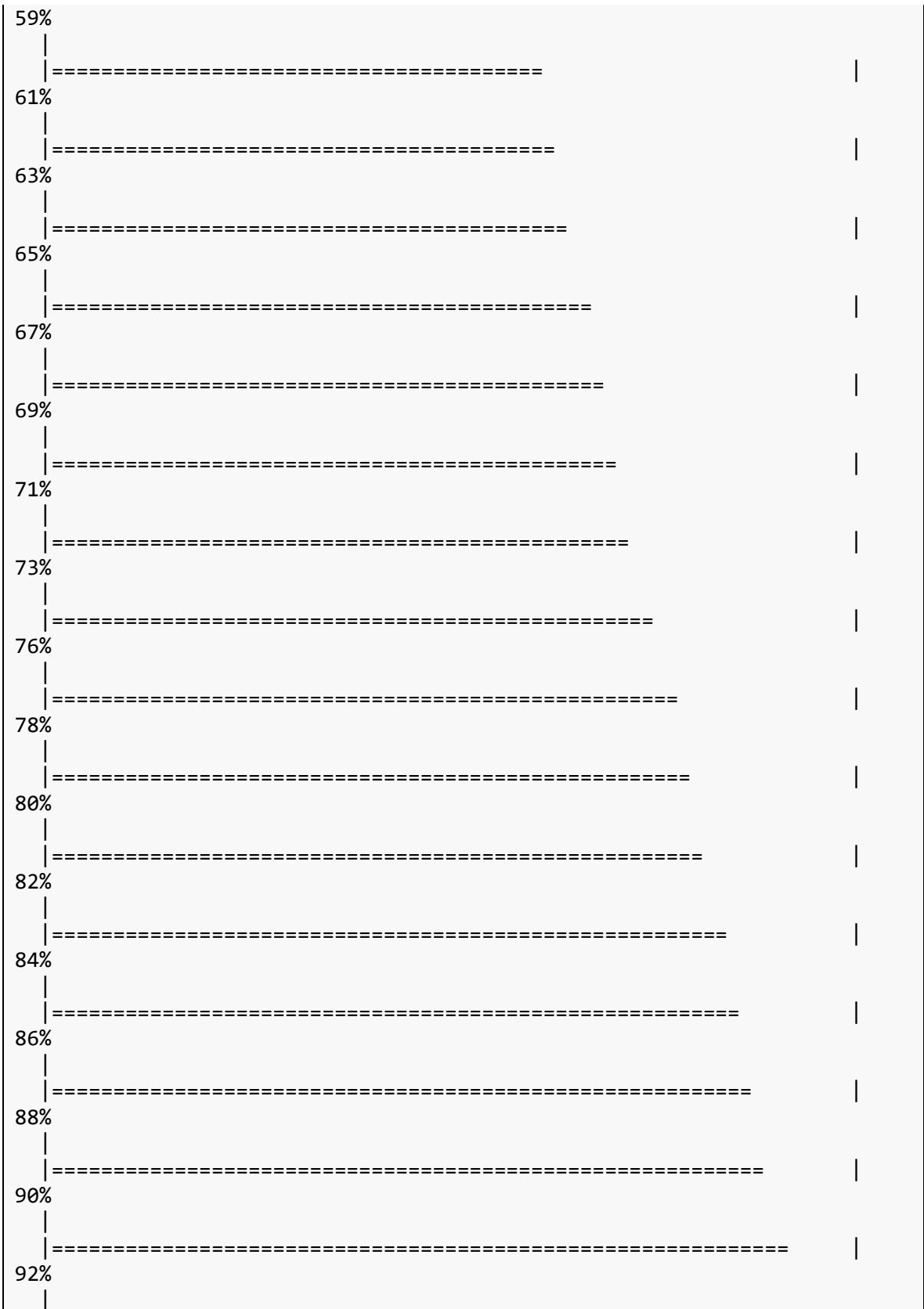



```
##          Sum Sq Df Mean Sq F value Pr(>F)
## fact1      52.086  2  26.0428  15.334  0.02 *
## Residuals 56.047 33  1.6984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 crossed fixed factors with interaction







```

=====
94%
|
=====
96%
|
=====
98%
|
=====
100%

## Permutation Analysis of Variance Table
##
## Response: response
## 49 permutations
##      Sum Sq Df Mean Sq F value Pr(>F)
## fact1    52.086  2  26.0428  13.7025  0.02 *
## fact2     0.273  2   0.1366   0.0719  0.86
## fact1:fact2  4.458  4   1.1145   0.5864  0.70
## Residuals  51.316 27   1.9006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

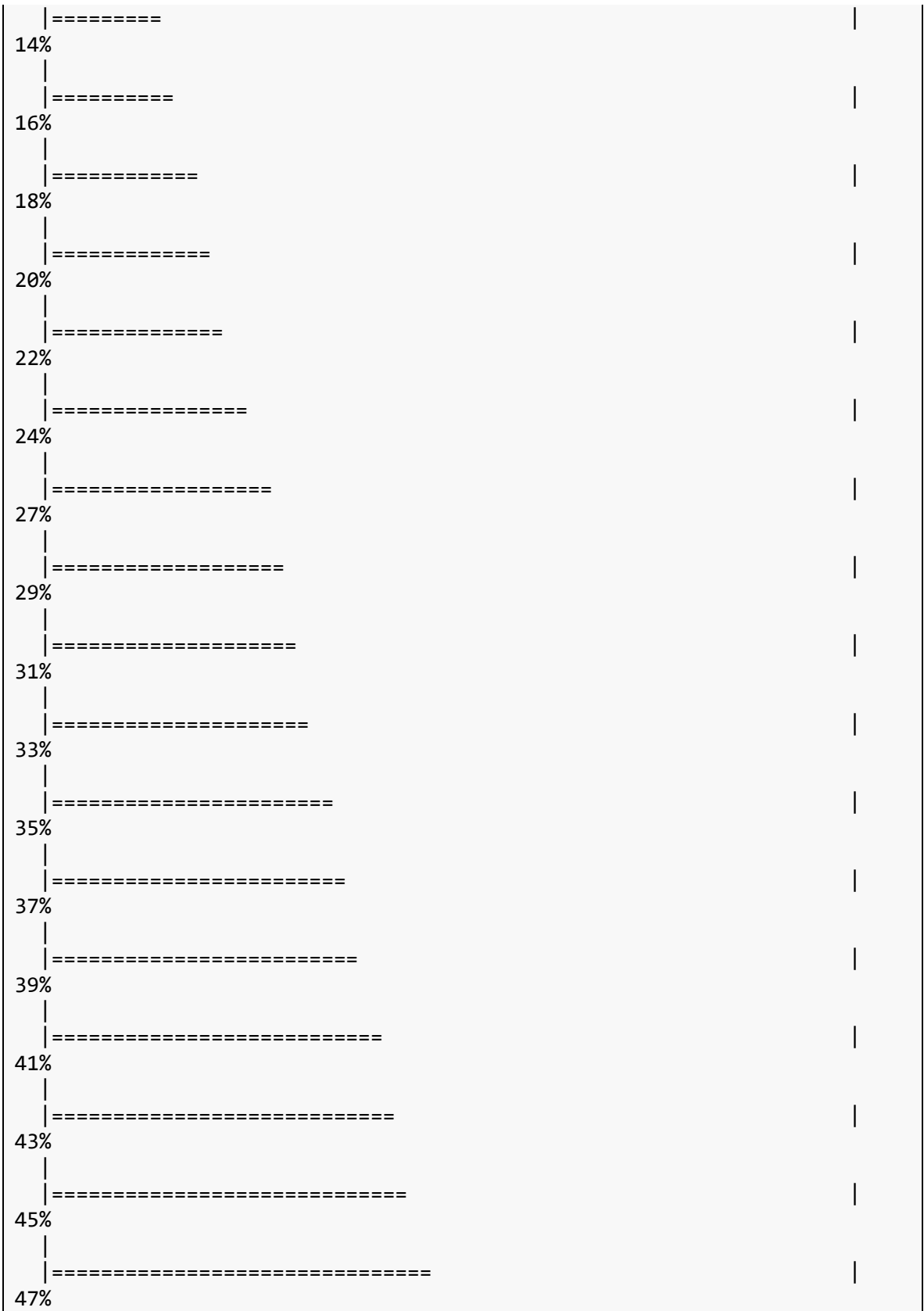
```

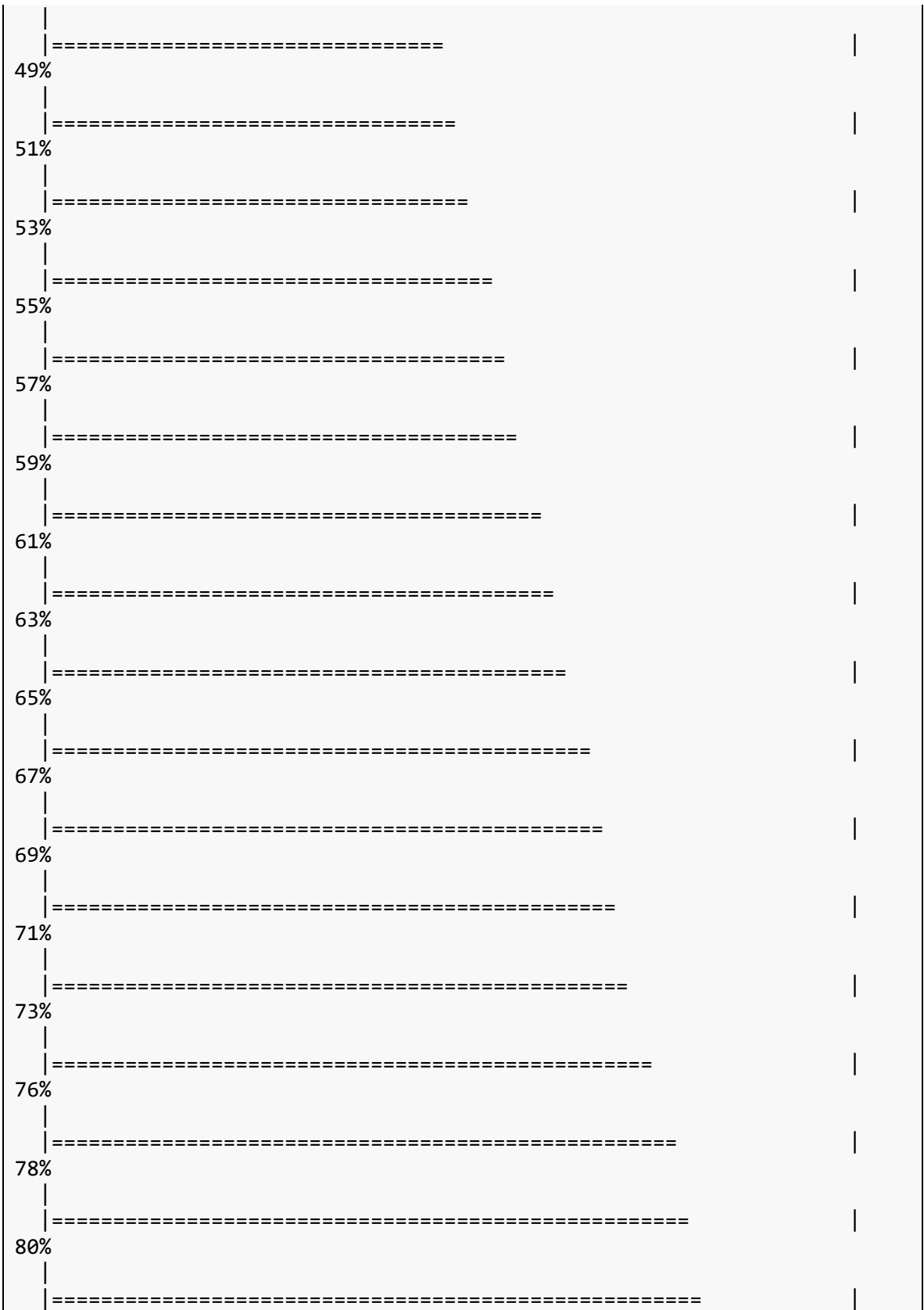
2 nested fixed factors

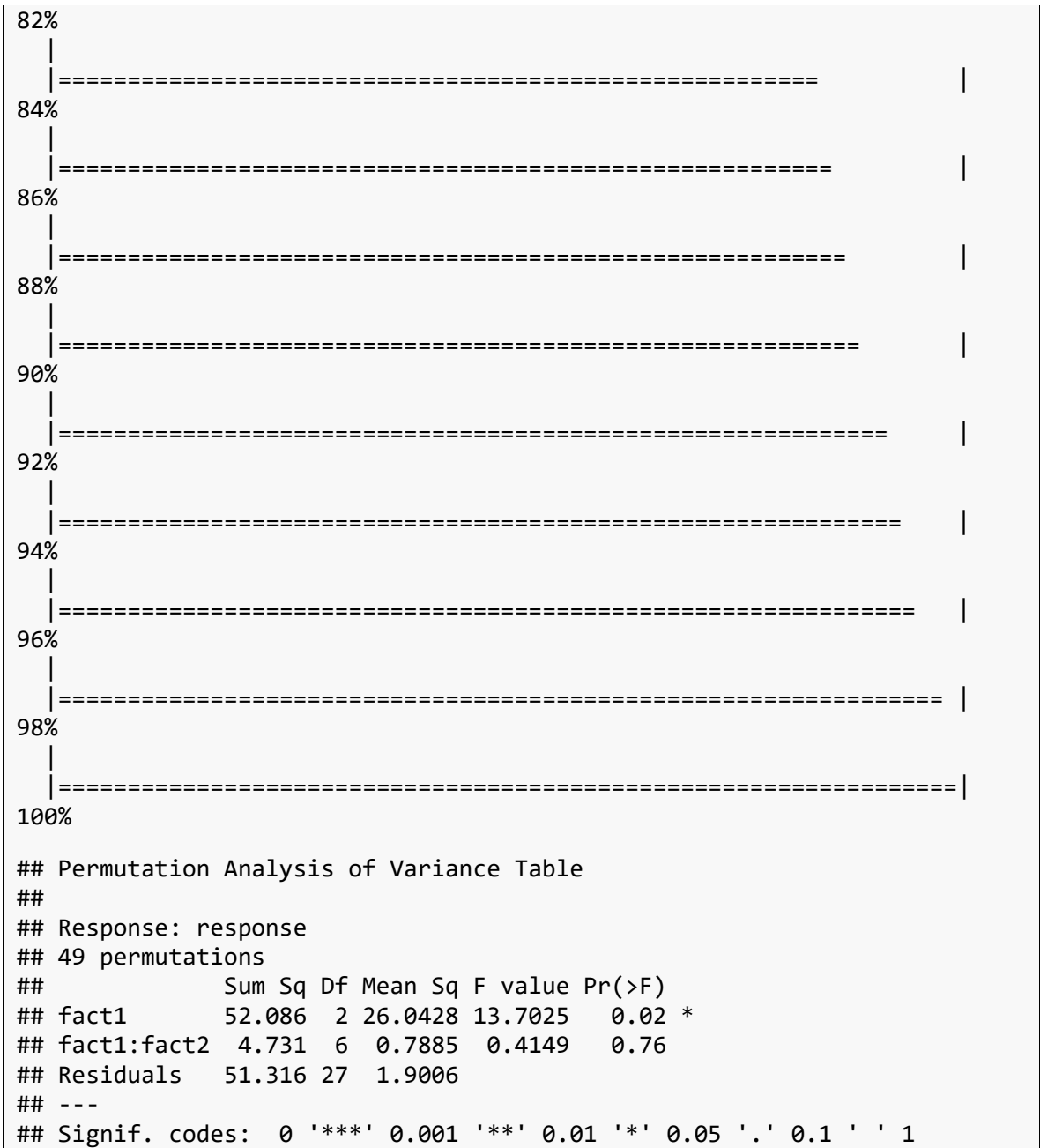
```

perm.anova(response~fact1/fact2,nperm=49)
##
|
0%
|
|=
2%
|
===
4%
|
=====
6%
|
=====
8%
|
=====
10%
|
=====
12%
|

```







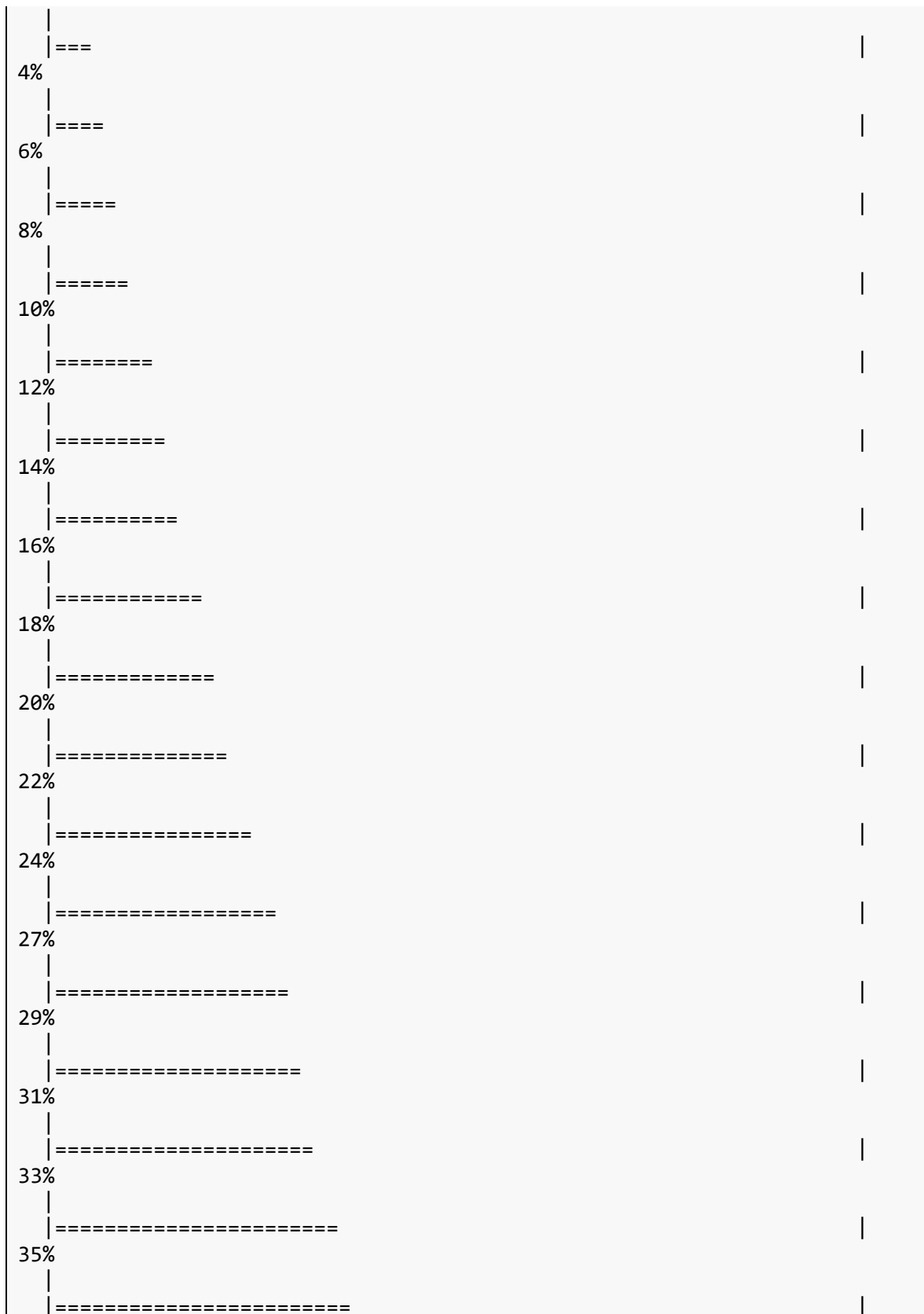
2 nested factors, fact2 being random

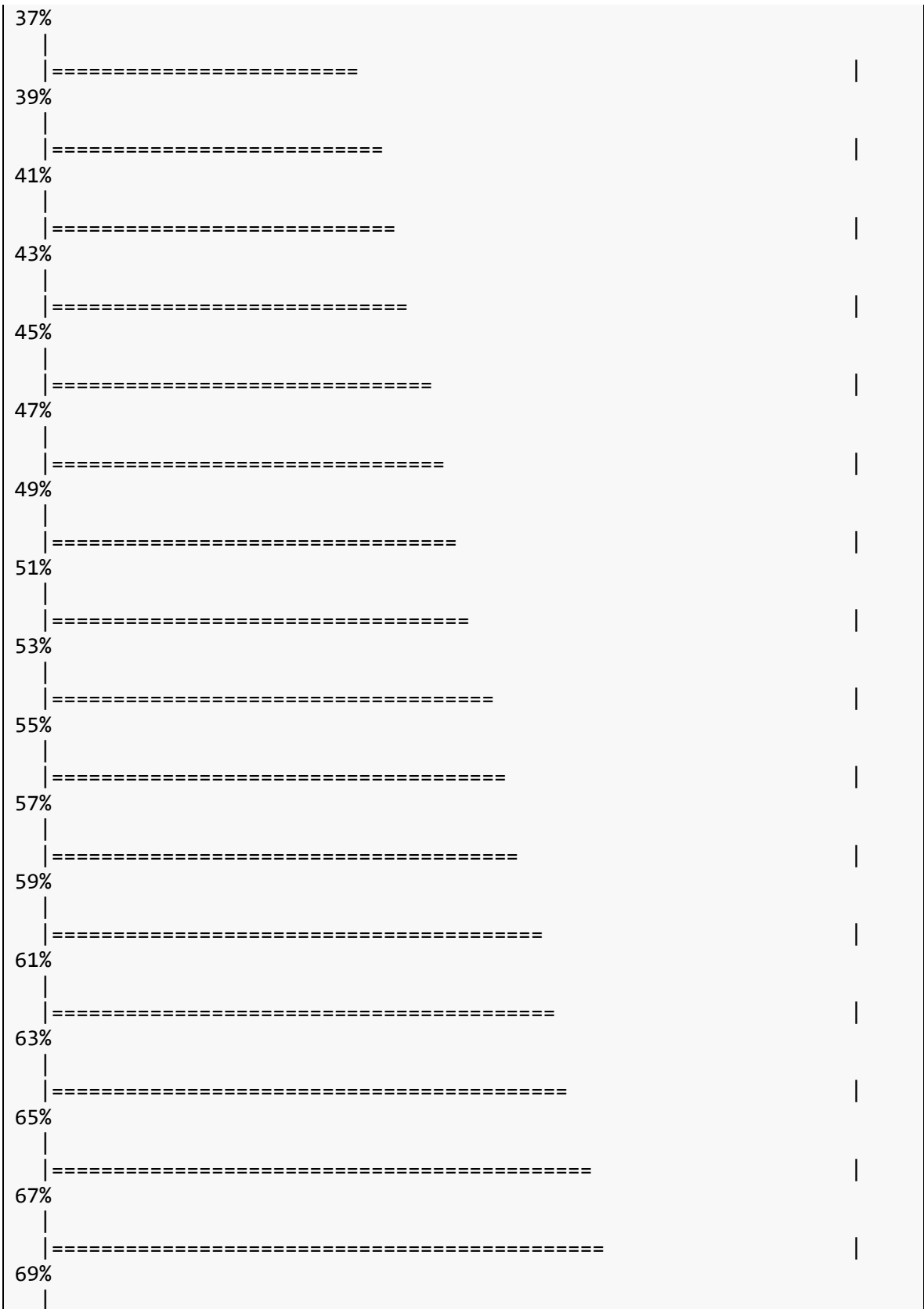
```
perm.anova(response~fact1/fact3, nest.f2="random", nperm=49)
```

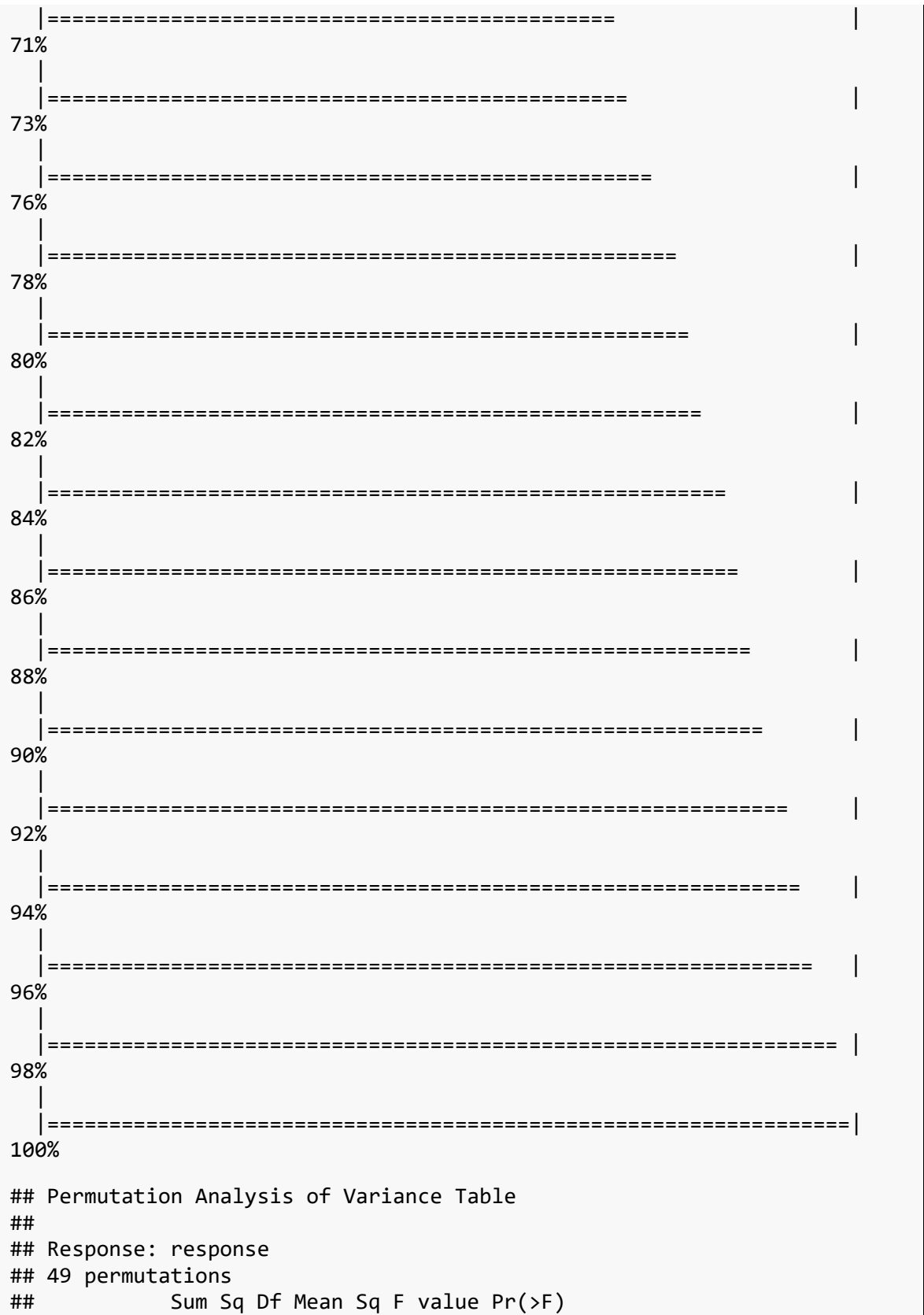
##

0%

2%

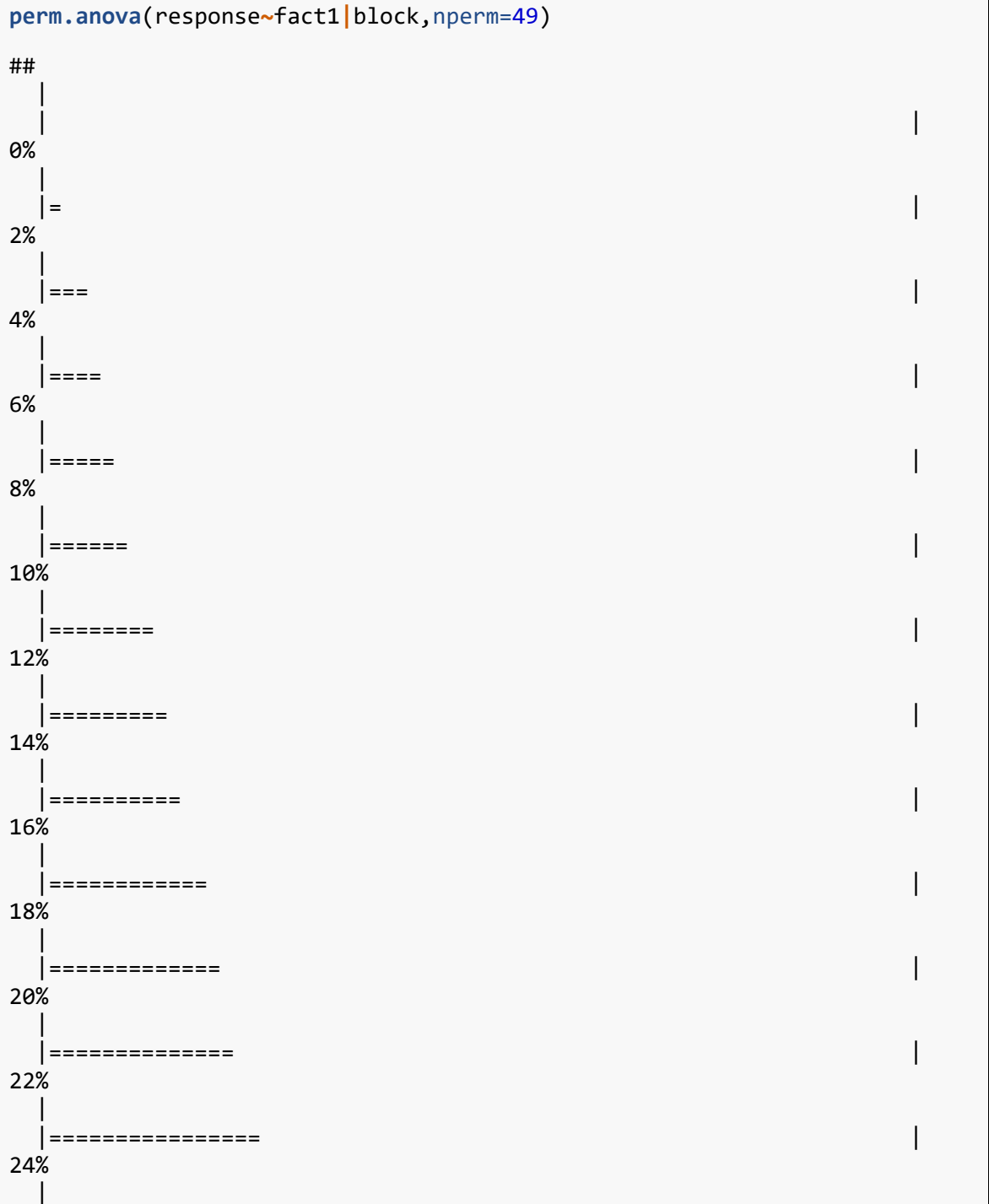


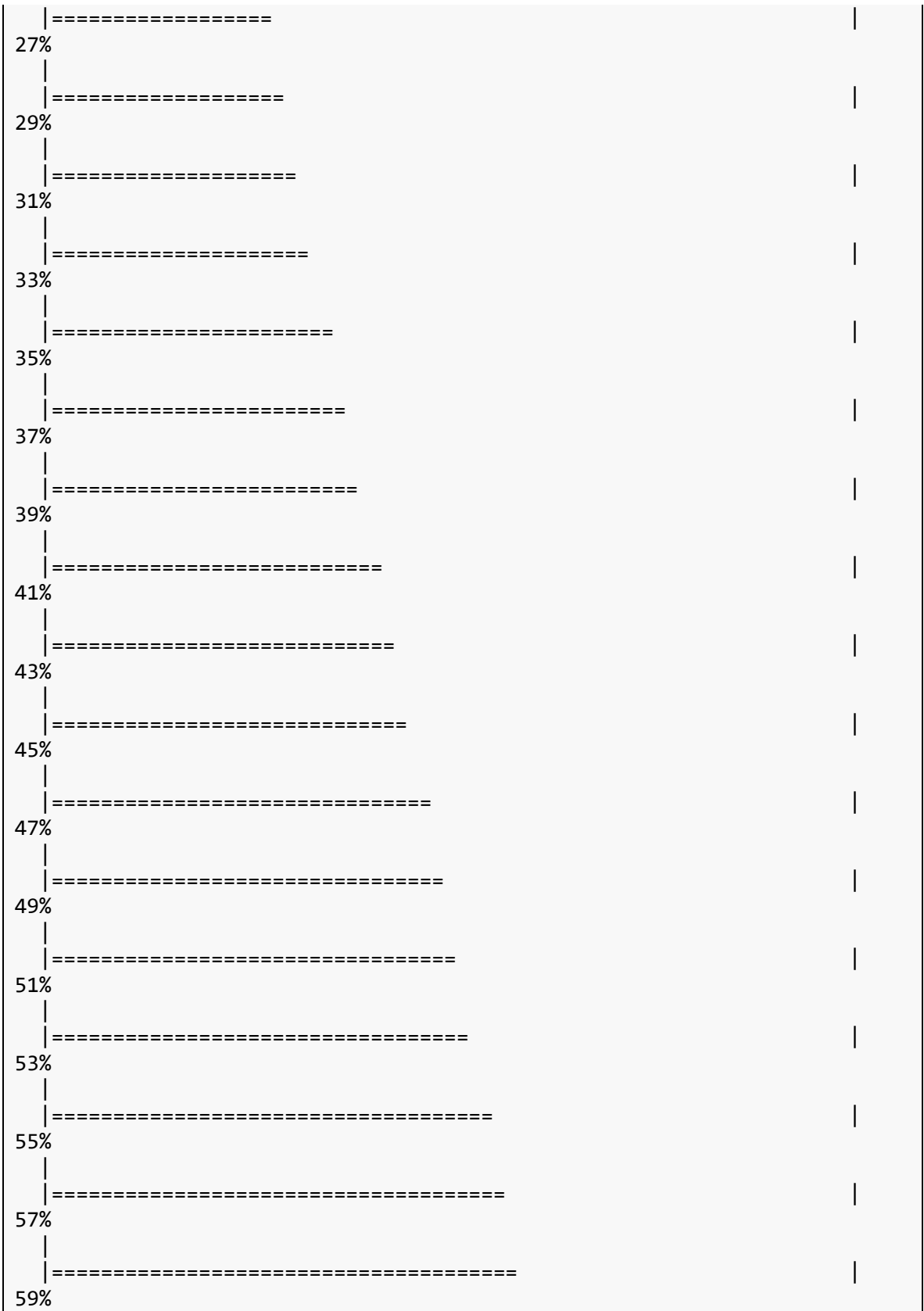


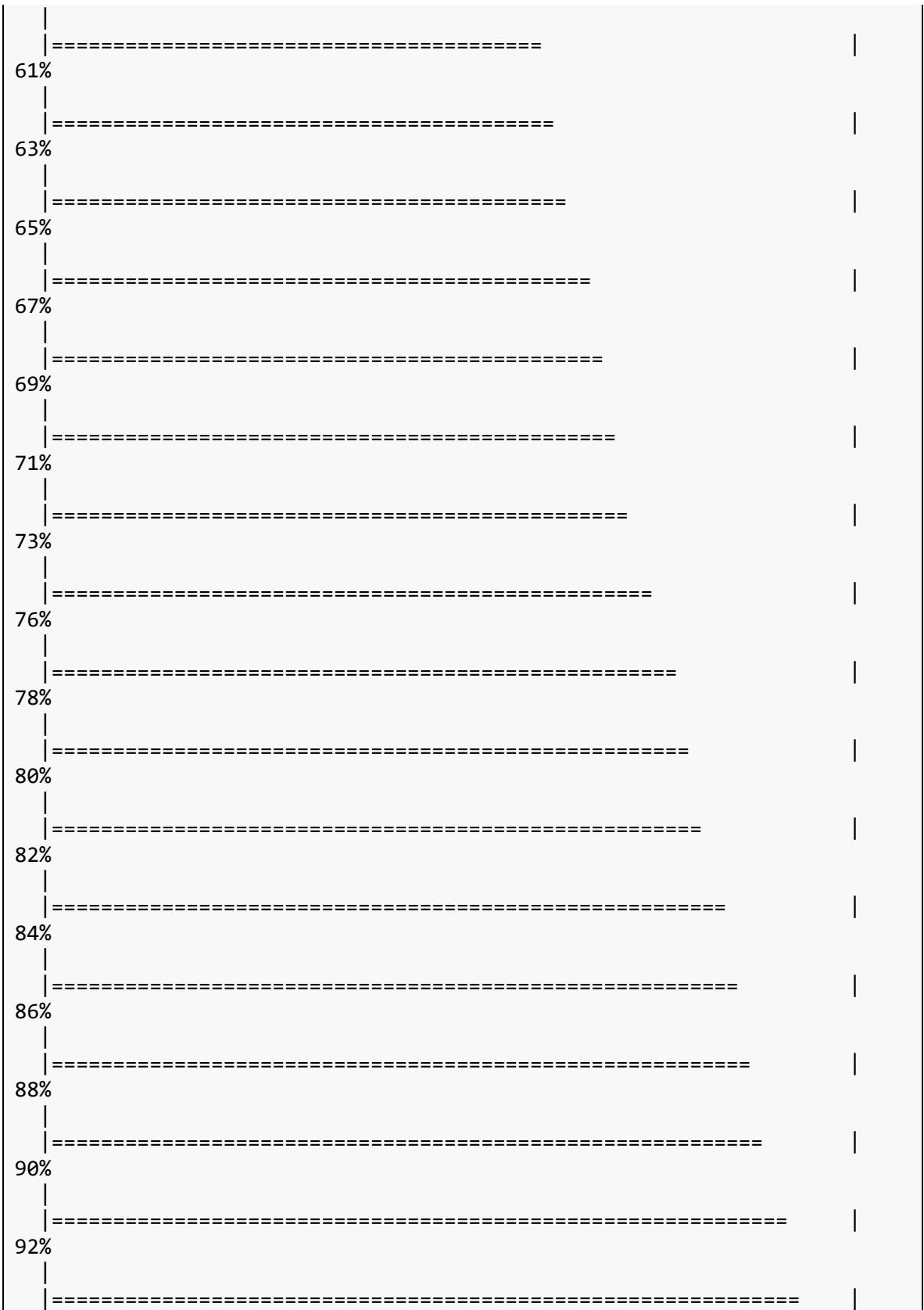


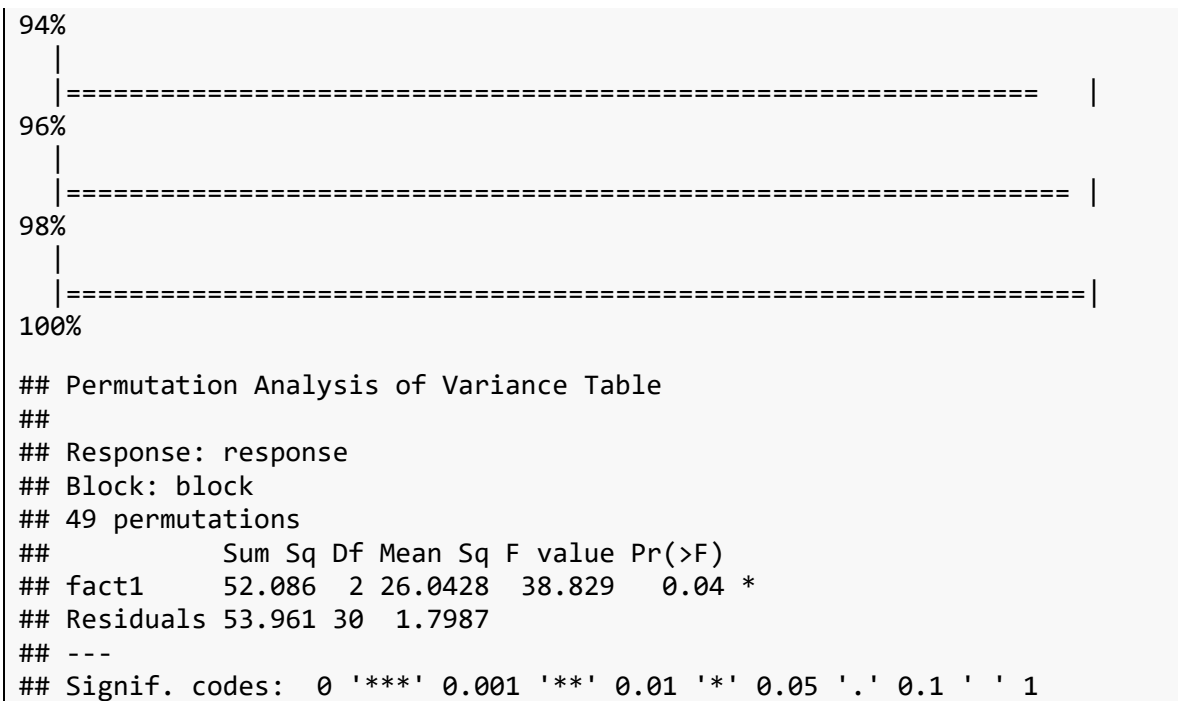
```
## fact1      52.086  2 26.0428  37.455   0.02  *
## Residuals 53.961 30  1.7987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1 fixed factor and 1 blocking (random) factor



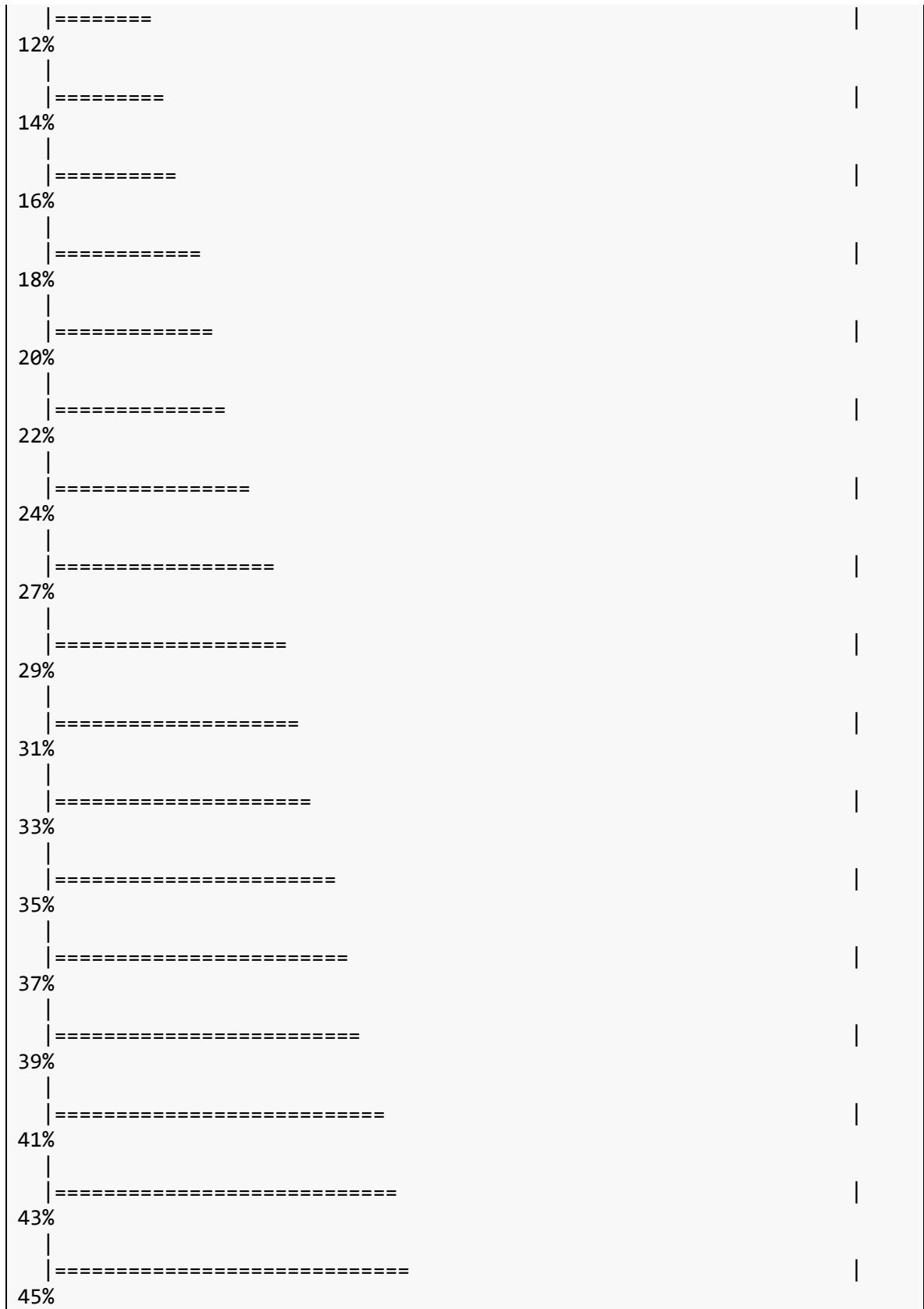


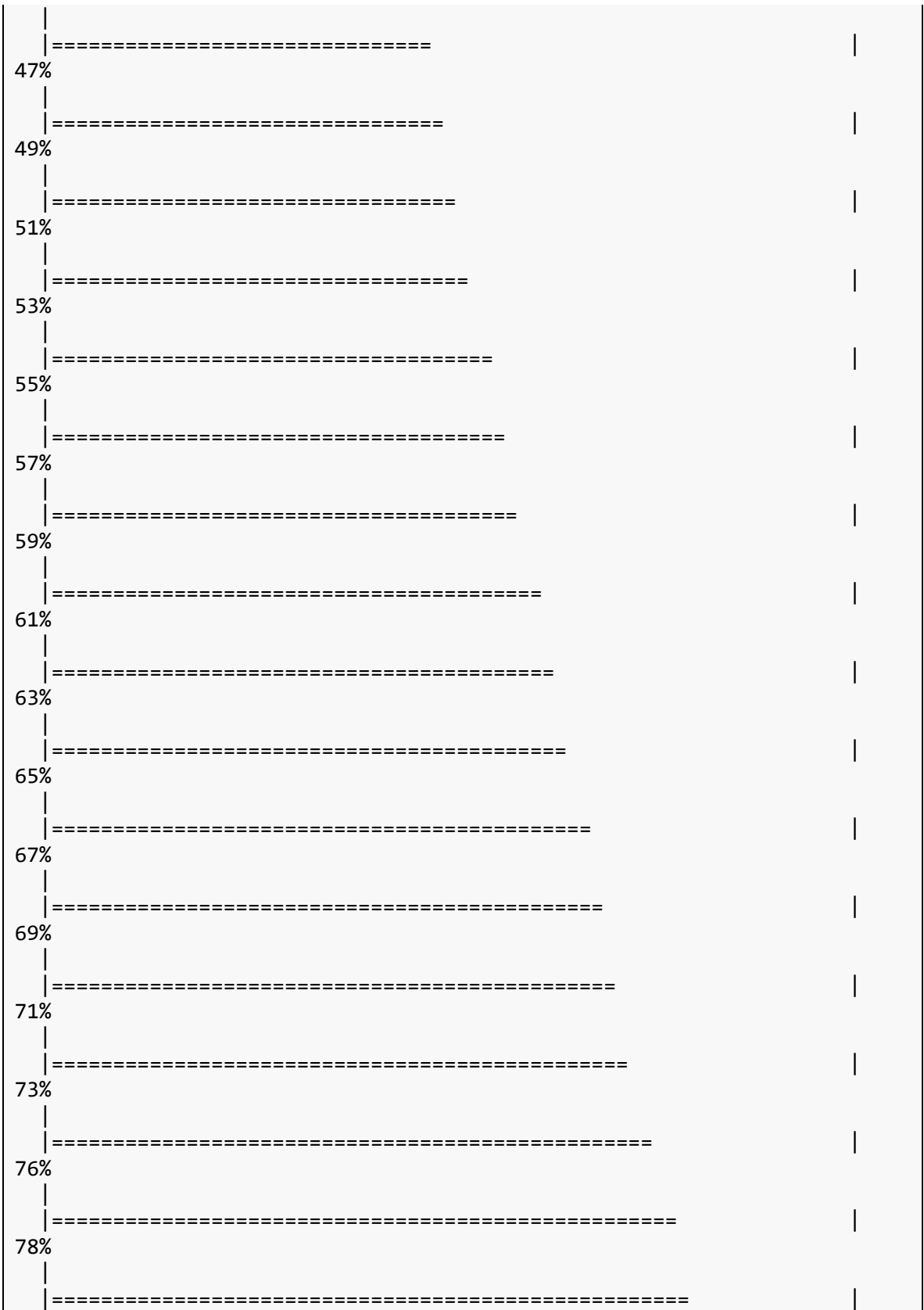


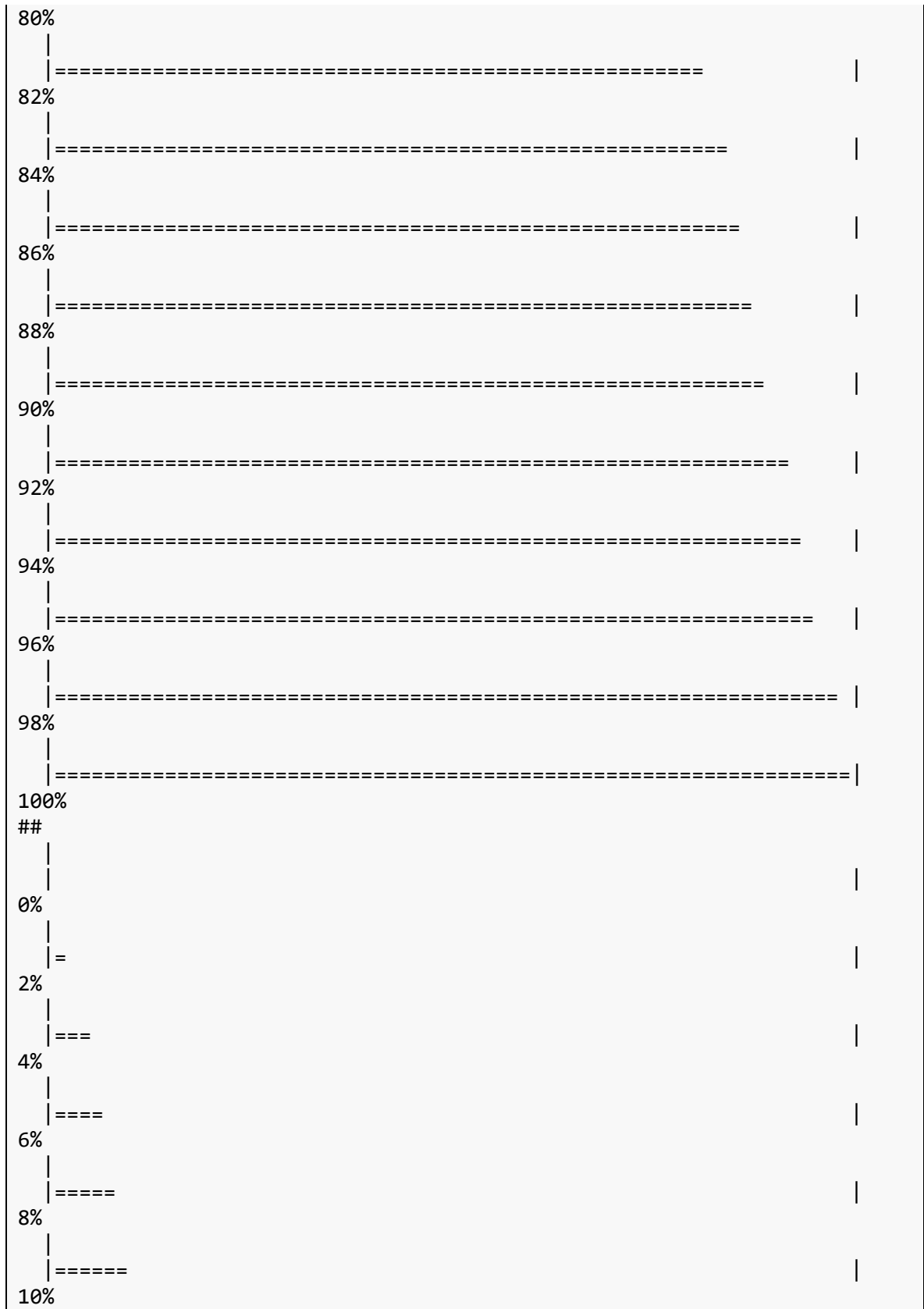


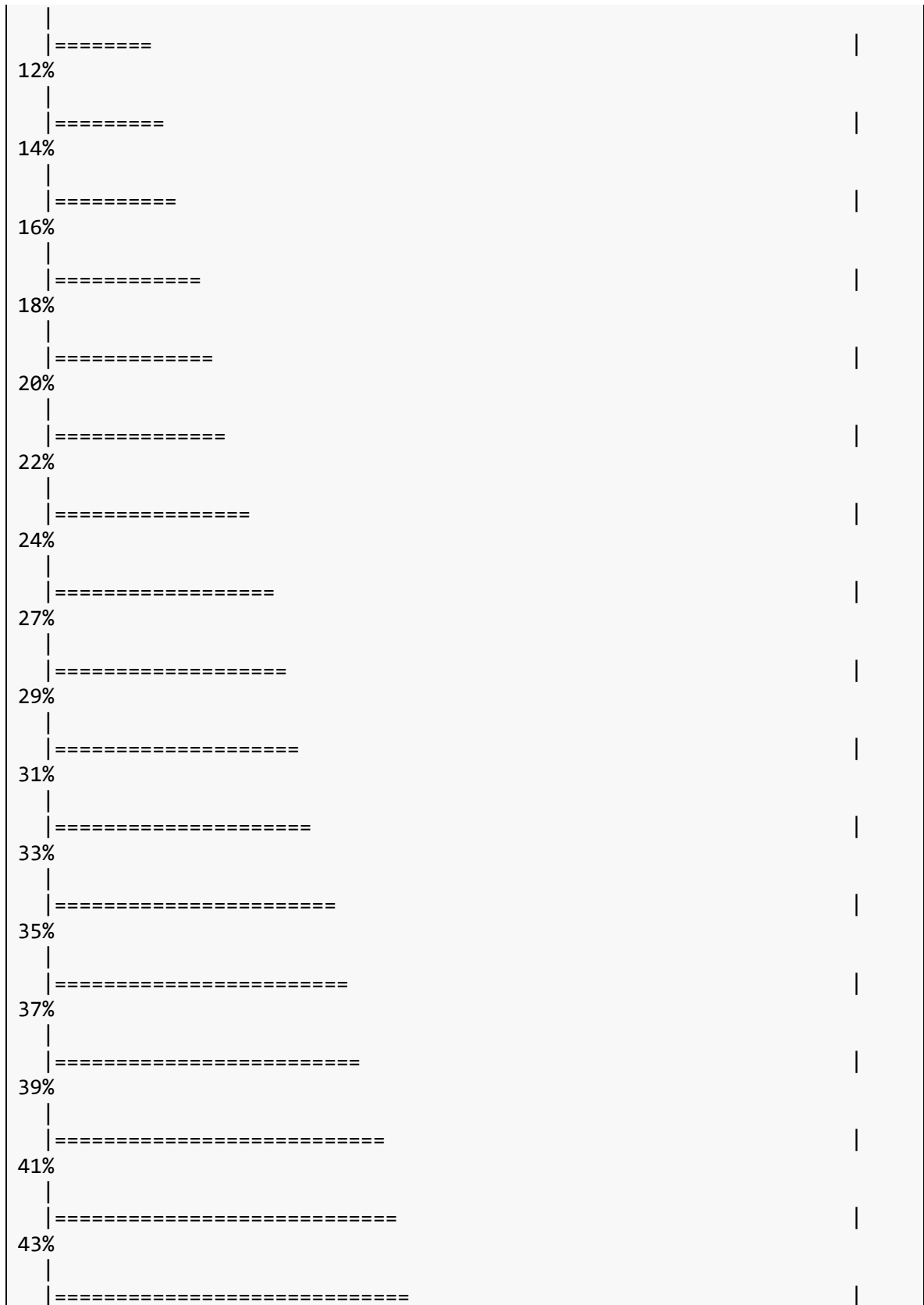
Este análisis ANOVA de permutaciones debería complementarse en el caso de $p < 0.05$ con el correspondiente análisis MCP de permutaciones utilizando la función `pairwise.perm.t.test()` que realiza una prueba t-Student de comparación de medias mediante permutaciones, y con una corrección del p-valor:

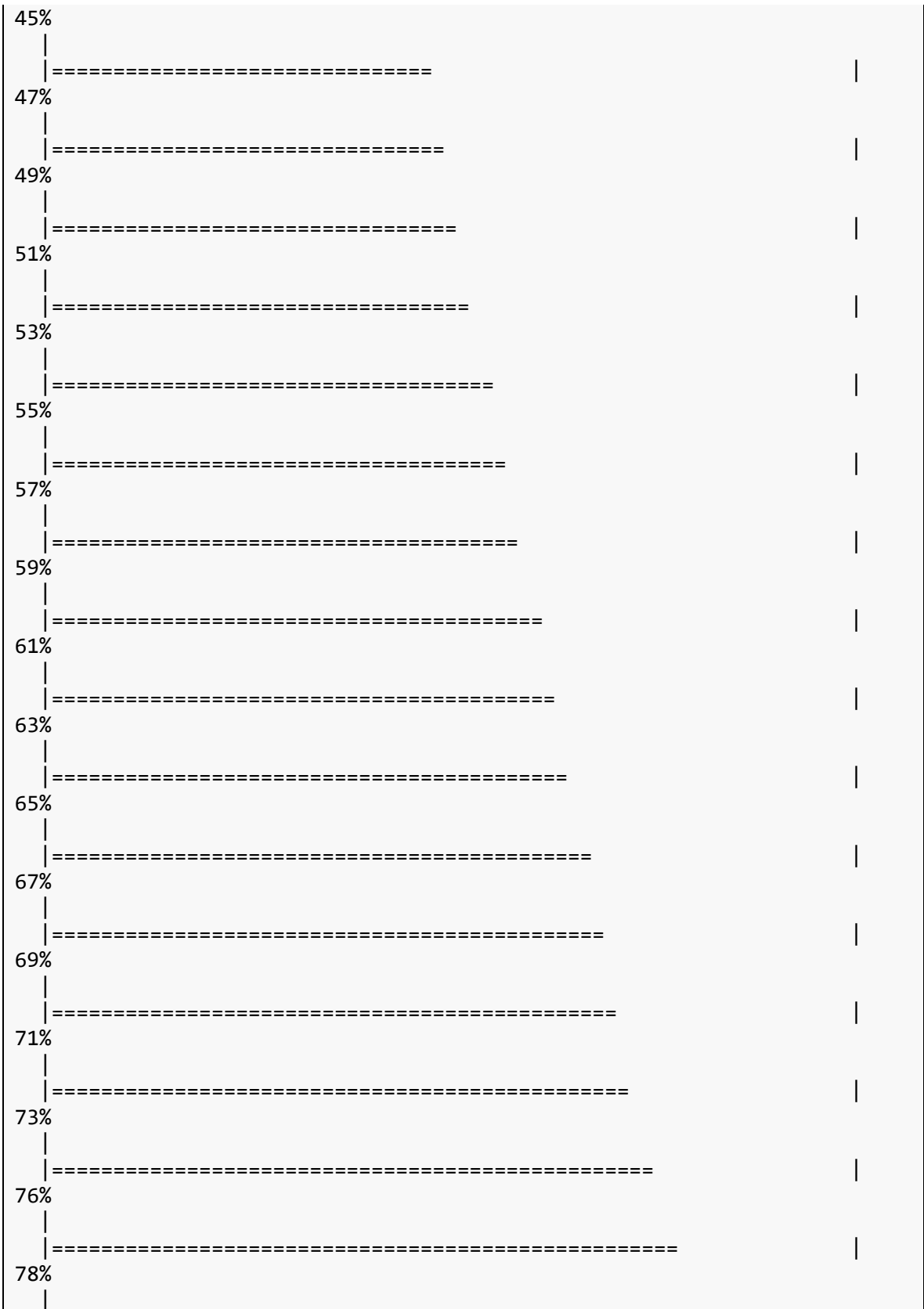


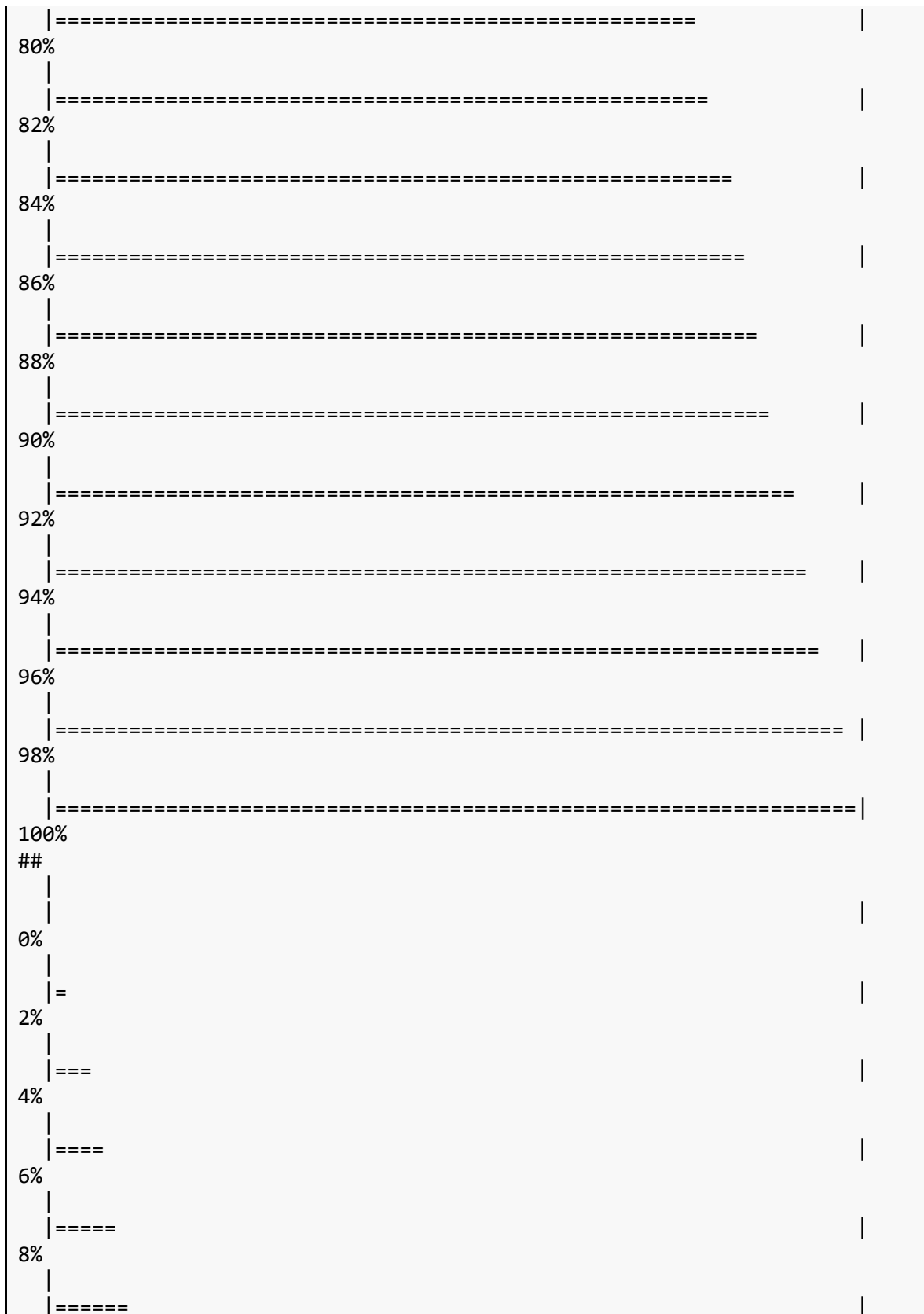


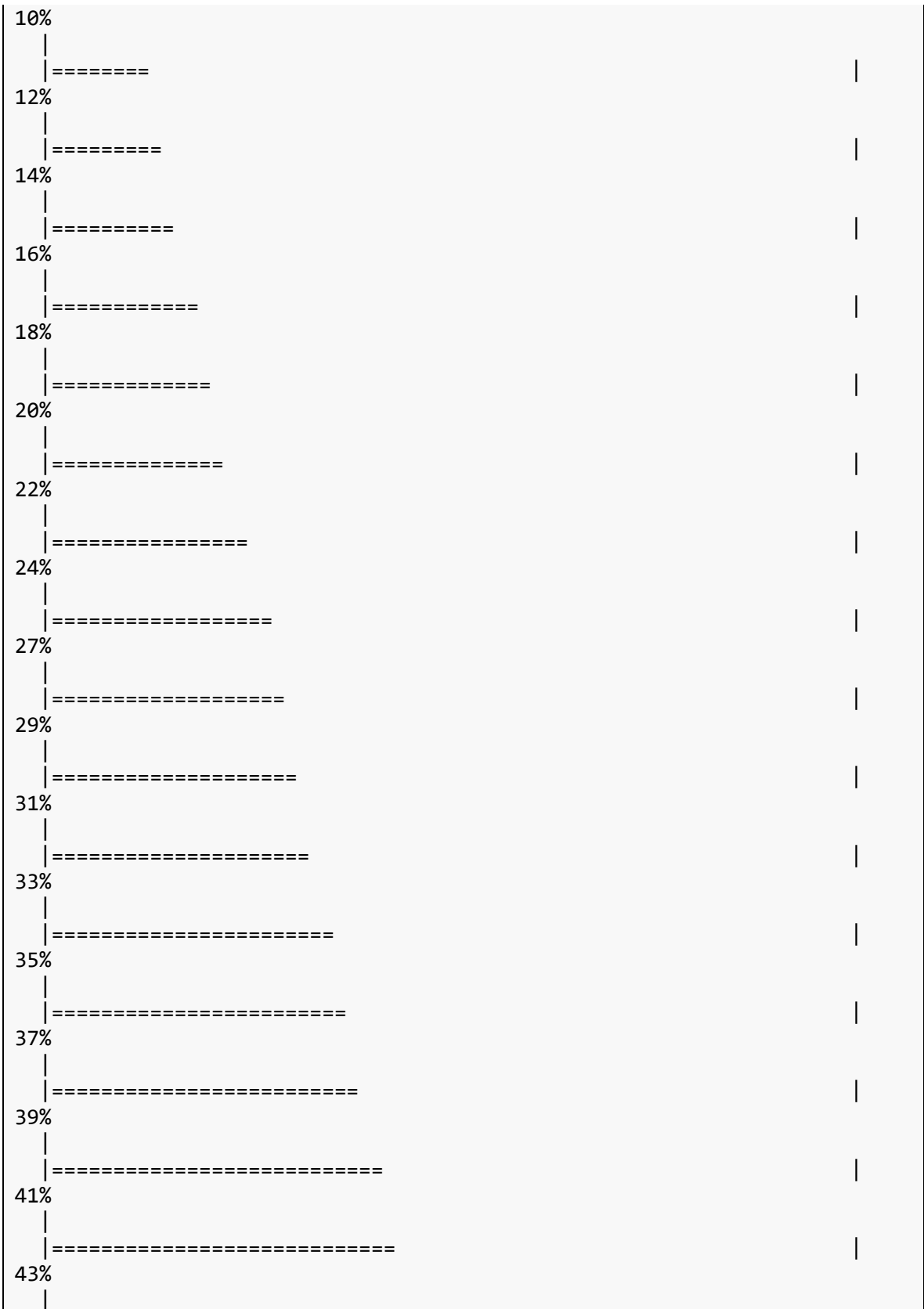


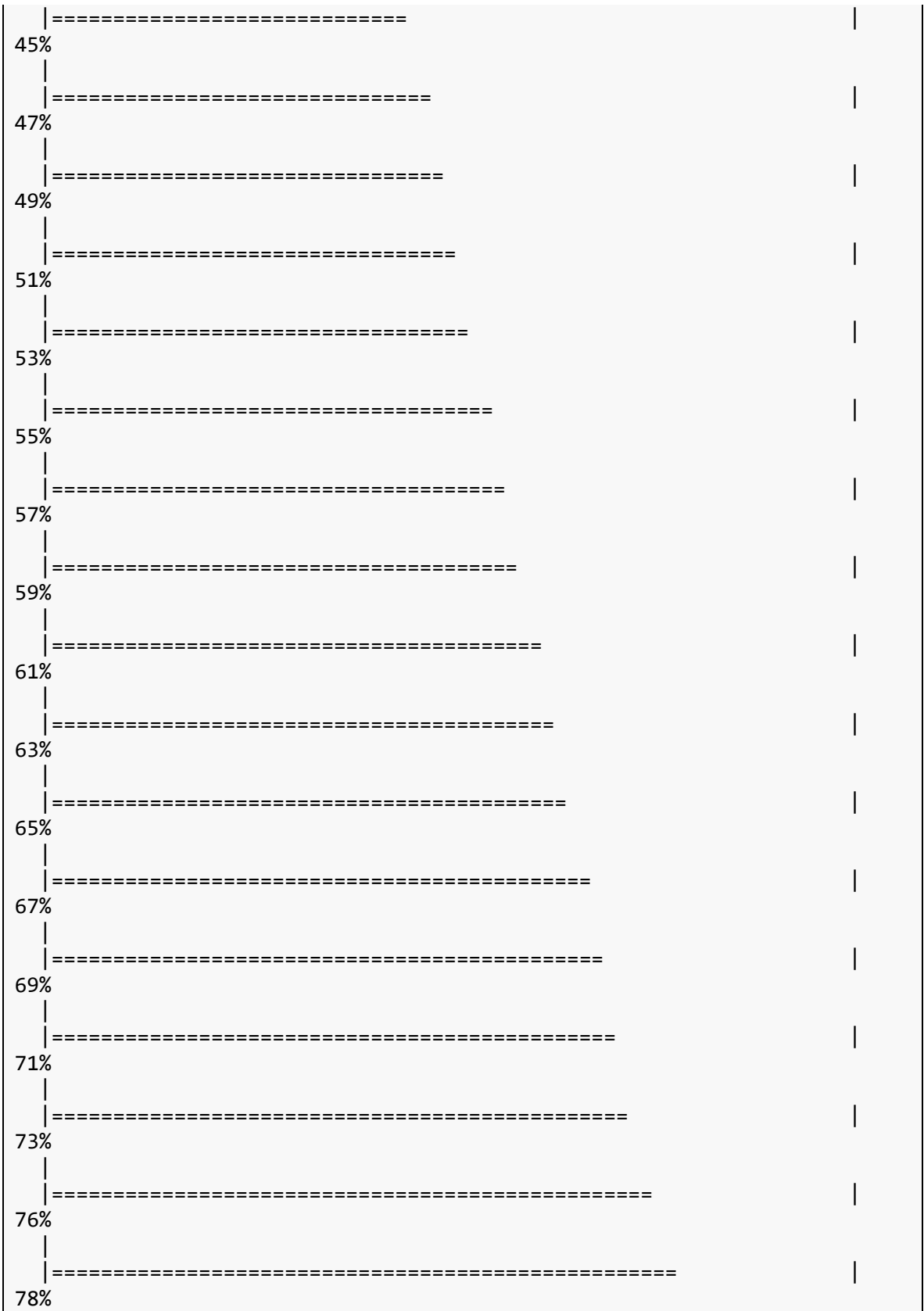


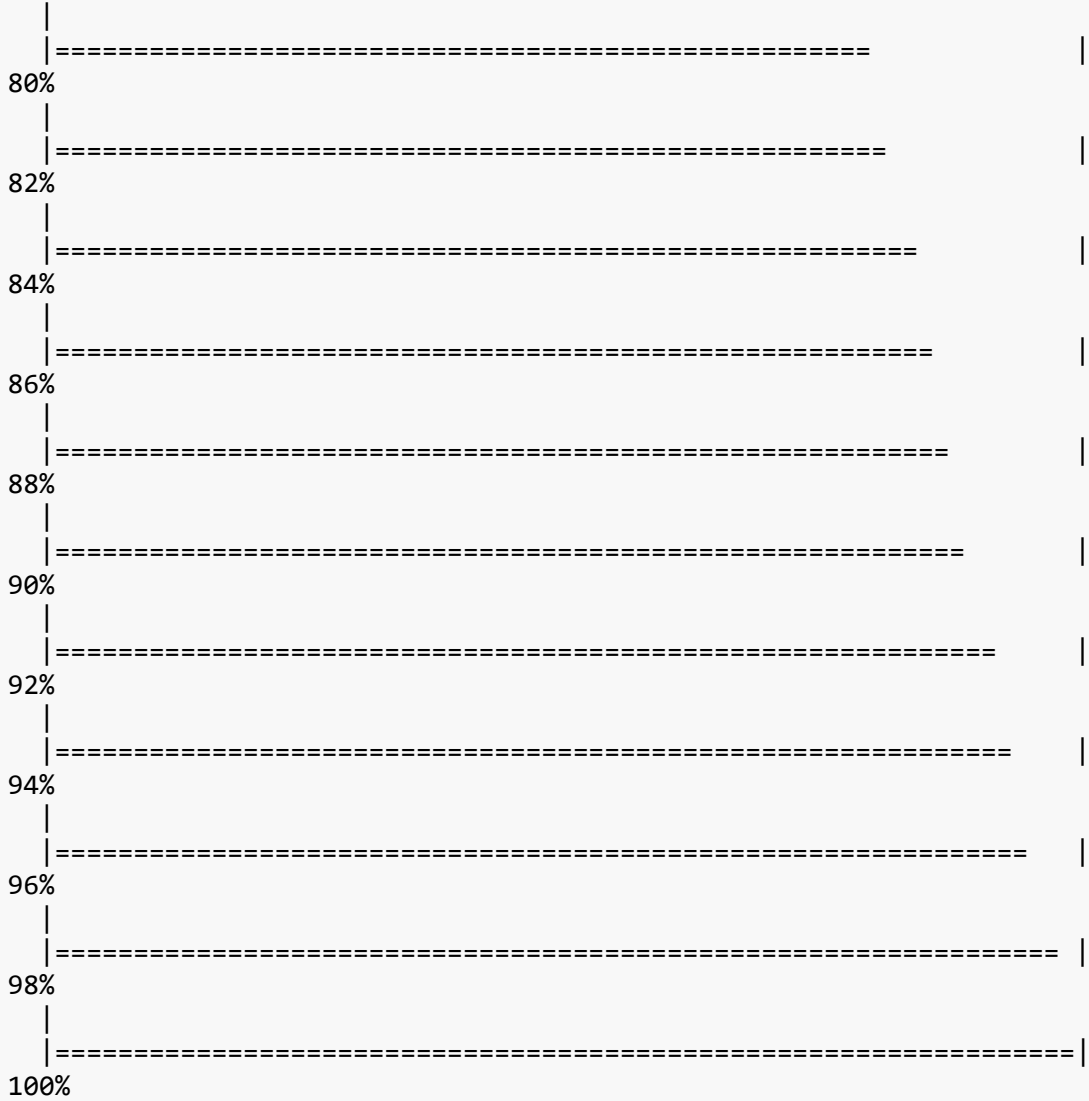












```
##
## Pairwise comparisons using permutation t tests
##
## data: response and fact1
## 49 permutations
##
## A B
## B 0.06 -
## C 0.12 0.06
##
## P value adjustment method: fdr
```

2.9.7. ANOVA multifactorial robusto: test de permutaciones función adonis() librería Vegan.

Otro test ANOVA con permutaciones es adonis, que se encuentra en la librería vegan. donde pueden encontrarse numerosos ejemplos como el que aquí se presenta con los datos denominados “dune”. Pueden utilizarse los comandos adonis y adonis.II:

Veamos un ejemplo de ANOVA de 2 factores con interacción procedente del ejemplo dune.

```
require(vegan)

## Loading required package: vegan
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.5-4

data(dune)
data(dune.env)

adonis(dune~Management*A1,data=dune.env,permutations=99)

##
## Call:
## adonis(formula = dune ~ Management * A1, data = dune.env, permutations
= 99)
##
## Permutation: free
## Number of permutations: 99
##
## Terms added sequentially (first to last)
##
##              Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## Management    3   1.4686 0.48953  3.2629 0.34161  0.01 **
## A1              1   0.4409 0.44089  2.9387 0.10256  0.03 *
## Management:A1  3   0.5892 0.19639  1.3090 0.13705  0.22
## Residuals     12   1.8004 0.15003           0.41878
## Total         19   4.2990                1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

adonis(dune~A1*Management,data=dune.env,permutations=99)

##
## Call:
## adonis(formula = dune ~ A1 * Management, data = dune.env, permutations
```



```

= 99)
##
## Permutation: free
## Number of permutations: 99
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## A1           1     0.7230 0.72295  4.8187 0.16817  0.01 **
## Management   3     1.1865 0.39551  2.6362 0.27600  0.03 *
## A1:Management 3     0.5892 0.19639  1.3090 0.13705  0.25
## Residuals   12     1.8004 0.15003           0.41878
## Total       19     4.2990           1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

adonis.II(dune~Management*A1,data=dune.env,permutations=99)

## Permutation: free
## Number of permutations: 99
##
## Type II tests
## Response: dune
##           Sum Sq Mean Sq Df      F Pr(>F)
## Management  1.1865 0.39551  3 2.6362  0.03 *
## A1           0.4409 0.44089  1 2.9387  0.03 *
## Management:A1 0.5892 0.19639  3 1.3090  0.22
## Residuals   1.8004 0.15003 12
## Total       4.2990           19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

adonis.II(dune~A1*Management,data=dune.env,permutations=99)

## Permutation: free
## Number of permutations: 99
##
## Type II tests
## Response: dune
##           Sum Sq Mean Sq Df      F Pr(>F)
## A1           0.4409 0.44089  1 2.9387  0.02 *
## Management   1.1865 0.39551  3 2.6362  0.02 *
## A1:Management 0.5892 0.19639  3 1.3090  0.21
## Residuals   1.8004 0.15003 12
## Total       4.2990           19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

El análisis ANOVA con el comando adonis se basa en permutaciones con cálculo de distancias.

2.9.10-Anova basado en Bootstrap (Aova multifactorial con diseños no balanceados)

Veamos el caso, bastante común, en diseños con varios factores donde nos encontramos factores fijos en diseños no balanceados (diferente cantidad de réplicas/individuos en las condiciones experimentales)

Esta función utiliza dos métodos como son el “residual Bootstrap” según lo descrito por Efron (1979) y el “wild Bootstrap” como lo describe Wu (1986) para la prueba de hipótesis ANOVA. Se permiten modelos lineales que incorporan variables predictoras categóricas y / o cuantitativas con una respuesta cuantitativa. La salida de la función crea la distribución nula de arranque para cada término que se va a probar. La estimación se realiza a través de mínimos cuadrados y solo se calcula la suma de cuadrados Tipo I y se da el p.valor bajo la hipótesis nula de diferencia de medias entre niveles del factor. Es muy interesante en el caso no balanceado ya que al realizar Bootstrap tiene en cuenta la incertidumbre causada por diseños no balanceados.

A continuación se presenta su uso con un ejemplo de la librería lmbot() con el conjunto de datos sobre motores y coches data(mtcars):

```
data(mtcars)           #Load an example dataset
library(lmbot)
myANOVA2 <- ANOVA.boot(mpg~as.factor(cyl)*as.factor(am), data=mtcars, B =
1000, type = "residual", wild.dist = "normal")
myANOVA2$p-values`    #bootstrap p-values for 2-way interactions model

## [1] 0.000 0.045 0.273 # p-value under null hypothesis (factor A, B,
A*B)
```

Los resultados obtenidos pueden visualizarse para cada iteración del bootstrap realizado y pueden verse a continuación, por ejemplo para los 50 primeros casos

```
#results are
myANOVA2 <- ANOVA.boot(mpg~as.factor(cyl)*as.factor(am), data=mtcars, B =
50, type = "residual", wild.dist = "normal")
myANOVA2 #50 p.values are represented

## $terms
## [1] "as.factor(cyl)           " "as.factor(am)           "
## [3] "as.factor(cyl):as.factor(am)" "Residuals             "
##
## $df
## [1]  2  1  2 26
##
## $origFStats
## [1] 44.851657  3.998759  1.383233
```

```

##
## $origSSE
## [1] 239.0592
##
## $origSSTr
## [1] 824.78459 36.76692 25.43651
##
## $bootFStats
##           [,1]           [,2]           [,3]
## [1,] 2.8579109645 0.5761706638 0.09819778
## [2,] 0.4065393187 2.1402871431 0.03264614
## [3,] 0.3102139568 0.5824949513 0.12465694
## [4,] 1.2692344043 1.8332530041 0.15810549
## [5,] 2.1162948998 0.5634056017 1.85413710
## [6,] 0.6874482124 0.3594568378 0.32959821
## [7,] 0.3306617538 0.2012075338 0.38690098
## [8,] 0.5365247807 0.0869359048 0.85938624
## [9,] 0.3077164161 2.1463464517 0.50555323
## [10,] 2.9122316015 0.0965686100 2.20846956
## [11,] 1.3378830802 2.4375361431 0.81442208
## [12,] 0.7297260802 2.8008492615 1.08634409
## [13,] 0.8132503193 1.2457687454 0.18526922
## [14,] 1.3157078480 2.4572828016 1.32496205
## [15,] 0.0664247273 0.0140108573 1.63865756
## [16,] 0.0326335858 0.3754723545 2.51937542
## [17,] 2.2554769171 0.0469476001 1.29993257
## [18,] 0.1781039562 0.5305872062 0.36687833
## [19,] 0.4220354499 0.9793275109 3.36524183
## [20,] 0.0852786387 0.1960433681 2.67346440
## [21,] 0.6862117838 1.9047010673 0.91888228
## [22,] 0.3486747738 0.7487833035 1.10202293
## [23,] 0.1786454497 2.2850558756 2.50211955
## [24,] 8.6870770462 0.4333736766 1.43118729
## [25,] 2.7867871523 0.8742515895 2.77793266
## [26,] 5.2039042369 1.1560464960 1.03996152
## [27,] 0.1700969364 5.7082431391 4.28077036
## [28,] 0.3450193191 2.4942001603 0.06748508
## [29,] 0.0518422078 1.0771684824 0.28866191
## [30,] 2.1691243347 0.0012515744 0.56949402
## [31,] 0.1163524833 1.5370216993 1.75862485
## [32,] 1.8378902524 0.1150866278 1.00737075
## [33,] 0.6588393639 0.4525928524 2.29078123
## [34,] 0.0209608384 0.5385135241 0.14942388
## [35,] 1.1265850201 3.0924620492 0.21655789
## [36,] 0.2101020215 0.0306569667 1.08280589
## [37,] 0.0008998878 0.2105556535 0.26485351
## [38,] 3.0321944314 1.6166348869 3.14456078
## [39,] 1.2564268111 0.5569948711 1.21323500
## [40,] 1.3779869104 0.1434760858 0.28728663
## [41,] 1.0474845628 0.0926556778 1.61131703

```

```
## [42,] 4.8703677171 0.0530248134 1.23746963
## [43,] 0.2509797945 0.0008335917 1.22838775
## [44,] 0.2173973395 0.6530125215 0.66415829
## [45,] 1.2424066212 6.7275342080 2.14404651
## [46,] 0.1103164133 0.2742436747 0.83147628
## [47,] 0.6978955078 2.4420256607 0.61544127
## [48,] 1.1297005853 2.9686791953 3.71235725
## [49,] 0.8203006993 0.0001916133 0.76124408
## [50,] 0.8226618909 1.7682655276 1.80301035
##
## $bootSSE
##          [,1]
## [1,] 305.55698
## [2,] 189.62474
## [3,] 172.59488
## [4,] 244.42471
## [5,] 125.87079
## [6,] 121.80078
## [7,] 155.94350
## [8,] 219.36154
## [9,] 188.06807
## [10,] 129.94198
## [11,] 239.55998
## [12,] 156.27124
## [13,] 165.95240
## [14,] 246.19906
## [15,] 257.42852
## [16,]  99.51745
## [17,] 185.52119
## [18,] 246.44826
## [19,] 245.44385
## [20,] 183.23317
## [21,] 177.35098
## [22,] 173.17039
## [23,]  79.91588
## [24,] 140.10834
## [25,] 156.58727
## [26,] 180.17409
## [27,] 190.23355
## [28,] 253.03469
## [29,] 162.86438
## [30,] 152.05331
## [31,] 113.15092
## [32,] 135.88565
## [33,] 246.33424
## [34,] 184.30467
## [35,] 185.46109
## [36,] 217.11850
## [37,] 216.54629
## [38,] 206.15416
```

```

## [39,] 252.83794
## [40,] 252.25128
## [41,] 178.09779
## [42,] 131.34106
## [43,] 197.76366
## [44,] 223.59381
## [45,] 122.09638
## [46,] 189.97508
## [47,] 151.82240
## [48,] 177.23148
## [49,] 232.22544
## [50,] 171.08458
##
## $bootSSTr
##           [,1]           [,2]           [,3]
## [1,] 57.49568066  4.847990109  2.3080783
## [2,]  8.38689057 10.758230986  0.4761935
## [3,] 10.18828885  4.943979481  1.6550116
## [4,] 35.13011036 12.516644573  2.9726838
## [5,] 33.19370231  5.548612903 17.9524386
## [6,] 13.44184340  4.043599392  3.0881015
## [7,]  8.04438470  2.565443694  4.6411301
## [8,] 11.31352463  0.647192638 14.5012531
## [9,]  4.61487373 15.802911849  7.3137245
## [10,] 61.94265693  0.813672935 22.0748398
## [11,] 26.19621807 18.920022727 15.0079181
## [12,]  9.07522617 15.200373080 13.0587949
## [13,] 18.01168145  8.515749334  2.3650670
## [14,] 25.72514583 35.078443031 25.0926466
## [15,]  1.38345426  0.121733532 32.4490141
## [16,]  0.79131888  3.865777082 19.2862938
## [17,] 47.37674776  0.434688371 18.5511563
## [18,]  3.89990291  3.779517884  6.9551172
## [19,]  8.94364207  6.080079280 63.5367620
## [20,]  2.26991677  1.985144788 37.6821046
## [21,]  6.54371755 14.986710852 12.5357437
## [22,]  6.58455884  9.600759082 14.6798265
## [23,]  3.85685935 15.876240401 15.3814681
## [24,] 90.49635853  3.137958315 15.4247140
## [25,] 50.90235772  6.518344038 33.4606843
## [26,] 101.71869202 4.676588168 14.4133940
## [27,]  3.45971887 36.167979632 62.6420121
## [28,]  8.47784935 24.852884105  1.3135435
## [29,]  0.73558609  7.978844366  3.6163648
## [30,] 37.82078561  0.014674097  6.6610348
## [31,]  2.65116933  9.683342244 15.3069242
## [32,] 42.74987415  0.683384543 10.5297867
## [33,]  9.19506305  4.460855374 43.4075276
## [34,]  0.40509458  7.062518475  2.1184245
## [35,] 23.94585187 22.885670550  3.0894663

```

```
## [36,] 5.84092916 0.310945372 18.0843990
## [37,] 0.01908256 2.107269785 4.4117727
## [38,] 39.62679293 12.184302329 49.8664823
## [39,] 28.62253946 4.795406150 23.5962949
## [40,] 23.43336858 1.315877131 5.5744940
## [41,] 19.00227161 0.914300953 22.0747689
## [42,] 61.50974357 0.311807991 12.5023514
## [43,] 5.38844278 0.011890665 18.6869580
## [44,] 4.81285568 8.177369601 11.4232066
## [45,] 36.91010656 54.414866993 20.1369475
## [46,] 2.20735705 2.011470552 12.1507519
## [47,] 15.62615394 24.678546995 7.1875207
## [48,] 15.24606013 17.620474409 50.6112742
## [49,] 12.21037797 0.001876187 13.5984803
## [50,] 12.98450268 11.661516170 23.7282514
##
## $`p-values`
## [1] 0.00 0.04 0.34
```

Para un solo factor

```
myANOVA1 <- ANOVA.boot(mpg~as.factor(cyl), data=mtcars)
myANOVA1$p-values #bootstrap p-values for 1-way model
## [1] 0 # p-value under null hypothesis (factor A)
```

Para dos factores, uno de ellos fijo y el otro de tipo bloque sin interacción:

```
myANOVA2a <- ANOVA.boot(mpg~as.factor(cyl)+as.factor(am), data=mtcars)
myANOVA2a$p-values #bootstrap p-values for 1-way additive model
## [1] 0.000 0.058 # p-value under null hypothesis (Factor A, B)
```

Ver en <https://rdrr.io/cran/lmboot/> y en <https://rdrr.io/cran/lmboot/>

El principal incentivo para utilizar el método ANOVA es determinar si las discrepancias entre los promedios de tratamiento son mayores de lo que podría esperarse razonablemente de la variación que se produce dentro de las clasificaciones de tratamiento.

ANOVA bloques aleatorizados

2.10.-Diseños con bloques aleatorizados

Es evidente que existen numerosos diseños, tantos como la imaginación humana y las necesidades de los experimentadores que diseñan los experimentos, es por ello que no pueden abarcarse todos en este texto.

Vamos a ver en este capítulo algunos diseños muy utilizados y no comentados todavía. Algunos de ellos están comentados en “Monleón-Getino T. 2017. Diseño y planificación de estudios científicos: Calidad de datos (data management) y principios de diseño experimental)”.

La aleatorización se utiliza en el diseño experimental para crear grupos de tratamiento homogéneos y reducir el sesgo (Monleón, 2017). Esta parte se ilustrará el enfoque de diseño completamente aleatorizado.

Ahora procedemos a estudiar el concepto de bloques aleatorizados. Los bloques aleatorios se introducen normalmente cuando hay cierta similitud entre los diferentes grupos de prueba. Al introducir el bloqueo podemos controlar mejor las fuentes de variabilidad que no nos interesan (que han sido bloqueadas).

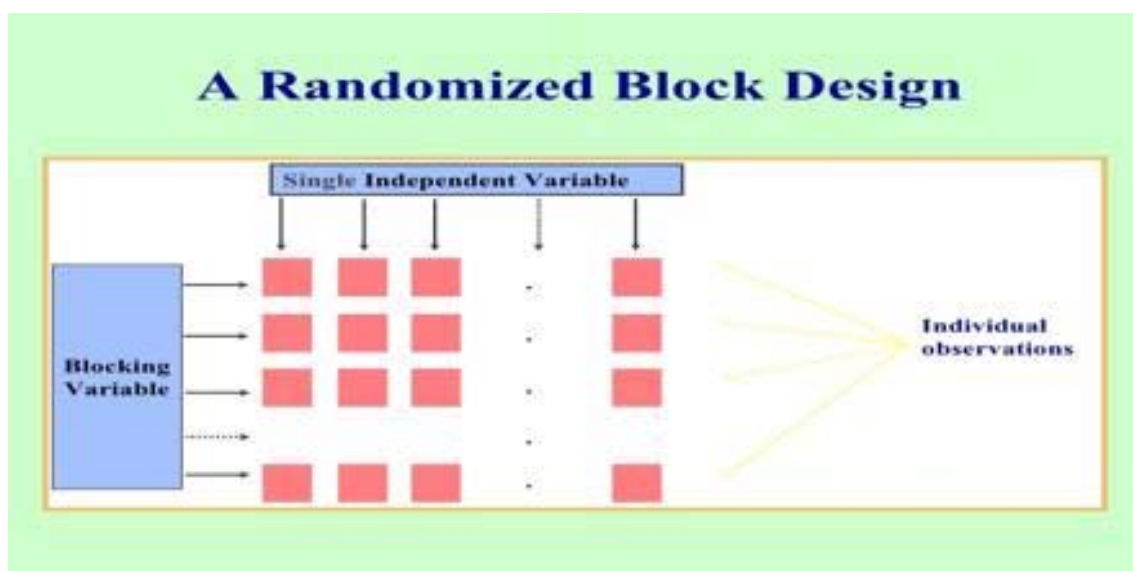


Imagen procedente de <http://www.assignmentpoint.com/wp-content/uploads/2016/07/Randomized-Block-Design.jpg>

Este tipo de diseño es muy utilizado para minimizar el efecto de la variabilidad cuando se asocia con unidades discretas (por ejemplo, localización, lote (batch), tiempo, operario, planta de producción, etc). “Normalmente distribuiremos aleatoriamente una réplica del experimento de cada combinación de tratamientos dentro de cada bloque. Por lo general, no hay un interés intrínseco en los bloques, y se considera que éstos son factores aleatorios. La suposición habitual es que el bloque por interacción de tratamiento es cero, y esta interacción pasa a ser el término de error para probar los efectos del tratamiento. Si designa la variable de bloqueo como Bloque, los términos en el modelo serían entonces Bloque, A, B y A*B. También especificaría el Bloque como el factor aleatorio.” (Ver en <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/supporting-topics/anova-models/what-are-randomized-block-designs-and-latin-square-designs/>)

2.10.1. Palillos en la alimentación

Vamos a presentar un ejemplo (Procedente de http://www.maths.lu.se/fileadmin/maths/matematisk_statistik/MASC05FMS072/F_r_s_ksplanering_lab_3new.pdf) donde se desea estudiar cómo la longitud de los palillos afecta la eficiencia de la alimentación. Estamos estudiando 6 longitudes diferentes de palillos, 180 mm, 210 mm, 240 mm, 270 mm, 300 mm, 330 mm en 31 sujetos diferentes (bloques). La eficiencia alimenticia se midió como la eficiencia de pellizco de los alimentos (el número de cacahuets recolectados y colocados en una taza).



Imagen procedente de https://images-na.ssl-images-amazon.com/images/I/71MZiDXZm8L_SX466_jpg

```
ChData<-read.table("chopstick2_rcb.dat")  
names(ChData)<-c("efficiency", "length group", "subject")  
attach(ChData)  
#download file chopstick2_rcb.dat from:  
http://www.maths.lu.se/fileadmin/maths/matematisk\_statistik/MASC05FMS072/chopstick2\_rcb.dat
```

A continuación, asignamos nuevas variables para las longitudes de palillos y el número de bloques de control.

```
f=c("L1", "L2", "L3", "L4", "L5","L6") #chopstick lengths  
k=6 #number of chopstick levels  
n=31 #number of control blocks
```

Se crea un vector de factores de tratamiento que corresponde a cada elemento en el vector de respuesta Y.

```
tm = gl(k, 1, n*k, factor(f)) #matching chopsticks tm
```

De manera similar, creamos un vector de factores de bloqueo para cada elemento en los datos de respuesta

```
blk = gl(n, k, k*n) # blocking factor
```

Ahora estamos listos para ejecutar ANOVA en los bloques aleatorizados

```
av = aov(ChData$efficiency ~ tm + blk)  
summary(av)  
detach(ChData)
```

```
> summary(av)
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|-----|--------|---------|---------|--------|
| tm | 5 | 54.1 | 10.82 | 0.694 | 0.629 |
| bl k | 30 | 626.8 | 20.89 | 1.340 | 0.130 |
| Residuals | 150 | 2338.1 | 15.59 | | |

¿Qué puede decir sobre el efecto de la longitud del palillo en la eficiencia de la alimentación?

El análisis anterior nos indica por una parte que no existe efecto del tamaño del palillo ($p=0.629$) y por otra parte el factor bloque “sujetos” no influye, por tanto, no convenía bloquear. Según algunos autores en estos casos convendría eliminar los bloques, pero esto no lo hemos sabido hasta analizar el experimento.

Rehacer el análisis anterior sin usar el bloqueo (seguir los pasos presentados en la sección anterior). Describir la diferencia entre el uso de bloques aleatorios y un resultado unidireccional completamente aleatorio.

Ver este link para más información:

<http://www.statmethods.net/stats/anova.html>

http://www.maths.lu.se/fileadmin/maths/matematisk_statistik/MASC05FMS072/F_rs_ksplanering_lab_3new.pdf

2.10.2. Diseño de probióticos en biotecnología

Con un diseño de bloques al azar, el experimentador divide a los sujetos en subgrupos llamados bloques, de modo que la variabilidad dentro de los bloques es menor que la variabilidad entre los bloques. Luego, los sujetos dentro de cada bloque se asignan aleatoriamente a las condiciones de tratamiento. En comparación con un diseño completamente aleatorizado, este diseño reduce la variabilidad dentro de las condiciones de tratamiento y la posibilidad de confusión, produciendo una mejor estimación de los efectos del tratamiento.

Based partilly in <http://www.r-tutor.com/elementary-statistics/analysis-variance/randomized-block-design>

En un diseño de bloques al azar, solo hay un factor principal bajo consideración en el experimento. Los sujetos de prueba similares se agrupan en bloques. Cada bloque se prueba contra todos los niveles de tratamiento del factor primario en orden aleatorio. Esto tiene la intención de eliminar la posible influencia por otros factores extraños que pueden ser desconocidos.

other interesting link <https://onlinecourses.science.psu.edu/stat502/node/176>

Un ejemplo que se presenta a continuación es del desarrollo de un probiótico a partir de bacterias.



Imagen procedente de

https://www.docprobiotics.net/uploads/1/0/3/8/103885584/show-animals-1-doc-lg_orig.jpg

Un producto biotecnológico obtenido de bacterias (probióticos) para tratar una enfermedad porcina contiene como principio activo, 3 nuevos probióticos (factor Probiotic). Para saber si tienen la misma precisión (Accuracy), se seleccionan al azar 6 granjas de cerdos para participar en el estudio. De acuerdo con el diseño de bloques al azar, cada granja se someterá a prueba de comercialización de los 3 probióticos nuevos. Además, una granja pondrá a prueba el mercado solo un probiótico por semana, y tardará 3 semanas en probar todos los probióticos. El orden de prueba de los probióticos para cada granja también se asigna al azar.

Problema: Supongamos que cada fila en la siguiente tabla representa la medida de precisión de los 3 probióticos en una granja de cerdos después de una semana de mercadotecnia de prueba. A un nivel de significación del .05, comprueba si la precisión (accuracy) media de los 3 probióticos nuevos es igual.

Solución La solución consiste en los siguientes pasos:

```
Accuracy <-c(31, 27, 24, 31, 28, 31, 45, 29, 46, 21, 18, 48 , 42, 36, 46,
32, 17, 40)
Probiotic <- c(1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3)
blk <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6)
experiment <- data.frame(Accuracy, Probiotic=as.factor(Probiotic),
```

```
blk=as.factor(blk))
View(experiment)
```

Se debe aplicar la función aov a una fórmula que describa la respuesta r por el factor de tratamiento Probiotic y el bloqueo de control de bloque (blk).

```
av = aov(Accuracy ~ Probiotic + blk, data=experiment)
summary(av)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## Probiotic   2  538.8   269.39   4.959 0.0319 *
## blk         5  559.8   111.96   2.061 0.1547
## Residuals  10  543.2    54.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que el valor p de 0.032 es menor que el nivel de significación de .05, rechazamos la hipótesis nula de que la precisión (accuracy) media de los nuevos probióticos es igual. El factor bloque no es significativo, lo que nos indica que no hay diferencias entre las medias de la precisión entre bloques y aunque así fuera se habría bloqueado al incorporar este factor.

2.10.3. Uso de cuadrados latinos en diseños de producción agrícola

El diseño del cuadrado latino se usa donde el investigador desea controlar la variación en un experimento que está relacionado con filas y columnas en el campo.

| | Position | | |
|--------|----------|---|---|
| Expt 1 | a | b | c |
| Expt 2 | b | c | a |
| Expt 3 | c | a | b |

Imagen de:

https://www.researchgate.net/profile/Lawrence_Sanford/publication/278399194/figure/fig2/AS:669061549674508@1536528112439/Experimental-design-of-a-latin-square-analysis-of-variance-adapted-after-Steel-Torrie.png

Tener en cuenta que:

- Los tratamientos se asignan al azar dentro de las filas y columnas, con cada tratamiento una vez por fila y una vez por columna.
- Hay un número igual de filas, columnas y tratamientos.
- Útil donde el experimentador desea controlar la variación en dos direcciones diferentes

#EJEMPLO EN: <http://www.r-bloggers.com/Latin-squares-design-in-r/>

#Veamos un ejemplo de fertilización en el campo donde la variable principal es frecuencia (como un tipo de producción) y donde se aleatorizarán mediante el uso de cuadrados latinos 3 fertilizantes y 3 tratamientos

```
fertil <- c(rep("fertil1",1), rep("fertil2",1), rep("fertil3",1),
rep("fertil4",1), rep("fertil5",1))
treat <- c(rep("treatA",5), rep("treatB",5), rep("treatC",5),
rep("treatD",5), rep("treatE",5))
seed <- c("A","E","C","B","D", "C","B","A","D","E", "B","C","D","E","A",
"D","A","E","C","B", "E","D","B","A","C")
freq <- c(42,45,41,56,47, 47,54,46,52,49, 55,52,57,49,45, 51,44,47,50,54,
44,50,48,43,46)
```

```
mydata <- data.frame(treat, fertil, seed, freq)
```

```
mydata
```

```
##      treat  fertil seed freq
## 1  treatA fertil1    A   42
## 2  treatA fertil2    E   45
## 3  treatA fertil3    C   41
## 4  treatA fertil4    B   56
## 5  treatA fertil5    D   47
## 6  treatB fertil1    C   47
## 7  treatB fertil2    B   54
## 8  treatB fertil3    A   46
## 9  treatB fertil4    D   52
## 10 treatB fertil5    E   49
## 11 treatC fertil1    B   55
## 12 treatC fertil2    C   52
## 13 treatC fertil3    D   57
```

```

## 14 treatC fertil4 E 49
## 15 treatC fertil5 A 45
## 16 treatD fertil1 D 51
## 17 treatD fertil2 A 44
## 18 treatD fertil3 E 47
## 19 treatD fertil4 C 50
## 20 treatD fertil5 B 54
## 21 treatE fertil1 E 44
## 22 treatE fertil2 D 50
## 23 treatE fertil3 B 48
## 24 treatE fertil4 A 43
## 25 treatE fertil5 C 46

matrix(mydata$seed, 5,5)

##      [,1] [,2] [,3] [,4] [,5]
## [1,] "A"  "C"  "B"  "D"  "E"
## [2,] "E"  "B"  "C"  "A"  "D"
## [3,] "C"  "A"  "D"  "E"  "B"
## [4,] "B"  "D"  "E"  "C"  "A"
## [5,] "D"  "E"  "A"  "B"  "C"

matrix(mydata$freq, 5,5)

##      [,1] [,2] [,3] [,4] [,5]
## [1,] 42  47  55  51  44
## [2,] 45  54  52  44  50
## [3,] 41  46  57  47  48
## [4,] 56  52  49  50  43
## [5,] 47  49  45  54  46

par(mfrow=c(2,2))
plot(freq ~ fertil+treat+seed, mydata)

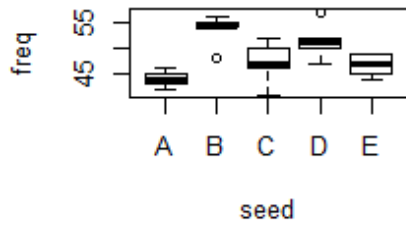
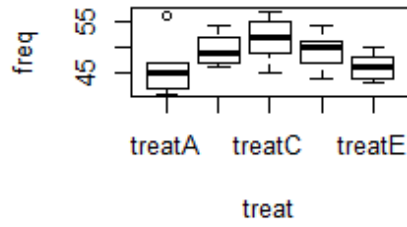
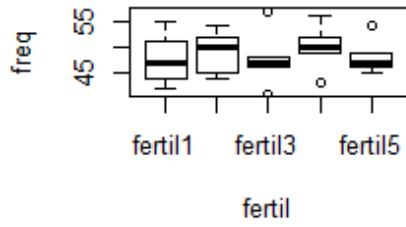
#forma de analizar el anova, donde los factores principales son
tratamiento y fertilizante
myfit <- lm(freq ~ fertil+treat+seed, mydata)
anova(myfit)

## Analysis of Variance Table
##
## Response: freq
##      Df Sum Sq Mean Sq F value    Pr(>F)
## fertil  4  17.76   4.440  0.7967 0.549839
## treat   4 109.36  27.340  4.9055 0.014105
## seed    4 286.16  71.540 12.8361 0.000271
## Residuals 12  66.88   5.573

#El resultado nos indica que existen diferencias estadísticamente
significativas entre las medias de producción de los tratamientos y que
en algunos de los casos se observan diferencias entre los bloques

```

utilizados (como era de esperar), pero ha quedado esta última aislada de los efectos principales.



ANOVA disceños jerárquicos (anidados)

2.11. Diseños jerárquicos o anidados con factores fijos y aleatorios

En algunas situaciones no se pueden combinar todos los niveles de un factor experimental con todos los niveles de otro, es decir, no se pueden determinar todos los posibles tratamientos que aparecen al cruzar los factores. Una buena introducción teórica al tema puede encontrarse en: <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Disenno/tema5de.pdf>

Este tema se tratará en profundidad en una futura ampliación de este texto.

Ver un ejemplo desarrollado en <https://www.r-bloggers.com/raccoon-ch-2-5-unbalanced-and-nested-anova/> Este ejemplo contiene datos recopilados de cinco tornos diferentes, cada uno de ellos utilizado por dos operadores diferentes. El objetivo del estudio es evaluar si existen diferencias significativas en el diámetro medio de los pernos entre tornos y / o operadores. Observe que aquí nos preocupa el efecto de los operadores, por lo que el diseño del experimento está anidado. Si nos preocupara el cambio en lugar del operador, el diseño habría sido al revés.



Torno industrial (Imagen de https://upload.wikimedia.org/wikipedia/commons/6/68/HwacheonCentreLathe_460x1000.jpg)

Veamos a continuación unos cuantos ejemplos de diseños jerárquicos, con dos o más factores fijos, mixtos o aleatorios.

2.11.1 Proveedores de nitrógeno y lotes (diseño mixto, con factores fijos y aleatorios)

En este apartado se presenta un ejemplo de diseño jerárquico donde se mide unos niveles diferenciales de nitrógeno de 3 proveedores (unidades inventadas que pueden ir de -5 a 5) y 4 lotes diferentes por proveedor. Se supondrá que el factor lote es un factor aleatorio ya que se coge de un gran conjunto de lotes que se producen y no tiene importancia en si, sino su variabilidad.

Lo que queremos contestar con este diseño experimental es si existen diferencias entre las medias de nitrógeno diferencial por proveedor y si el factor lote aporta o no variabilidad a la variable de respuesta



Imagen procedente de:

http://www.indura.cl/Content/Storage/CL/Contenido/Editor/Sitio_Vitivinicola/inertizacion_vitivinicola.jpg

Veamos cómo se cargan los datos y cómo se analiza el ejemplo mediante ANOVA:

```
op <- options(show.signif.stars=FALSE, contrasts=c("contr.sum",  
"contr.poly"))  
  
# Datos:  
nitrogen <- c(1, -2, -2, 1, -1, -3, 0, 4, 0, -4, 1, 0,
```

```

      1,0,-1,0,-2,4,0,3,-3,2,-2,2,
      2,-2,1,3,4,0,-1,2,0,2,2,1)
proveedor <- gl(3, 12 , labels=c("prov1","prov2","prov3"))
lot <- gl(4, 1, 36)

proveedor
## [1] prov1 prov1 prov1 prov1 prov1 prov1 prov1 prov1 prov1 prov1 prov1
## [12] prov1 prov2 prov2 prov2 prov2 prov2 prov2 prov2 prov2 prov2 prov2
## [23] prov2 prov2 prov3 prov3 prov3 prov3 prov3 prov3 prov3 prov3 prov3
## [34] prov3 prov3 prov3
## Levels: prov1 prov2 prov3

lot
## [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1
2 3
## [36] 4
## Levels: 1 2 3 4

# Modelo con 2 factores cruzados 1 fijo, 1 aleatorio, jerárquico

#Imaginemos que tratamos los factores como factores fijos:
fix.aov <- aov(nitrogen ~ proveedor + lot %in% proveedor)
anova(fix.aov)

## Analysis of Variance Table
##
## Response: nitrogen
##          Df Sum Sq Mean Sq F value Pr(>F)
## proveedor      2 15.056   7.5278   2.8526 0.07736
## proveedor:lot   9 69.917   7.7685   2.9439 0.01667
## Residuals     24 63.333   2.6389

# Forma equivalente, pero concisa
fix.aov <- aov(nitrogen ~ proveedor/lot)
anova(fix.aov)

## Analysis of Variance Table
##
## Response: nitrogen
##          Df Sum Sq Mean Sq F value Pr(>F)
## proveedor      2 15.056   7.5278   2.8526 0.07736
## proveedor:lot   9 69.917   7.7685   2.9439 0.01667
## Residuals     24 63.333   2.6389

# La tabla ANOVA anterior no es correcta, ya que el segundo factor es
aleatorio.

# Tabla ANOVA correcta dado que el primer factor es fijo y el otro
aleatorio
# Se puede obtener con la función 'anova2nest ()'

```

```

# Para disponer de la función 'anova2nest ()' es necesario que el archivo
"anova3nest.R" sea
# En el directorio de trabajo de R:
source("anova2nest.R")

anova2nest(fix.aov, random = c("lot"))

## Analysis of Variance Table
##
## Response: nitrogen
##           Df Sum Sq Mean Sq F value Pr(>F)
## proveedor    2 15.056   7.5278   0.9690 0.41578
## proveedor:lot  9 69.917   7.7685   2.9439 0.01667
## Residuals    24 63.333   2.6389

# No hay diferencias entre proveedores, pero el efecto lote es
significativo e indica que aporta varianza a la variable de respuesta Y
(nitrógeno diferencial)

# Sólo valen los parámetros del factor fijo (efectos del factor). Veamos
cómo se estiman estos efectos:
coef(fix.aov)

##           (Intercept)           proveedor1           proveedor2
##           0.36111111          -0.77777778          -0.02777778
## proveedorprov1:lot1 proveedorprov2:lot1 proveedorprov3:lot1
##           0.41666667          -1.66666667           0.83333333
## proveedorprov1:lot2 proveedorprov2:lot2 proveedorprov3:lot2
##           -2.58333333           1.66666667          -1.16666667
## proveedorprov1:lot3 proveedorprov2:lot3 proveedorprov3:lot3
##           0.08333333          -1.33333333          -0.50000000

# También :

dummy.coef(fix.aov)

## Full coefficients are
##
## (Intercept):           0.3611111
## proveedor:           prov1           prov2           prov3
##           -0.77777778  -0.02777778  0.80555556
## proveedor:lot:           prov1:1           prov2:1           prov3:1           prov1:2
##           0.41666667  -1.66666667  0.83333333  -2.58333333
##
## (Intercept):
## proveedor:
##
## proveedor:lot:           prov2:2           prov3:2           prov1:3           prov2:3
##           1.66666667  -1.16666667  0.08333333  -1.33333333

```

```
##
## (Intercept):
## proveedor:
##
## proveedor:lot:      prov3:3      prov1:4      prov2:4      prov3:4
##                    -0.50000000  2.08333333  1.33333333  0.83333333
```

0 model.tables()

```
model.tables(fix.aov)
```

```
## Tables of effects
##
## proveedor
## proveedor
##   prov1  prov2  prov3
## -0.7778 -0.0278  0.8056
##
## proveedor:lot
##       lot
## proveedor 1      2      3      4
##   prov1  0.4167 -2.5833  0.0833  2.0833
##   prov2 -1.6667  1.6667 -1.3333  1.3333
##   prov3  0.8333 -1.1667 -0.5000  0.8333
```

También pueden estimarse las medias para cada nivel del factor principal:

```
model.tables(fix.aov, type = "means")
```

```
## Tables of means
## Grand mean
##
## 0.3611111
##
## proveedor
## proveedor
##   prov1  prov2  prov3
## -0.4167  0.3333  1.1667
##
## proveedor:lot
##       lot
## proveedor 1      2      3      4
##   prov1  0.0000 -3.0000 -0.3333  1.6667
##   prov2 -1.3333  2.0000 -1.0000  1.6667
##   prov3  2.0000  0.0000  0.6667  2.0000
```

Una parte muy interesante de la estimación de parámetros de este modelo es la varianza de la variable de respuesta, que en este caso será la varianza residual (within o MSerror) y la asociada al factor lote, un

factor aleatorio. Cada una de estas varianzas de senomina: Componentes de la varianza.

#Veamos cómo se estima en este modelo:

```
anovaTab <- anova(fix.aov)
anovaTab

## Analysis of Variance Table
##
## Response: nitrogen
##           Df Sum Sq Mean Sq F value Pr(>F)
## proveedor    2  15.056   7.5278   2.8526 0.07736
## proveedor:lot  9  69.917   7.7685   2.9439 0.01667
## Residuals    24  63.333   2.6389

ms.proveedor <- anovaTab[1,3]
ms.lot <- anovaTab[2,3]
mse <- anovaTab[3,3]

# gLL.proveedor <- anovaTab[1,1]
# gLL.lot <- anovaTab[2,1]
# gLL.error <- anovaTab[3,1]

# Componentes de La varianza
var.lot <- (ms.lot - mse) / 3
var.lot # Componente de La varianza asociado a Lote:

## [1] 1.709877

var.error <- mse
var.error # Componente de La varianza asociado a error:

## [1] 2.638889

var.total <- var.lot + var.error # Componente de La varianza total
round(100 * var.lot / var.total, 2) # Componente de La varianza de Lote %

## [1] 39.32

# En forma de porcentajes:
round(100 * var.error / var.total, 2) # Componente de La varianza
residual %

## [1] 60.68
```

2.11.2 Medida de la salinidad en pozos de diversas zonas agrícolas

Veamos otro ejemplo de modelo lineal y su resolución ANOVA, donde se mezclan factores fijos y aleatorios, así como un factor jerarquizado y/o cruzado a otro. Este ejemplo es uno de los ejemplos clásicos que se explican en la asignatura de “Diseño de experimentos y análisis de datos” en la Facultad de Biología de la UB.



Imagen de: https://comercialhidraulica.webnode.mx/_files/200000109-dbe8adce53/Pozo%20de%20Agua.jpg

```
# EJEMPLO DE ANOVA - FACTORES FIJOS Y ALEATORIOS, INTERACCIONES Y  
JERÁRQUICOS  
# Ejemplo medida de la salinidad en unos pozos en diferentes horas  
#####  
#####  
#Toni MonLeón. DEAD 2011-2012  
  
op <- options(show.signif.stars=T, contrasts=c("contr.sum",  
"contr.poly"))  
  
#Leer Los datos del ejemplo  
pous <- read.table("pous.txt", header=T)  
fix(pous) #ver Los datos y ver que son correctas  
  
#especificar factores  
class(pous$Hora)
```



```
## [1] "factor"
```

```
class(pous$Pou)
```

```
## [1] "factor"
```

```
#) Imaginemos que vamos a medir la salinidad de unos pozos en dos  
horarios diferentes, así hacemos medidas de los mismos pozos (tomamos más  
de una medida: réplicas) dos veces cada día, así tendremos un modelo  
lineal de dos factores mixtos cruzados.  
#) Primer apartado a resolver: Modelo con 2 factores cruzados: 1 fijo  
(factor horas) y 1 aleatorio (pozo)
```

```
#####  
##
```

```
source("anova2cross.R")
```

```
#Vamos a imaginar que los factores son tratados como factores fijos de  
manera incorrecta : Model ANOVA de 2 Factors fijos y cruzados  
pous.aov <- aov(Salinitat ~ Hora*Pou, data=pous)  
anova(pous.aov) #Existen diferencias estadísticas entre pozos (p-valor  
= 0.003325) y entre horas (3.334e-10 ***), pero no en la interacción  
(0.086948 . )
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: Salinitat
```

```
##           Df Sum Sq Mean Sq  F value    Pr(>F)  
## Hora      1 20.7770 20.7770 5866.4494 3.334e-10 ***  
## Pou       2  0.1211  0.0606  17.0988  0.003325 **  
## Hora:Pou  2  0.0267  0.0134   3.7718  0.086948 .  
## Residuals 6  0.0212  0.0035
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# La tabla ANOVA anterior no es correcta !!!!!!!!!!!!!!!!.
```

```
# Vamos a ver entonces como calcular correctamente los p-valores:
```

```
# Modelo ANOVA de 2 Factores mixtos (1 fijo y 1 aleatorio), cruzados  
# 'Pou' aleatorio, 'Hora' fijo. Existen 2 posibilidades para el cálculo  
de los F y p-valores: 1º modelo restringido (por defecto. Es un modelo  
problemático si no hay balanceo. Ver Montgomery, Diseño de experimentos).  
Para el cálculo vamos a utilizar la función anova2cross()
```

```
anova2cross(pous.aov, random = "Pou")
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: Salinitat
```

```
##           Df Sum Sq Mean Sq  F value    Pr(>F)
```

```

## Hora      1 20.7770 20.7770 1555.3593 0.0006423 ***
## Pou       2  0.1211  0.0606   17.0988 0.0033255 **
## Hora:Pou  2  0.0267  0.0134    3.7718 0.0869477 .
## Residuals 6  0.0212  0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# equivalente a:
anova2cross(pous.aov, random = "Pou", restricted = TRUE) # Se detectan
diferencias entre horas (p-valor =0.0006423) y variabilidad significativa
de los pozos (p-valor = 0.0033255)

## Analysis of Variance Table
##
## Response: Salinitat
##          Df Sum Sq Mean Sq  F value    Pr(>F)
## Hora      1 20.7770 20.7770 1555.3593 0.0006423 ***
## Pou       2  0.1211  0.0606   17.0988 0.0033255 **
## Hora:Pou  2  0.0267  0.0134    3.7718 0.0869477 .
## Residuals 6  0.0212  0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Otra posibilidad es 'Pou' aleatorio, 'Hora' fijo, modelo no restringido
: no sale por defecto en la función anova2cross() y hay que solicitarlo
específicamente.
anova2cross(pous.aov, random = "Pou", restricted = FALSE) # Diferencias
entre horas (p-valor =0.0006423)

## Analysis of Variance Table
##
## Response: Salinitat
##          Df Sum Sq Mean Sq  F value    Pr(>F)
## Hora      1 20.7770 20.7770 1555.3593 0.0006423 ***
## Pou       2  0.1211  0.0606    4.5334 0.1807215
## Hora:Pou  2  0.0267  0.0134    3.7718 0.0869477 .
## Residuals 6  0.0212  0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#¿Qué es más correcto? Restringido O NO restringido. Si tenemos en cuenta
el efecto sobre las interacciones, tenemos que pensar que es más correcto
el restringido. En caso contrario es preferible al otro. El modelo
restringido es problemático sin balanceo

# EL MODELO NO RESTRINGIDO ES EL QUE UTILIZAN OTROS PAQUETES ESTADISTICO
COMO EL STATGRAPHICS POR DEFECTO Y Se EXPLICA EN LOS LIBROS DE TEXTO.

#Así para el cálculo del valor F la forma correcta sería:

#Cociente F correcto sobre el efecto del factor aleatorio (aquí designado
como B):

```

```

#Model no restringido: F_factor= dividir MS_factor/MSAB
#Model restringido: F_factor= dividir MS_factor/MSE

# COMPONENTES DE LA VARIANZA
#Ara calculamos la estimación de las componentes de la varianza del
modelo ANOVA no restringido (por ejemplo)
# En futuros cálculos, nos resultará más cómodo indicar como a, b i n
# el número de niveles de cada factor y el tamaño muestral en cada
casilla:
a <- length(levels(pous[,"Hora"])) # niveles de hora
b <- length(levels(pous[,"Pou"])) # niveles de pous
# Solo válido para este caso balanceado:
n <- nrow(pous) %/% (a*b) # son las réplicas

a
## [1] 2

b
## [1] 3

n
## [1] 2

#ANOVA
anovaTab <- anova2cross(pous.aov, random = "Pou", restricted = FALSE)
anovaTab

## Analysis of Variance Table
##
## Response: Salinitat
##          Df Sum Sq Mean Sq  F value    Pr(>F)
## Hora      1  20.7770  20.7770  1555.3593 0.0006423 ***
## Pou       2   0.1211   0.0606    4.5334 0.1807215
## Hora:Pou  2   0.0267   0.0134    3.7718 0.0869477 .
## Residuals 6   0.0212   0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Sumas de cuadrados medios
ms.pou <- anovaTab[2,3]
ms.hora_pou <- anovaTab[3,3]
ms.resid <- anovaTab[4,3]
#graus de llibertat
gll.hora <- anovaTab[1,1]
gll.pou <- anovaTab[2,1]
gll.hora_pou <- anovaTab[3,1]
gll.resid <- anovaTab[4,1]

```

```

# Componentes de La varianza (factor pouzo e interacción horas pouzo):
# Componentes de La varianza de pou:

var.pou <- (ms.pou - ms.hora_pou) / (a*n)
var.pou      #componente de La varianza de POU

## [1] 0.0118

var.hora_pou <- (ms.hora_pou - ms.resid) / n
var.hora_pou  #componente de La variancia de HORA-POU

## [1] 0.004908333

var.resid <- ms.resid
var.resid    #componente de La variancia de RESIDUO (ERROR)

## [1] 0.003541667

# En forma de tantos por ciento:
var.total <- var.pou + var.hora_pou + var.resid #total
round(100 * var.pou / var.total, 2) #componente de La varianza de POU

## [1] 58.27

round(100 * var.hora_pou / var.total, 2) #componente de La varianza de
HORA-POU

## [1] 24.24

round(100 * var.resid / var.total, 2) #componente de La varianza RESIDUAL

## [1] 17.49

#Estimación de La media global y de Los efectos del factor hora:
coef(pous.aov)

## (Intercept)          Hora1          Pou1          Pou2  Hora1:Pou1
## 6.054166667 -1.315833333 -0.129166667  0.013333333  0.060833333
## Hora1:Pou2
## -0.006666667

#Estimación de La media global y de Los efectos del factor hora:
dummy.coef(pous.aov)

## Full coefficients are
##
## (Intercept):          6.054167
## Hora:                21h              7h
##                      -1.315833      1.315833
## Pou:                  Pou1            Pou2            Pou3
##                      -0.1291667    0.01333333    0.11583333
## Hora:Pou:            21h:Pou1      7h:Pou1        21h:Pou2      7h:Pou2
##                      0.06083333 -0.06083333 -0.006666667  0.006666667

```

```

##
## (Intercept):
## Hora:
##
## Pou:
##
## Hora:Pou:          21h:Pou3      7h:Pou3
##                   -0.054166667  0.054166667

#Estimación de La media global y de Los efectos del factor hora:

model.tables(pous.aov)

## Tables of effects
##
## Hora
## Hora
##   21h    7h
## -1.3158 1.3158
##
## Pou
## Pou
##   Pou1    Pou2    Pou3
## -0.12917 0.01333 0.11583
##
## Hora:Pou
##   Pou
## Hora Pou1    Pou2    Pou3
##   21h 0.06083 -0.00667 -0.05417
##   7h -0.06083 0.00667 0.05417

#Estimación de La media global y medias de Los niveles de Los factores:

model.tables(pous.aov, type = "means")

## Tables of means
## Grand mean
##
## 6.054167
##
## Hora
## Hora
##   21h    7h
## 4.738 7.370
##
## Pou
## Pou
##   Pou1  Pou2  Pou3
## 5.925 6.068 6.170
##
## Hora:Pou

```

```

##      Pou
## Hora  Pou1  Pou2  Pou3
##   21h 4.670 4.745 4.800
##   7h  7.180 7.390 7.540

#####
#) Ahora va a suponer que los pozos muestreados son diferentes en cada
una de las horas, por tanto se trata claramente de un factor (factor
aleatorio pou) jerarquizado a hora (factor fijo). Apartado 2: Modelo con
2 factores: 1 fijo, 1 aleatorio, jerárquico o anidado
#####

# imaginamos que son tratados como factores fijos incorrectamente:
pous1.aov <- aov(Salinitat ~ Hora + Pou %in% Hora, data=pous)
anova(pous1.aov) # factor hora p-valor = 3.334e-10; factor Pou(hora) p-
valor =0.007192

## Analysis of Variance Table
##
## Response: Salinitat
##          Df Sum Sq Mean Sq F value    Pr(>F)
## Hora      1 20.7770 20.7770 5866.449 3.334e-10 ***
## Hora:Pou  4  0.1478  0.0370  10.435  0.007192 **
## Residuals 6  0.0212  0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Forma equivalente, más concisa
pous1.aov <- aov(Salinitat ~ Hora/Pou, data=pous)
anova(pous1.aov) # factor hora p-valor = 3.334e-10; factor Pou(hora) p-
valor =0.007192

## Analysis of Variance Table
##
## Response: Salinitat
##          Df Sum Sq Mean Sq F value    Pr(>F)
## Hora      1 20.7770 20.7770 5866.449 3.334e-10 ***
## Hora:Pou  4  0.1478  0.0370  10.435  0.007192 **
## Residuals 6  0.0212  0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# La tabla ANOVA anterior no es correcta!!!!!!!!!!!!!!!!!!!!!!
# Tabla ANOVA correcta dado que el primer factor es fijo (hora) y el otro
# factor es aleatorio (Factor pozo). Paso a paso sería:

# Para poder calcular correctamente los valores F disponemos de la
función 'anova2nest'. Es necesario que el fichero "anova2nest.R" esté en
el directorio de trabajo de R:

```

```

source("anova2nest.R")

# Mucho más directo :
anova2nest(pous1.aov, random = c("Pou")) # factor hora p-valor = 1.876e-
05; factor Pou(hora) p-valor =0.007192

## Analysis of Variance Table
##
## Response: Salinitat
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Hora       1 20.7770  20.7770  562.174 1.876e-05 ***
## Hora:Pou   4  0.1478   0.0370   10.435 0.007192 **
## Residuals  6  0.0212   0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Ahora calculamos la estimación de las componentes de la varianza del
# modelo ANOVA
# En futuros cálculos, nos resultará más cómodo indicar como a, b i n
# el número de niveles de cada factor y el tamaño muestral en cada
# casilla:
a <- length(levels(pous[,"Hora"])) # niveles de hora
b <- length(levels(pous[,"Pou"])) # niveles de pozos
# Solo válido para este caso balanceado:
n <- nrow(pous) %/% (a*b) # son las réplicas

a
## [1] 2

b
## [1] 3

n
## [1] 2

#ANOVA
anovaTab <- anova2nest(pous1.aov, random = c("Pou"))
anovaTab

## Analysis of Variance Table
##
## Response: Salinitat
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Hora       1 20.7770  20.7770  562.174 1.876e-05 ***
## Hora:Pou   4  0.1478   0.0370   10.435 0.007192 **
## Residuals  6  0.0212   0.0035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# Sumas de cuadrados medios
ms.pou_aniuat_hora <- anovaTab[2,3]
ms.resid <- anovaTab[3,3]

# Componentes de La varianza (factor POZO anidado a HORA):
var.pou_aniuat_hora <- (ms.pou_aniuat_hora - ms.resid) / n
var.pou_aniuat_hora      #componente de La varianza de POZO anidado a HORA
## [1] 0.01670833

var.resid <- ms.resid
var.resid      #componente de La varianza de RESIDU
## [1] 0.003541667

# En forma de porcentajes:
var.total <- var.pou_aniuat_hora + var.resid
round(100 * var.pou_aniuat_hora / var.total, 2)
## [1] 82.51

round(100 * var.resid / var.total, 2)
## [1] 17.49

# Estimación de La media global y Los efectos del factor hora:
coef(pous.aov)
## (Intercept)          Hora1          Pou1          Pou2  Hora1:Pou1
## 6.054166667 -1.315833333 -0.129166667  0.013333333  0.060833333
## Hora1:Pou2
## -0.006666667

dummy.coef(pous.aov)
## Full coefficients are
##
## (Intercept):          6.054167
## Hora:                21h          7h
##                    -1.315833    1.315833
## Pou:                  Pou1          Pou2          Pou3
##                    -0.12916667  0.01333333  0.11583333
## Hora:Pou:            21h:Pou1    7h:Pou1    21h:Pou2    7h:Pou2
##                    0.060833333 -0.060833333 -0.006666667  0.006666667
##
## (Intercept):
## Hora:
##
## Pou:
##

```



```
## Hora:Pou:          21h:Pou3      7h:Pou3
##                   -0.054166667  0.054166667
```

```
model.tables(pous.aov)
```

```
## Tables of effects
```

```
##
```

```
## Hora
```

```
## Hora
```

```
##    21h      7h
```

```
## -1.3158  1.3158
```

```
##
```

```
## Pou
```

```
## Pou
```

```
##    Pou1      Pou2      Pou3
```

```
## -0.12917  0.01333  0.11583
```

```
##
```

```
## Hora:Pou
```

```
##    Pou
```

```
## Hora  Pou1      Pou2      Pou3
```

```
##    21h  0.06083 -0.00667 -0.05417
```

```
##    7h  -0.06083  0.00667  0.05417
```

```
model.tables(pous.aov, type = "means")
```

```
## Tables of means
```

```
## Grand mean
```

```
##
```

```
## 6.054167
```

```
##
```

```
## Hora
```

```
## Hora
```

```
##    21h      7h
```

```
## 4.738 7.370
```

```
##
```

```
## Pou
```

```
## Pou
```

```
##    Pou1  Pou2  Pou3
```

```
## 5.925 6.068 6.170
```

```
##
```

```
## Hora:Pou
```

```
##    Pou
```

```
## Hora  Pou1  Pou2  Pou3
```

```
##    21h 4.670 4.745 4.800
```

```
##    7h  7.180 7.390 7.540
```

2.11.3: Heredabilidad de las cualidades anaerobias (3 factores anidados)

El ejercicio anaeróbico consiste en realizar actividades de alta intensidad como carreras cortas de gran velocidad, abdominales, o cualquier ejercicio que precise mucho esfuerzo durante poco tiempo. Los músculos entrenados con el ejercicio anaeróbico ofrecen mayor rendimiento al realizar actividades de corta duración y gran intensidad, por lo que este tipo de ejercicio se utiliza para adquirir potencia y masa muscular, y sirve para fortalecer el sistema musculoesquelético (Ver en <https://www.webconsultas.com/ejercicio-y-deporte/vida-activa/tipos-de-deporte/el-ejercicio-anaerobico-1888>)

En un estudio sobre la heredabilidad de las cualidades anaeróbicas se consideran 4 parejas de hermanos gemelos monocigotos y 4 parejas dicigotos. Entre otras variables se mide la potencia anaeróbica desarrollada en 15" (W/Kg), obteniendo dos medidas para cada individuo. Se desea estudiar si estas cualidades anaerobias se heredan y por lo tanto las medias por grupos son diferentes, pero teniendo en cuenta que en el experimento los individuos monocigotos están jerárquicamente ligados a cada grupo estudiado, son por ello diferentes y además, son parejas de hermanos.

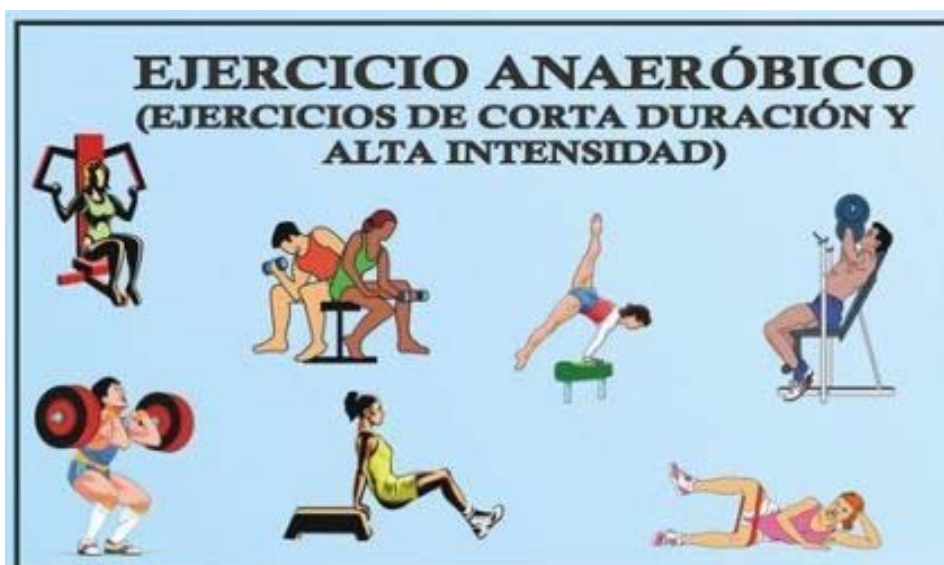


Imagen de <http://2.bp.blogspot.com/-HwidcLES2rE/UKZ33DvN7oI/AAAAAAAAAFA/F3v-BEmdEm4/s1600/ejercicio-anaerobico.jpg>

Los datos obtenidos son los siguientes:

#Monocigotos:

| | | |
|------------|------|-------|
| #P.1 Ind.1 | 20.6 | 20.62 |
| #P.1.Ind.2 | 20.3 | 20.7 |
| #P.2.Ind.1 | 22.1 | 22.4 |
| #P.2.Ind.2 | 22.3 | 23 |
| #P.3.Ind.1 | 21.4 | 23.4 |
| #P.3.Ind2. | 24 | 23.2 |
| #P.4.Ind.1 | 24 | 23.1 |
| #P.4.Ind.2 | 23.5 | 22 |

#Dicigotos

| | | |
|------------|------|-------|
| #P.1 Ind.1 | 21.6 | 22.62 |
| #P.1.Ind.2 | 20.5 | 21.7 |
| #P.2.Ind.1 | 22.2 | 24.4 |
| #P.2.Ind.2 | 23.3 | 25 |
| #P.3.Ind.1 | 24.4 | 22.4 |
| #P.3.Ind2. | 21 | 21.2 |
| #P.4.Ind.1 | 20 | 21.1 |
| #P.4.Ind.2 | 22.1 | 22 |

```
op <- options(show.signif.stars=FALSE, contrasts=c("contr.sum",  
"contr.poly"))
```

#Para los cálculos del estadístico F es necesario disponer de la función anova3nest() creada por el Dr Jordi Ocaña.

#Para disponer de la función 'anova3nest'. Es necesario que el archivo "anova3nest.R" esté situado

en el directorio de trabajo de R:

```
source("anova2i3.R")
```

```
bessons <- read.table("bessons.txt", header=T)
```

```
fix(bessons)
```

```
bessons[,1:3] <- lapply(bessons[,1:3], as.factor)
```

```
sapply(bessons, class)
```

```
##      zigosi      parella      individu CAnaerobia
```

```
## "factor" "factor" "factor" "numeric"
```

En futuros cálculos, nos resultará más cómodo indicar como a, b, d i n el número de niveles de cada factor y el tamaño muestral en cada casilla:

```
a <- length(levels(bessons[, "zigosi"]))
```

```
b <- length(levels(bessons[, "parella"]))
```

```

d <- length(levels(bessons[,"individu"]))
# Solo válido para este caso balanceado:
n <- nrow(bessons) %/% (a*b*d)

# La forma más explícita de indicar el modelo lineal jerarquizado en el
que
# 'parella' es jerárquicamente dentro de 'zigosi' y 'individu' está
dentro de las combinaciones de 'zigosi' y 'parella' es:
bessons.aov <- aov(CAnaerobia ~ zigosi + parella %in% zigosi + individu
%in% (parella %in% zigosi), data=bessons)
# o bien:
bessons.aov <- aov(CAnaerobia ~ zigosi + parella %in% zigosi + individu
%in% (zigosi/parella), data=bessons)

anova(bessons.aov)

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##           Df Sum Sq Mean Sq F value    Pr(>F)
## zigosi      1  0.0378   0.0378   0.0506 0.8248622
## zigosi:parella 6 31.0803   5.1800   6.9325 0.0009065
## zigosi:parella:individu 8 11.5347   1.4418   1.9296 0.1252416
## Residuals  16 11.9554   0.7472

# Forma equivalente, más concisa
bessons.aov <- aov(CAnaerobia ~ zigosi/parella/individu, data=bessons)
anova(bessons.aov)

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##           Df Sum Sq Mean Sq F value    Pr(>F)
## zigosi      1  0.0378   0.0378   0.0506 0.8248622
## zigosi:parella 6 31.0803   5.1800   6.9325 0.0009065
## zigosi:parella:individu 8 11.5347   1.4418   1.9296 0.1252416
## Residuals  16 11.9554   0.7472

# La tabla ANOVA anterior no es correcta.
# Tabla ANOVA correcta dado que el primer factor es fijo y los dos
últimos
# factores son aleatorios. Paso a paso sería:
anovaTab <- anova(bessons.aov)
anovaTab

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##           Df Sum Sq Mean Sq F value    Pr(>F)
## zigosi      1  0.0378   0.0378   0.0506 0.8248622

```

```

## zigosi:parella      6 31.0803  5.1800  6.9325 0.0009065
## zigosi:parella:individu 8 11.5347  1.4418  1.9296 0.1252416
## Residuals          16 11.9554  0.7472

ms.zigosi <- anovaTab[1,3]
ms.parella <- anovaTab[2,3]
ms.individu <- anovaTab[3,3]
ms.resid <- anovaTab[4,3]

gll.zigosi <- anovaTab[1,1]
gll.parella <- anovaTab[2,1]
gll.individu <- anovaTab[3,1]

anovaTab[1, 4] <- f.zigosi <- ms.zigosi / ms.parella
anovaTab[2, 4] <- f.parella <- ms.parella / ms.individu

anovaTab[1, 5] <- pf(f.zigosi, gll.zigosi, gll.parella, lower.tail =
FALSE)
anovaTab[2, 5] <- pf(f.parella, gll.parella, gll.individu, lower.tail =
FALSE)

anovaTab

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##           Df Sum Sq Mean Sq F value Pr(>F)
## zigosi      1  0.0378   0.0378   0.0073 0.93469
## zigosi:parella 6 31.0803   5.1800   3.5927 0.04958
## zigosi:parella:individu 8 11.5347   1.4418   1.9296 0.12524
## Residuals  16 11.9554   0.7472

# Mucho más directo:
anova3nest(bessons.aov, random = c("parella", "individu"))

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##           Df Sum Sq Mean Sq F value Pr(>F)
## zigosi      1  0.0378   0.0378   0.0073 0.93469
## zigosi:parella 6 31.0803   5.1800   3.5927 0.04958
## zigosi:parella:individu 8 11.5347   1.4418   1.9296 0.12524
## Residuals  16 11.9554   0.7472

# Equivalente a:
anova3nest(bessons.aov, random = c("individu", "parella"))

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##           Df Sum Sq Mean Sq F value Pr(>F)

```

```

## zigosi          1  0.0378  0.0378  0.0073  0.93469
## zigosi:parella  6 31.0803  5.1800  3.5927  0.04958
## zigosi:parella:individu  8 11.5347  1.4418  1.9296  0.12524
## Residuals      16 11.9554  0.7472

# (El orden en el que se enumeran los factores aleatorios no importa)

# También equivalente a:
anova3nest(bessons.aov, random = "parella")

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##          Df Sum Sq Mean Sq F value Pr(>F)
## zigosi    1  0.0378  0.0378  0.0073 0.93469
## zigosi:parella  6 31.0803  5.1800  3.5927 0.04958
## zigosi:parella:individu  8 11.5347  1.4418  1.9296 0.12524
## Residuals  16 11.9554  0.7472

# (Todos los factores jerárquicamente inferiores de un factor aleatorio
# los considera aleatorios)

# En cambio, no sería lo mismo:
anova3nest(bessons.aov, random = "individu")

## Analysis of Variance Table
##
## Response: CAnaerobia
##
##          Df Sum Sq Mean Sq F value Pr(>F)
## zigosi    1  0.0378  0.0378  0.0262 0.87537
## zigosi:parella  6 31.0803  5.1800  3.5927 0.04958
## zigosi:parella:individu  8 11.5347  1.4418  1.9296 0.12524
## Residuals  16 11.9554  0.7472

# (ahora consideraría que el único aleatorio es el último
# jerárquicamente)

# Componentes de la varianza:
var.parella <- (ms.parella - ms.individu) / (d*n)
var.parella

## [1] 0.9345521

var.individu <- (ms.individu - ms.resid) / n
var.individu

## [1] 0.3473125

```

```

var.resid <- ms.resid
var.resid

## [1] 0.7472125

# En forma de porcentajes:
var.total <- var.parella + var.individu + var.resid
round(100 * var.parella / var.total, 2)

## [1] 46.06

round(100 * var.individu / var.total, 2)

## [1] 17.12

round(100 * var.resid / var.total, 2)

## [1] 36.83

coef(bessons.aov)

##                (Intercept)                zigosimono:parella1
##                22.254375                -0.034375
##                zigosidi:parella1                zigosimono:parella2
##                -0.615000                -1.733750
##                zigosidi:parella2                zigosimono:parella3
##                1.505000                0.161250
##                zigosidi:parella3                zigosimono:parella4
##                0.030000                0.711250
##                zigosidi:parella1:individu1 zigosimono:parella1:individu1
##                0.505000                0.055000
##                zigosidi:parella2:individu1 zigosimono:parella2:individu1
##                -0.425000                -0.200000
##                zigosidi:parella3:individu1 zigosimono:parella3:individu1
##                1.150000                -0.600000
##                zigosidi:parella4:individu1 zigosimono:parella4:individu1
##                -0.750000                0.400000

dummy.coef(bessons.aov)

## Full coefficients are
##
## (Intercept):                22.25437
## zigosimono:                di                mono
##                -0.034375                0.034375
## zigosimono:parella:                di:1                mono:1                di:2                mono:2
## di:3
##                -0.61500                -1.73375                1.50500                0.16125
## 0.03000
## zigosimono:parella:individu:                di:1:1                mono:1:1                di:2:1                mono:2:1
## di:3:1
##                0.505                0.055                -0.425                -0.200

```

```

1.150
##
## (Intercept):
## zigosi:
##
## zigosi:parella:          mono:3      di:4      mono:4
##                          0.71125 -0.92000  0.86125
## zigosi:parella:individu: mono:3:1    di:4:1    mono:4:1  di:1:2    mono:1:2
##                          -0.600   -0.750   0.400 -0.505   -0.055
##
## (Intercept):
## zigosi:
##
## zigosi:parella:
##
## zigosi:parella:individu: di:2:2 mono:2:2 di:3:2 mono:3:2 di:4:2
##                          0.425    0.200 -1.150    0.600  0.750
##
## (Intercept):
## zigosi:
##
## zigosi:parella:
##
## zigosi:parella:individu: mono:4:2
##                          -0.400

model.tables(bessons.aov)

## Tables of effects
##
## zigosi
## zigosi
##      di      mono
## -0.03437  0.03437
##
## zigosi:parella
##      parella
## zigosi 1      2      3      4
## di -0.6150  1.5050  0.0300 -0.9200
## mono -1.7338  0.1613  0.7113  0.8613
##
## zigosi:parella:individu
## , , individu = 1
##
##      parella
## zigosi 1      2      3      4
## di  0.505 -0.425  1.150 -0.750
## mono 0.055 -0.200 -0.600  0.400
##
## , , individu = 2

```



```

##
##      parella
## zigosi 1      2      3      4
##   di  -0.505  0.425 -1.150  0.750
##   mono -0.055  0.200  0.600 -0.400

model.tables(bessons.aov, type = "means")

## Tables of means
## Grand mean
##
## 22.25437
##
## zigosi
## zigosi
##   di  mono
## 22.220 22.289
##
## zigosi:parella
##      parella
## zigosi 1      2      3      4
##   di  21.605 23.725 22.250 21.300
##   mono 20.555 22.450 23.000 23.150
##
## zigosi:parella:individu
## , , individu = 1
##
##      parella
## zigosi 1      2      3      4
##   di  22.11 23.30 23.40 20.55
##   mono 20.61 22.25 22.40 23.55
##
## , , individu = 2
##
##      parella
## zigosi 1      2      3      4
##   di  21.10 24.15 21.10 22.05
##   mono 20.50 22.65 23.60 22.75

```

ANOVA diseños de medidas repetidas

2.12.1. Diseños de experimentos longitudinales (medidas repetidas)

El ANOVA de medidas repetidas es el equivalente del ANOVA “one-way”, pero para grupos relacionados, no independientes, y es la extensión de la prueba t dependiente (apareada) a más de dos niveles del factor/es estudiado/s.

Un ANOVA de medidas repetidas también se conoce como ANOVA o ANOVA dentro de los sujetos para muestras correlacionadas. Todos estos nombres implican la naturaleza de las medidas repetidas ANOVA, la de una prueba para detectar cualquier diferencia general entre las medias relacionadas. Hay muchos diseños complejos que pueden hacer uso de medidas repetidas, pero sólo presentaré un ejemplo simple. Esta prueba particular requiere una variable independiente y una variable dependiente.

En el diseño medidas repetidas todos los sujetos de la muestra reciben todos los tratamientos en el tiempo. La tendencia en el tiempo de las respuestas individuales al tratamiento es un aspecto importante para muchos experimentos, como por ejemplo, en un ensayo clínico, cuando la respuesta medida de los pacientes a un tratamiento se mide regularmente. Otro ejemplo podría ser el caso en los que se pesan animales cada semana para supervisar su crecimiento bajo diferentes condiciones de nutrición.

En los diseños ANOVA de medidas repetidas, la variable independiente tiene categorías llamadas niveles o grupos relacionados. Cuando las mediciones se repiten a lo largo del tiempo, como cuando se miden los cambios en la presión arterial debido a un tipo de programa de entrenamiento con ejercicios, la variable independiente es el tiempo. Cada nivel (o grupo relacionado) es un punto de tiempo específico. Por lo tanto, para el estudio de entrenamiento con ejercicios, habría una serie de puntos de tiempo y cada punto de tiempo es un nivel de la variable independiente.

Vamos a presentar un ejemplo obtenido de <https://stats.idre.ucla.edu/r/seminars/repeated-measures-analysis-with-r/>

Hay una serie de situaciones que pueden surgir cuando el análisis incluye los efectos de los grupos y los efectos de los sujetos. El ejemplo versa sobre dos grupos de pacientes con depresión en 3 momentos de tiempo y cómo evoluciona esta enfermedad, se trata de averiguar si existen diferencias entre grupos y ver si el tiempo es un factor relevante

En este ejemplo, los dos grupos comienzan siendo bastante diferentes en cuanto a los niveles de depresión, pero terminan siendo bastante cercanos a la depresión.

La prueba entre grupos indica que el grupo de variables es significativo, por lo tanto, en el gráfico, vemos que las líneas para los dos grupos están bastante separadas. La prueba dentro del sujeto indica que hay un efecto de tiempo significativo, en otras palabras, los grupos cambian con el tiempo, ambos grupos se deprimen menos con el

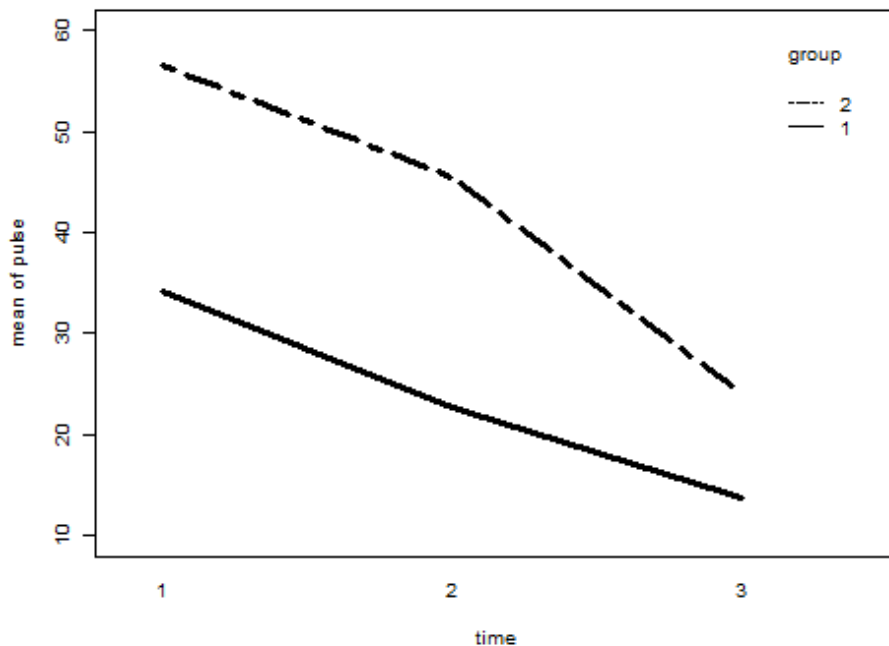
tiempo. Además, la interacción de tiempo y grupo es significativa, lo que significa que los grupos están cambiando con el tiempo, pero están cambiando de diferentes maneras, lo que significa que en el gráfico las líneas no serán paralelas. En la gráfica vemos que los grupos tienen líneas no paralelas que disminuyen con el tiempo y se van acercando progresivamente entre sí.

2.12.2 Ejemplo simple de medidas repetidas

```
#Example from UCLA
demo3 <- read.csv("https://stats.idre.ucla.edu/stat/data/demo3.csv")
## Convert variables to factor
demo3 <- within(demo3, {
  group <- factor(group)
  time <- factor(time)
  id <- factor(id)
})

par(cex = .6)

with(demo3, interaction.plot(time, group, pulse,
  ylim = c(10, 60), lty = c(1, 12), lwd = 3,
  ylab = "mean of pulse", xlab = "time", trace.label = "group"))
```



```
demo3.aov <- aov(pulse ~ group * time + Error(id), data = demo3)
summary(demo3.aov)
```

```
##
## Error: id
##           Df Sum Sq Mean Sq F value Pr(>F)
## group      1 2035.0  2035.0   343.1 1.6e-06 ***
## Residuals  6   35.6    5.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## time       2 2830.3  1415.2   553.8 1.52e-12 ***
## group:time  2  200.3   100.2    39.2 5.47e-06 ***
## Residuals  12   30.7    2.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La prueba ANOVA entre grupos indica que el grupo de variables es significativo, por lo tanto, en el gráfico, vemos que las líneas para los dos grupos están bastante separadas.

La prueba dentro del sujeto indica que hay un efecto de tiempo significativo, en otras palabras, los grupos cambian con el tiempo, ambos grupos se deprimen menos con el tiempo.

Además, la interacción de tiempo y grupo es significativa, lo que significa que los grupos están cambiando con el tiempo, pero están cambiando de diferentes maneras, lo que significa que en el gráfico las líneas no serán paralelas. En el gráfico vemos que los grupos tienen líneas no paralelas que disminuyen con el tiempo y se van acercando progresivamente entre sí.

2.12.3 Crecimiento de los dientes en niños y niñas

Se va a utilizar un ejemplo procedente del conjunto de datos Orthodont (data(Orthodont)) de la librería library(nlme). Se trata de datos de la curva de crecimiento en una medida de ortodoncia que tiene 108 filas y 4 columnas del cambio en una medición de ortodoncia a lo largo del tiempo para varios sujetos jóvenes.

La variable utilizada en este estudio es la distancia desde la pituitaria hasta la fisura pterigomaxilar (mm). Estas distancias se miden en imágenes de rayos X del cráneo.



Imagen

https://img.europapress.es/fotoweb/fotonoticia_20151216070932_640.jpg

de

EJEMPLO OBTENIDO DE:
<http://stats.stackexchange.com/questions/7185/what-is-the-difference->

between-using-aov-and-lme-in-analyzing-a-longitudinal

#CRECIMIENTO DE LOS DIENTES DE UNOS NIÑOS/NIÑAS SEGUN EDAD Y SEXO

library(nlme)

data(Orthodont)

Orthodont

Grouped Data: distance ~ age | Subject

| ## | distance | age | Subject | Sex |
|-------|----------|-----|---------|------|
| ## 1 | 26.0 | 8 | M01 | Male |
| ## 2 | 25.0 | 10 | M01 | Male |
| ## 3 | 29.0 | 12 | M01 | Male |
| ## 4 | 31.0 | 14 | M01 | Male |
| ## 5 | 21.5 | 8 | M02 | Male |
| ## 6 | 22.5 | 10 | M02 | Male |
| ## 7 | 23.0 | 12 | M02 | Male |
| ## 8 | 26.5 | 14 | M02 | Male |
| ## 9 | 23.0 | 8 | M03 | Male |
| ## 10 | 22.5 | 10 | M03 | Male |
| ## 11 | 24.0 | 12 | M03 | Male |
| ## 12 | 27.5 | 14 | M03 | Male |
| ## 13 | 25.5 | 8 | M04 | Male |
| ## 14 | 27.5 | 10 | M04 | Male |
| ## 15 | 26.5 | 12 | M04 | Male |
| ## 16 | 27.0 | 14 | M04 | Male |
| ## 17 | 20.0 | 8 | M05 | Male |
| ## 18 | 23.5 | 10 | M05 | Male |
| ## 19 | 22.5 | 12 | M05 | Male |
| ## 20 | 26.0 | 14 | M05 | Male |
| ## 21 | 24.5 | 8 | M06 | Male |
| ## 22 | 25.5 | 10 | M06 | Male |
| ## 23 | 27.0 | 12 | M06 | Male |
| ## 24 | 28.5 | 14 | M06 | Male |
| ## 25 | 22.0 | 8 | M07 | Male |
| ## 26 | 22.0 | 10 | M07 | Male |
| ## 27 | 24.5 | 12 | M07 | Male |
| ## 28 | 26.5 | 14 | M07 | Male |
| ## 29 | 24.0 | 8 | M08 | Male |
| ## 30 | 21.5 | 10 | M08 | Male |
| ## 31 | 24.5 | 12 | M08 | Male |
| ## 32 | 25.5 | 14 | M08 | Male |
| ## 33 | 23.0 | 8 | M09 | Male |
| ## 34 | 20.5 | 10 | M09 | Male |
| ## 35 | 31.0 | 12 | M09 | Male |
| ## 36 | 26.0 | 14 | M09 | Male |
| ## 37 | 27.5 | 8 | M10 | Male |
| ## 38 | 28.0 | 10 | M10 | Male |
| ## 39 | 31.0 | 12 | M10 | Male |
| ## 40 | 31.5 | 14 | M10 | Male |

| | | | | |
|-------|------|----|-----|--------|
| ## 41 | 23.0 | 8 | M11 | Male |
| ## 42 | 23.0 | 10 | M11 | Male |
| ## 43 | 23.5 | 12 | M11 | Male |
| ## 44 | 25.0 | 14 | M11 | Male |
| ## 45 | 21.5 | 8 | M12 | Male |
| ## 46 | 23.5 | 10 | M12 | Male |
| ## 47 | 24.0 | 12 | M12 | Male |
| ## 48 | 28.0 | 14 | M12 | Male |
| ## 49 | 17.0 | 8 | M13 | Male |
| ## 50 | 24.5 | 10 | M13 | Male |
| ## 51 | 26.0 | 12 | M13 | Male |
| ## 52 | 29.5 | 14 | M13 | Male |
| ## 53 | 22.5 | 8 | M14 | Male |
| ## 54 | 25.5 | 10 | M14 | Male |
| ## 55 | 25.5 | 12 | M14 | Male |
| ## 56 | 26.0 | 14 | M14 | Male |
| ## 57 | 23.0 | 8 | M15 | Male |
| ## 58 | 24.5 | 10 | M15 | Male |
| ## 59 | 26.0 | 12 | M15 | Male |
| ## 60 | 30.0 | 14 | M15 | Male |
| ## 61 | 22.0 | 8 | M16 | Male |
| ## 62 | 21.5 | 10 | M16 | Male |
| ## 63 | 23.5 | 12 | M16 | Male |
| ## 64 | 25.0 | 14 | M16 | Male |
| ## 65 | 21.0 | 8 | F01 | Female |
| ## 66 | 20.0 | 10 | F01 | Female |
| ## 67 | 21.5 | 12 | F01 | Female |
| ## 68 | 23.0 | 14 | F01 | Female |
| ## 69 | 21.0 | 8 | F02 | Female |
| ## 70 | 21.5 | 10 | F02 | Female |
| ## 71 | 24.0 | 12 | F02 | Female |
| ## 72 | 25.5 | 14 | F02 | Female |
| ## 73 | 20.5 | 8 | F03 | Female |
| ## 74 | 24.0 | 10 | F03 | Female |
| ## 75 | 24.5 | 12 | F03 | Female |
| ## 76 | 26.0 | 14 | F03 | Female |
| ## 77 | 23.5 | 8 | F04 | Female |
| ## 78 | 24.5 | 10 | F04 | Female |
| ## 79 | 25.0 | 12 | F04 | Female |
| ## 80 | 26.5 | 14 | F04 | Female |
| ## 81 | 21.5 | 8 | F05 | Female |
| ## 82 | 23.0 | 10 | F05 | Female |
| ## 83 | 22.5 | 12 | F05 | Female |
| ## 84 | 23.5 | 14 | F05 | Female |
| ## 85 | 20.0 | 8 | F06 | Female |
| ## 86 | 21.0 | 10 | F06 | Female |
| ## 87 | 21.0 | 12 | F06 | Female |
| ## 88 | 22.5 | 14 | F06 | Female |
| ## 89 | 21.5 | 8 | F07 | Female |
| ## 90 | 22.5 | 10 | F07 | Female |


```

## 91      23.0  12    F07 Female
## 92      25.0  14    F07 Female
## 93      23.0   8    F08 Female
## 94      23.0  10    F08 Female
## 95      23.5  12    F08 Female
## 96      24.0  14    F08 Female
## 97      20.0   8    F09 Female
## 98      21.0  10    F09 Female
## 99      22.0  12    F09 Female
## 100     21.5  14    F09 Female
## 101     16.5   8    F10 Female
## 102     19.0  10    F10 Female
## 103     19.0  12    F10 Female
## 104     19.5  14    F10 Female
## 105     24.5   8    F11 Female
## 106     25.0  10    F11 Female
## 107     28.0  12    F11 Female
## 108     28.0  14    F11 Female

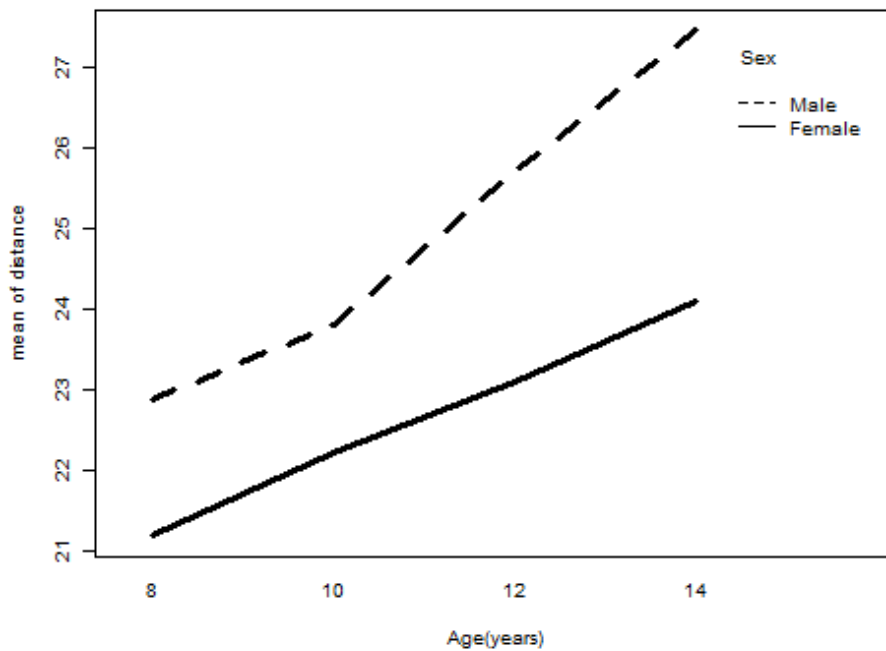
```

```
par(cex=.6)
```

```

with(Orthodont, interaction.plot(age, Sex, distance,
                                lwd = 3,
                                ylab = "mean of distance", xlab =
"Age(years)", trace.label = "Sex"))

```



```
#una manera de hacerlo con el paquete lme
res <- lme(distance ~ age*Sex, random = ~ 1 | Subject, data = Orthodont)
anova(res)
```

```
##              numDF denDF  F-value p-value
## (Intercept)      1    79 4123.156 <.0001
## age              1    79  122.450 <.0001
## Sex              1    25   9.292 0.0054
## age:Sex          1    79   6.303 0.0141
```

#TAMBIEN PUEDE HACERSE:

```
res <- aov(distance ~ age*Sex + Error(Subject), data = Orthodont)
summary(res)
```

```
##
## Error: Subject
##           Df Sum Sq Mean Sq F value Pr(>F)
## Sex        1  140.5  140.46   9.292 0.00538
## Residuals 25  377.9   15.12
##
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## age         1 235.36  235.36 122.450 <2e-16
## age:Sex     1  12.11   12.11   6.303 0.0141
## Residuals 79 151.84    1.92
```

*#For lme(), you used list(id=pdDiag(~time)), which not only adds a
#random intercept to the model, but also a random slope. Moreover,
#by using pdDiag, you are setting the correlation between the random
#intercept and slope to zero. This is a different model than what you
#specified via aov() and hence you get different results.*

#mas ejercicios en:

http://www.ats.ucla.edu/stat/r/seminars/Repeated_Measures/repeated_measures.htm

otros en: http://www.ccsr.ac.uk/staff/documents/Rworkshop_000.pdf

```
library(nlme)
```

```
library(lattice)
```

#The dataset we will be using in this workshop is from Chapter 7 of the Singer and

#Willet book (page 244, Table 7.1

#). On each of four days, spaced exactly one week

#apart, 35 people completed an inventory that assesses their performance on a timed

#cognitive task called 'opposites naming' At wave 1 each person also completed a

```

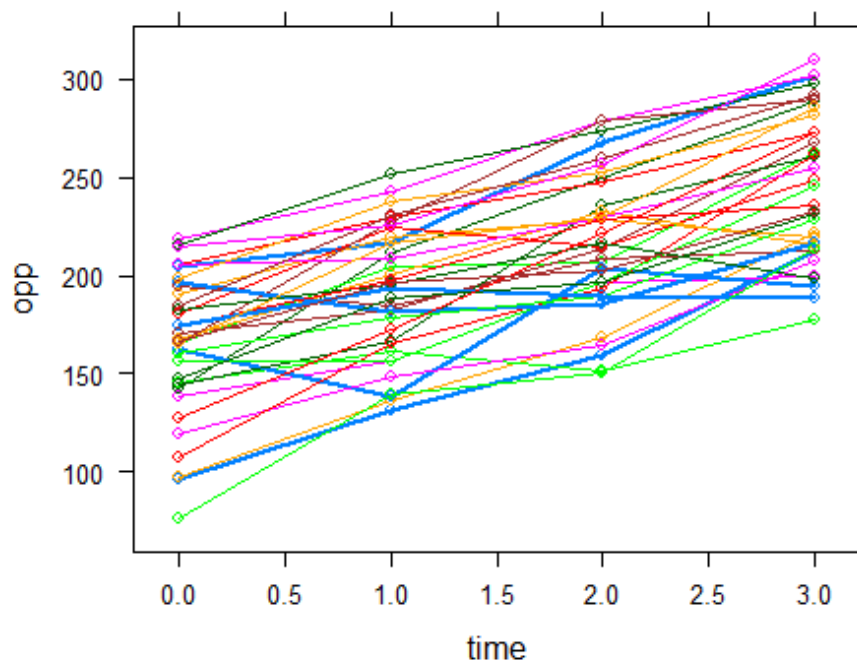
#standardised instrument assessing general co
#gnitive skill. The full dataset has 140
#records 4 for each person. The research question here is to
#see (a) whether the
#opposites naming skill improves with time and (b) whether the
#opposites naming
#skill increases more rapidly with time among individuals
#with stronger cognitive
#skills

#datos en: http://www.ccsr.ac.uk/staff/jk.htm

# Los datos estan en: http://www.ccsr.ac.uk/staff/opposites\_pp.txt

opposites <- read.table("opposites_pp.txt",header=TRUE,sep=",")
xyplot(opp~time, groups= id, type="o", panel=panel.superpose, data=
opposites)

```



```

fit.ar1 <- gls(opp ~time*ccog, corr=corCompSymm(form= ~ 1 | id ),
data=opposites)

summary(fit.ar1)

## Generalized least squares fit by REML
## Model: opp ~ time * ccog
## Data: opposites
## AIC BIC logLik

```

```

## 1299.048 1316.524 -643.5238
##
## Correlation Structure: Compound symmetry
## Formula: ~1 | id
## Parameter estimate(s):
## Rho
## 0.7309599
##
## Coefficients:
## Value Std.Error t-value p-value
## (Intercept) 164.37429 5.687004 28.903492 0.0000
## time 26.95998 1.375882 19.594686 0.0000
## ccog -0.11355 0.461855 -0.245862 0.8062
## time:ccog 0.43286 0.111739 3.873839 0.0002
##
## Correlation:
## (Intr) time ccog
## time -0.363
## ccog 0.000 0.000
## time:ccog 0.000 0.000 -0.363
##
## Standardized residuals:
## Min Q1 Med Q3 Max
## -2.5296418 -0.7420834 0.1047677 0.7473087 2.0926577
##
## Residual standard error: 35.09068
## Degrees of freedom: 140 total; 136 residual

#To get the estimated variance covariance matrix under this model type
getVarCov(fit.ar1)

## Marginal variance covariance matrix
## [,1] [,2] [,3] [,4]
## [1,] 1231.40 900.07 900.07 900.07
## [2,] 900.07 1231.40 900.07 900.07
## [3,] 900.07 900.07 1231.40 900.07
## [4,] 900.07 900.07 900.07 1231.40
## Standard Deviations: 35.091 35.091 35.091 35.091

```

3.PROBLEMAS DE CLASE PROPUESTOS/RESUELTOS

Este apartado se basa en el material docente desarrollado y utilizado en clase por el Dr Jordi Ocaña Rebull.

3.1 Problema: Variabilidad de los operarios en la fábrica de plásticos

En la industria del plástico, frecuentemente se utilizan grandes cantidades de resina. El proceso en el que se descarga de los vagones las cisternas cargadas de resina, en dirección a los depósitos de una fábrica, es un proceso delicado que se realiza mediante dispositivos de extracción algo complejos. En un experimento en el que se estudia los factores que afectan a la producción de plástico con cloruro de polivinilo (PVC), tres operarios seleccionados al azar, utilizaron ocho dispositivos de extracción de resina. Los dispositivos eran los ochos disponibles en la fábrica en la que se desarrollaba el estudio había interés en estudiar estos ocho dispositivos concretos. Cada operario utilizó cada uno de los ocho dispositivos. Para cada una de las 24 combinaciones, se fabricaron dos muestras de PVC. La respuesta observada se midió a partir de la partícula producida.

A partir de los listados y de los gráficos adjuntos, cuando sea necesario, responde a las siguientes cuestiones. Justifica siempre tus respuestas.



Fabricación de plástico (Imagen procedente de <https://www.cintex.com.mx/wp-content/uploads/2015/12/fabricacion-de-botellas-de-plastico-pet.jpg>)

1. Indica el diseño utilizado, los factores y su carácter (fijo o aleatorio), así como el número de niveles de cada factor. Indica también el modelo lineal asociado a su parametrización (supuestos, restricciones, ...). (1 punto)

Diseño factorial cruzado de dos factores: “operario” (“operator” en los listados), aleatorio con $a=3$ niveles; i “dispositivo extractor de resina” (“resin”) fijo con $g=8$ niveles. Es un diseño balanceado, con $n=2$. No hay indicación de aleatorización. Es de esperar, que de algún modo, se hizo como por ejemplo escogiendo aleatoriamente el orden en el cual los operarios utilizaron los dispositivos. Consideremos el caso no restringido.

El modelo lineal sería⁶:

Mida de la partícula en una mostra produïda per l'operari "i" amb el dispositiu "j" a la rèplica "k"

↓ Mitjana general

↓ ↓ Efecte aleatori de l'operari "i"

↓ ↓ ↓

$$Y_{ijk} = \mu + A_i + \beta_j + (A\beta)_{ij} + e_{ijk}$$

↑ ↑ ↑

↑ ↑ Residu de l'observació "ijk"

↑ Efecte aleatori de la interacció entre operari "i" i dispositiu "j"

Efecte fix del dispositiu "j"

On: $A_i \sim N(0, \sigma_A)$ independents i idènticament distribuïdes (iid),

$\sum_{j=1}^b \beta_j = 0$, $(A\beta)_{ij} \sim N(0, \sigma_{AB})$ iid, $e_{ijk} \sim N(0, \sigma)$ iid

i les A_i , $(A\beta)_{ij}$ i e_{ijk} mútuament independents.

Parametrització: el model té 12 paràmetres: μ , σ_A^2 , $\beta_1, \beta_2, \dots, \beta_8$ i σ_{AB}^2 (o 11).

2. Copia y completa aquí la tabla ANOVA correspondiente, y estudia la significación de los efectos principales del modelo, así como su interacción. Indica claramente las hipótesis planteadas y las conclusiones a las que se llega, dado un nivel de significación del 0.05. Puedes asumir que se verifican las condiciones de normalidad y de homogeneidad de varianzas. (2 puntos)

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | F taules |
|----------------|----|---------|---------|---------------------|--------|--------------------------|
| operator | 2 | 20.718 | 10.359 | 10.12 = 10.36/1.024 | < 0,05 | $F_{0,05}(2,14) = 3,74$ |
| resin | 7 | 283.946 | 40.564 | 39.61 = 40,56/1.024 | < 0,05 | $F_{0,05}(7,14) = 2,76$ |
| operator:resin | 14 | 14.335 | 1.024 | 0.69 = 1,024/1,478 | > 0,05 | $F_{0,05}(14,24) = 2,13$ |
| Residuals | 24 | 35.480 | 1.478 | | | |

$$H_0 : \sigma_A^2 = 0 \quad \text{vs} \quad H_1 : \sigma_A^2 > 0$$

⁶ Els factors s'indiquen en el mateix ordre que apareixen als llistats (per tant, primer el factor aleatori i després el fix), petita variant no intencionada respecte de lo explicado en clase. No debería de representar un gran problema.

Rechazamos H_0 , con una probabilidad de error de tipo I de 0,05, podemos concluir que la varianza asociada al operador no es nula.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_8 = 0 \text{ vs } H_1 : \beta_i \neq \beta_{i'}, \text{ per algun parell } i, i' = 1, \dots, 8, i \neq i'$$

Rechazamos H_0 , con una probabilidad de error de tipo I de 0,05, podemos concluir que el efecto del dispositivo no es nulo.

$$H_0 : \sigma_{AB}^2 = 0 \text{ vs } H_1 : \sigma_{AB}^2 > 0$$

No rechazamos H_0 , no tenemos evidencia de que haya interacción entre operario y dispositivo.

3. Estima todos los parámetros del modelo. ¿Tendría sentido dedicar recursos a la formación de los operarios con la finalidad de mejorar la calidad del producto, expresada en términos de menor variabilidad en la medida de las partículas de PVC? (2 puntos)

(Intercept): $32.35417 = \hat{\mu}$

Para los efectos alatorios se tienen que estimar las varianzas, no los coeficientes:

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_{AB}}{bn} = \frac{10,359 - 1,02}{8 \cdot 2} = 0,584$$

Los 5 primeros coeficientes de "resin" directamente de 'dummy.coef(...)':

| resin: j = | 1 | 2 | 3 | 4 | 5 |
|-------------------|-----------|-----------|------------|------------|------------|
| $\hat{\beta}_j =$ | 3.2958333 | 2.2625000 | -2.5041667 | -2.8875000 | -1.5041667 |

Los restantes se tienen que calcular:

$$\hat{\beta}_6 = \bar{Y}_{6..} - \bar{Y}_{...} = 30,20 - 32,35417 = -2,1541667$$

$$\hat{\beta}_7 = \bar{Y}_{7..} - \bar{Y}_{...} = 32,73 - 32,35417 = 0,3791667$$

$$\hat{\beta}_8 = -\sum_{j=1}^7 \hat{\beta}_j = 3,1125$$

Para los efectos aleatorios se tienen que estimar las varianzas, no los coeficientes:

$$\hat{\sigma}_{AB}^2 = \frac{MS_{AB} - MS_E}{n} = \frac{1,024 - 1,478}{2} < 0 \Rightarrow = 0$$

$$\hat{\sigma}^2 = MS_E = 1,478$$

Sí que tendría sentido dedicar ciertos recursos a formar al personal, ya que la varianza de operador claramente no es nula. De todas formas, posiblemente el resultado sería relativamente pobre, ya que parece que tiene un valor comparativamente pequeño respecto al de la varianza residual, por lo que se puede

deducir que contribuye de manera pequeña (pero no nula) a la varianza del diámetro de la partícula.

4. ¿Cómo podemos interpretar la interacción de este problema concreto? ¿Qué criterio decide si realmente hay interacción: los gráficos o el test? (1 punto)

Si la hubiera, la interacción significaría que cada combinación concreta de un operario con un dispositivo podría tener una contribución diferente a la medida de la partícula.

Dado que sería un efecto aleatorio, quedaría representado por la existencia de una cierta varianza σ_{AB}^2 que haría que este efecto no fuera constante (y nulo), sino un valor aleatorio alrededor de cero, que haría que para algunas combinaciones de operario con dispositivo, la medida del punto se incrementará o se reducirá.

El gráfico de interacción hace sospechar que podría haber una cierta interacción, ya que las líneas no son del todo paralelas, pero el resultado del test sería una prueba más definitiva de su posible existencia. A pesar de la impresión que tenemos a partir del gráfico, no podemos descartar que los cruces y los no paralelismos observados no puedan ser debidos simplemente al azar.

5. Si decidiéramos eliminar la interacción del modelo, ¿varían sustancialmente las conclusiones? (1 punto)

Ahora la tabla ANOVA correcta sería:

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | F taules |
|-----------|----|--------|---------|--------------------|--------|-----------------------|
| operator | 2 | 20.72 | 10.36 | 7.91 = 10.36/1.31 | < 0.05 | $F_{0.05}(2,38)=3.24$ |
| resin | 7 | 283.9 | 40.56 | 30.96 = 40.56/1.31 | < 0.05 | $F_{0.05}(7,38)=2.26$ |
| Residuals | 38 | 49.8 | 1.31 | | | |

Por tanto, con las mismas dos primeras hipótesis que hicimos en la pregunta 2, la conclusión sería la misma, en ambos casos se rechaza la hipótesis nula. Las conclusiones no variarían sustancialmente.

6. Hemos supuesto que se verifican las condiciones habituales de normalidad y de homogeneidad de varianzas. Pero, a la vista de los listados y gráficos adjuntos, ¿es cierto? ¿Por qué? (1 punto)

La prueba de Shapiro-Wilks de ajuste a la distribución normal da un p-valor muy pequeño, por debajo de los niveles de significación habituales. Por tanto se puede tomar como una prueba de que la distribución de los residuos no es normal.

Como ocurre con todas estas pruebas, en realidad no dan una idea de si el alejamiento de la normalidad es muy importante, tanto como para alterar seriamente los resultados (en el sentido indicado en la pregunta 8) o no; pero ciertamente hemos detectado un alejamiento de la normalidad. El Q-Q-plot corrobora el resultado anterior, y sugiere que el problema está en unos cuantos valores que parece que discrepan claramente del modelo normal. Quizás estaría bien revisado qué pasó en los experimentos correspondientes.

La prueba de homogeneidad de varianzas de Levene no aporta ninguna prueba de que las varianzas no sean iguales (no rechazamos H_0 de homogeneidad de varianzas ya que el p-valor es $0,8193 > 0,05$), a pesar de que sabiendo esto, no garantiza que realmente lo sean. El gráfico que muestra la dispersión de los residuos sugiere que podría haber operarios con más varianza residual que otros, también convendría revisar estos casos. En resumen, tanto la homoscedasticidad (únicamente a partir del análisis gráfico), como la normalidad, son dudosas.

7. Para un nivel de significación del 0.01, y mediante la prueba HSD de Turkey, indica para qué pares de dispositivos (resin), la medida media de las partículas es realmente diferente. ¿Existe algún o algunos dispositivos de extracción con una medida media más pequeña que los otros? **(1 punto)**

Hay una prueba de diferencia (se rechaza la hipótesis nula de igualdad de medias entre la producción asociada a los correspondientes dispositivos), para las parejas de dispositivos con p-valor (en negrita) inferior a 0.01 (no sería necesario copiar el listado):

| | | | | | |
|-------|------------|----------|-----|-------------|-------------|
| 1 - 3 | 5.8000000 | 0.000000 | *** | 2.97808370 | 8.62191630 |
| 1 - 4 | 6.1833333 | 0.000000 | *** | 3.36141703 | 9.00524963 |
| 1 - 5 | 4.8000000 | 0.000011 | *** | 1.97808370 | 7.62191630 |
| 1 - 6 | 5.4500000 | 0.000001 | *** | 2.62808370 | 8.27191630 |
| 1 - 7 | 2.9166667 | 0.007258 | ** | 0.09475037 | 5.73858297 |
| 2 - 3 | 4.7666667 | 0.000012 | *** | 1.94475037 | 7.58858297 |
| 2 - 4 | 5.1500000 | 0.000004 | *** | 2.32808370 | 7.97191630 |
| 2 - 5 | 3.7666667 | 0.000381 | *** | 0.94475037 | 6.58858297 |
| 2 - 6 | 4.4166667 | 0.000040 | *** | 1.59475037 | 7.23858297 |
| 3 - 7 | -2.8833333 | 0.008127 | ** | -5.70524963 | -0.06141703 |
| 3 - 8 | -5.6166667 | 0.000001 | *** | -8.43858297 | -2.79475037 |

```

4 - 7 -3.2666667 0.002176 ** -6.08858297 -0.44475037
4 - 8 -6.0000000 0.000000 *** -8.82191630 -3.17808370
5 - 8 -4.6166667 0.000021 *** -7.43858297 -1.79475037
6 - 8 -5.2666667 0.000002 *** -8.08858297 -2.44475037

```

A partir del gráfico de medias sospechamos que los dispositivos 3, 4, 5 y 6, están asociados a medias menores. Siempre son significativamente diferentes de las restantes. Pero no podemos afirmar que dentro de estos grupos haya alguno significativamente diferente a los otros. Esto queda muy bien expresado en el listado encabezado con "Groups, Treatments and means" donde se ve que 3, 4, 5 i 6 forman un posible grupo aparte, de medias inferiores al resto.

8. Supongamos que no se verifica la condición de normalidad de los residuos. Vuelve a escribir la tabla ANOVA de la pregunta 2 y tacha todos los valores de la tabla que probablemente han dejado de ser correctos. (1 punto)

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | F taules |
|----------------|----|---------|---------|---------------------|---------|--------------------------|
| operator | 2 | 20.718 | 10.359 | 10.12 = 10.36/1.024 | < 0,05? | $F_{0,05}(2,14) = 3,74$ |
| resin | 7 | 283.946 | 40.564 | 39.61 = 40,56/1.024 | < 0,05? | $F_{0,05}(7,14) = 2,76$ |
| operator:resin | 14 | 14.335 | 1.024 | 0.69 = 1,024/1,478 | > 0,05? | $F_{0,05}(14,24) = 2,13$ |
| Residuals | 24 | 35.480 | 1.478 | | | |

Los grados de libertad, las sumas de cuadrados y los cuadrados medios no tienen nada que ver con la distribución de los residuos. Por lo tanto, los valores de F calculados sobre nuestra serie serían correctos. La tabla original de los listados quedaría igual, no sería necesario tachar nada. La única cosa que afectaría la normalidad sería la distribución del estadístico F. Los valores que ya no serían correctos serían los de las tablas (o los p-valores en una tabla que los mostrara), ya que sería como si miráramos una tabla equivocada. Por lo tanto, tampoco estaríamos seguros de si el p-valor es inferior o no a 0.05.

LISTADOS ANEXOS

```

> pvc.aov <- aov(psi ze ~ operator*resi n)
> summary(pvc.aov)
          Df Sum Sq Mean Sq F value  Pr(>F)
operator    2   20.72    10.36
resin       7  283.95    40.56
operator:resi n 14   14.34     1.02
Residuals  24   35.48     1.48
> dummy.coef(pvc.aov)
Full coefficients are

```

```

(Intercept):      32.35417
operator:         1      2      3
                 0.5895833  0.3270833 -0.9166667
resi n:          1      2      3      4      5
                 3.2958333  2.2625000 -2.5041667 -2.8875000 -1.5041667
operator: resi n: 1:1      2:1      3:1      1:2      2:2
                 0.01041667 -0.57708333  0.56666667 -0.05625000  0.40625000

```

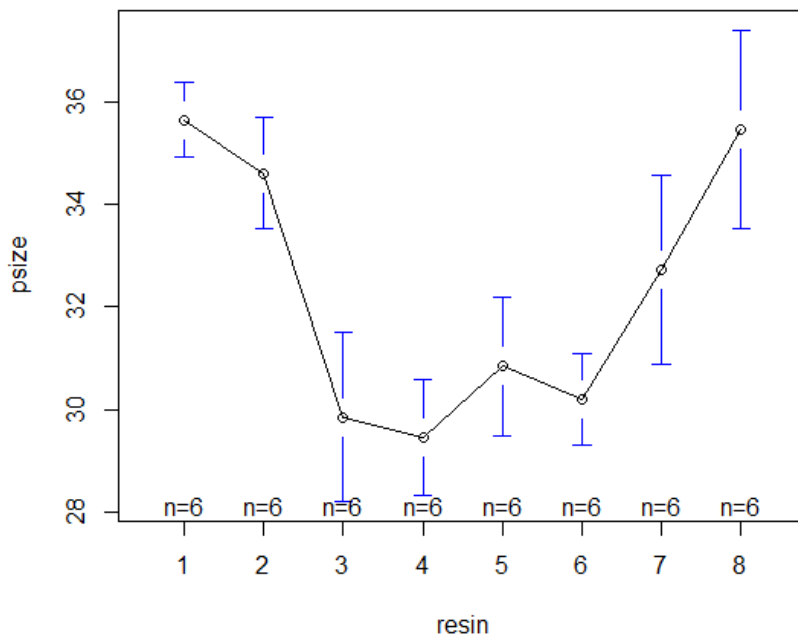
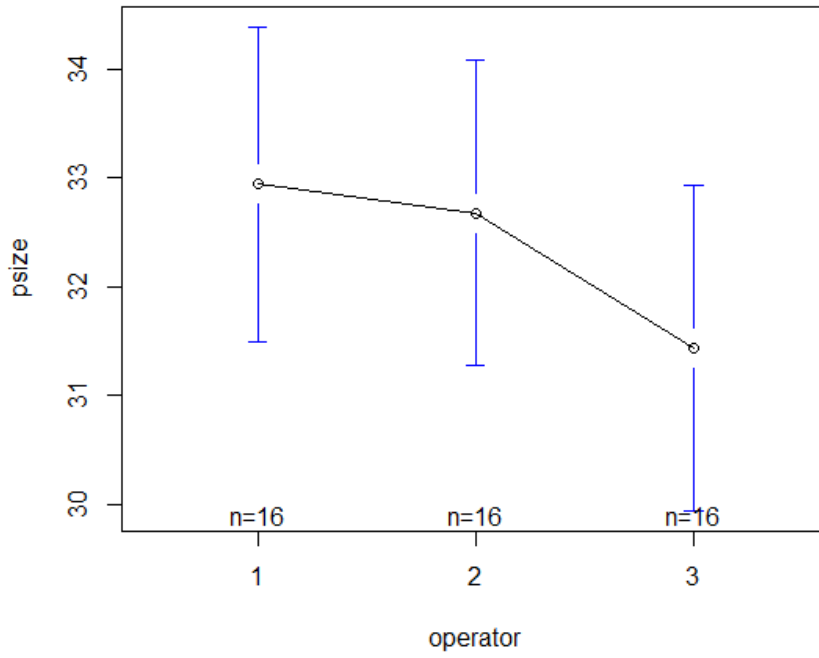
```

> pvc.aov1 <- aov(psi ze ~ operator+resi n)
> summary(pvc.aov1)

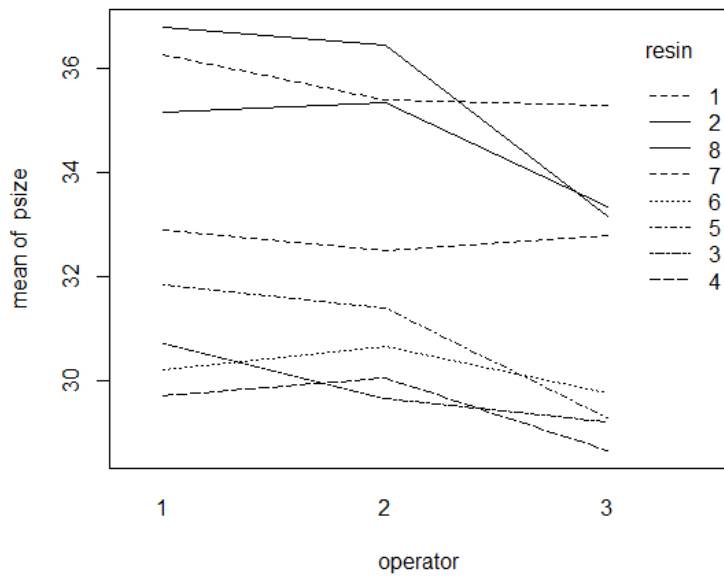
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------|----|--------|---------|---------|--------|
| operator | 2 | 20.72 | 10.36 | | |
| resi n | 7 | 283.95 | 40.56 | | |
| Resi dual s | 38 | 49.82 | 1.31 | | |

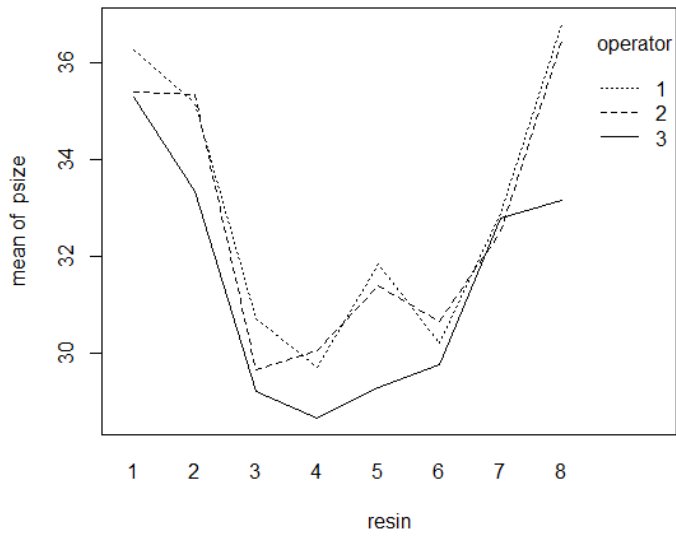
Gráficos de medias y niveles de factor



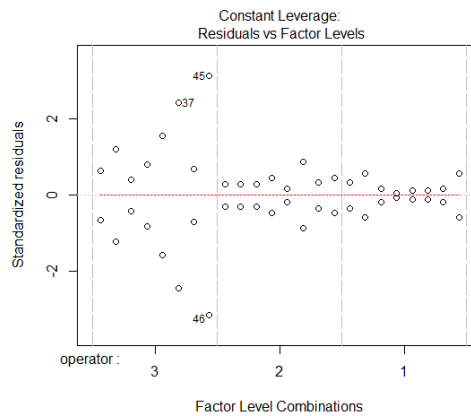
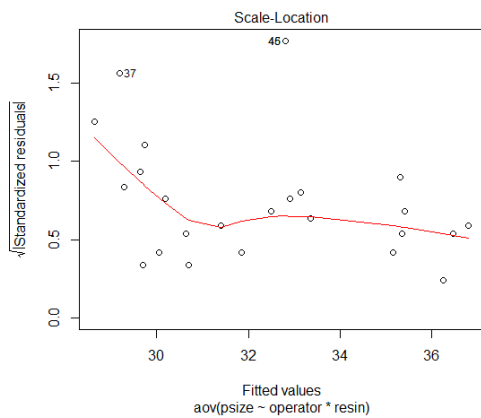
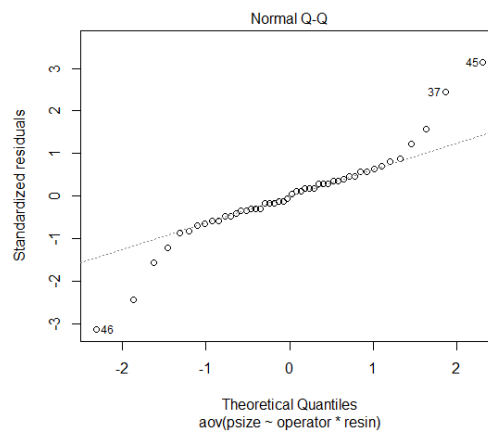
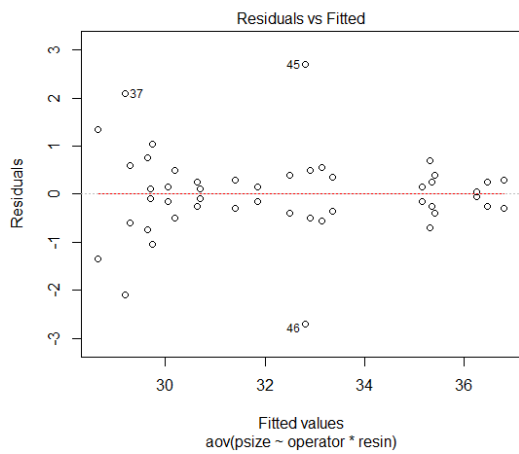
`> interaction.plot(operator, resin, psize)`



```
> interaction.plot(resin, operator, psize)
```



```
> plot(pvc.aov)
```



```
> #normalidad
> Shapiro.test(residuals(pvc.aov))
```

Shapiro-Wilk normality test

data: residuals(pvc.aov)
W = 0.9174, p-value = 0.002401

```
> #homocedasticidad
> levene.test(psize, resin)
```

modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median

data: psize
Test Statistic = 0.5129, p-value = 0.8193

```
> model.tables(pvc.aov, type="mean")
Tables of means
Grand mean
```

32.35417


```
operator
operator
  1    2    3
32.94 32.68 31.44
```

```
resi n
resi n
  1    2    3    4    5    6    7    8
35.65 34.62 29.85 29.47 30.85 30.20 32.73 35.47
```

```
operator: resi n
resi n
operator 1    2    3    4    5    6    7    8
  1 36.25 35.15 30.70 29.70 31.85 30.20 32.90 36.80
  2 35.40 35.35 29.65 30.05 31.40 30.65 32.50 36.45
  3 35.30 33.35 29.20 28.65 29.30 29.75 32.80 33.15
```

```
> require(agricolae)
> HSD.test(pvc.aov, "resi n", group=F, alpha=0.01)
```

Study:

HSD Test for psi ze

Mean Square Error: 1.478333

resi n, means

| | psi ze | std. err | replication |
|---|----------|-----------|-------------|
| 1 | 35.65000 | 0.2825479 | 6 |
| 2 | 34.61667 | 0.4190598 | 6 |
| 3 | 29.85000 | 0.6412228 | 6 |
| 4 | 29.46667 | 0.4409586 | 6 |
| 5 | 30.85000 | 0.5277310 | 6 |
| 6 | 30.20000 | 0.3483293 | 6 |
| 7 | 32.73333 | 0.7204937 | 6 |
| 8 | 35.46667 | 0.7557189 | 6 |

alpha: 0.01 ; Df Error: 24
Critical Value of Studentized Range: 5.685041

Comparison between treatments means

| | Di fference | pval ue | sig | LCL | UCL |
|-------|-------------|----------|-----|-------------|-------------|
| 1 - 2 | 1.0333333 | 0.814239 | | -1.78858297 | 3.85524963 |
| 1 - 3 | 5.8000000 | 0.000000 | *** | 2.97808370 | 8.62191630 |
| 1 - 4 | 6.1833333 | 0.000000 | *** | 3.36141703 | 9.00524963 |
| 1 - 5 | 4.8000000 | 0.000011 | *** | 1.97808370 | 7.62191630 |
| 1 - 6 | 5.4500000 | 0.000001 | *** | 2.62808370 | 8.27191630 |
| 1 - 7 | 2.9166667 | 0.007258 | ** | 0.09475037 | 5.73858297 |
| 1 - 8 | 0.1833333 | 0.999994 | | -2.63858297 | 3.00524963 |
| 2 - 3 | 4.7666667 | 0.000012 | *** | 1.94475037 | 7.58858297 |
| 2 - 4 | 5.1500000 | 0.000004 | *** | 2.32808370 | 7.97191630 |
| 2 - 5 | 3.7666667 | 0.000381 | *** | 0.94475037 | 6.58858297 |
| 2 - 6 | 4.4166667 | 0.000040 | *** | 1.59475037 | 7.23858297 |
| 2 - 7 | 1.8833333 | 0.176093 | | -0.93858297 | 4.70524963 |
| 2 - 8 | -0.8500000 | 0.920796 | | -3.67191630 | 1.97191630 |
| 3 - 4 | 0.3833333 | 0.999207 | | -2.43858297 | 3.20524963 |
| 3 - 5 | -1.0000000 | 0.837212 | | -3.82191630 | 1.82191630 |
| 3 - 6 | -0.3500000 | 0.999561 | | -3.17191630 | 2.47191630 |
| 3 - 7 | -2.8833333 | 0.008127 | ** | -5.70524963 | -0.06141703 |
| 3 - 8 | -5.6166667 | 0.000001 | *** | -8.43858297 | -2.79475037 |
| 4 - 5 | -1.3833333 | 0.520087 | | -4.20524963 | 1.43858297 |
| 4 - 6 | -0.7333333 | 0.962153 | | -3.55524963 | 2.08858297 |
| 4 - 7 | -3.2666667 | 0.002176 | ** | -6.08858297 | -0.44475037 |

```

4 - 8 -6.0000000 0.000000 *** -8.82191630 -3.17808370
5 - 6 0.6500000 0.980295 -2.17191630 3.47191630
5 - 7 -1.8833333 0.176093 -4.70524963 0.93858297
5 - 8 -4.6166667 0.000021 *** -7.43858297 -1.79475037
6 - 7 -2.5333333 0.025919 * -5.35524963 0.28858297
6 - 8 -5.2666667 0.000002 *** -8.08858297 -2.44475037
7 - 8 -2.7333333 0.013452 * -5.55524963 0.08858297
> HSD.test(pvc.aov, "resi n", group=T, alpha=0.01)

```

Study:

HSD Test for psi ze

Mean Square Error: 1.478333

resi n, means

| | psi ze | std. err | replication |
|---|----------|-----------|-------------|
| 1 | 35.65000 | 0.2825479 | 6 |
| 2 | 34.61667 | 0.4190598 | 6 |
| 3 | 29.85000 | 0.6412228 | 6 |
| 4 | 29.46667 | 0.4409586 | 6 |
| 5 | 30.85000 | 0.5277310 | 6 |
| 6 | 30.20000 | 0.3483293 | 6 |
| 7 | 32.73333 | 0.7204937 | 6 |
| 8 | 35.46667 | 0.7557189 | 6 |

alpha: 0.01 ; Df Error: 24

Critical Value of Studentized Range: 5.685041

Honestly Significant Difference: 2.821916

Means with the same letter are not significantly different.

| Groups, | Treatments | and means |
|---------|------------|-----------|
| a | 1 | 35.65 |
| ab | 8 | 35.47 |
| ab | 2 | 34.62 |
| bc | 7 | 32.73 |
| cd | 5 | 30.85 |
| cd | 6 | 30.2 |
| d | 3 | 29.85 |
| d | 4 | 29.47 |

3.2. Influencia del genotipo en la actividad del enzima Manosa-6-fosfat isomerasa (MPI) en los crustáceos anfípodos *Platorchestia platensis*

| Subject | Sex | Genotype | Activity |
|---------|--------|----------|----------|
| 1 | male | ff | 1.884 |
| 2 | male | ff | 2.283 |
| 3 | male | fs | 2.396 |
| 4 | female | ff | 2.838 |
| 5 | male | fs | 2.956 |
| 6 | female | ff | 4.216 |
| 7 | female | ss | 3.620 |
| 10 | male | fs | 3.105 |
| 12 | female | fs | 3.087 |
| 13 | male | ff | 4.939 |
| 14 | male | ff | 3.486 |
| 16 | male | fs | 2.649 |
| 17 | female | fs | 1.943 |
| 19 | female | ff | 4.198 |
| 24 | female | fs | 2.200 |
| 26 | male | ss | 4.801 |
| 28 | male | ss | 4.921 |
| 30 | female | fs | 4.281 |
| 34 | female | ss | 3.586 |
| 36 | female | ff | 3.944 |
| 41 | male | ss | 5.275 |
| 43 | female | ss | 2.963 |
| 46 | female | ss | 3.236 |
| 49 | male | ss | 4.110 |

Problema 1. En un estudio se quería determinar si el genotipo podría influir en la actividad del enzima Manosa-6-fosfat isomerasa (MPI) en los crustáceos anfípodos *Platorchestia platensis*. En estos crustáceos, el locus MPI puede presentar tres tipos de genotipo posibles: Mpi^{ff} , Mpi^{fs} , i Mpi^{ss} . El estudio se basó en una muestra aleatoria de 24 individuos (no hagáis caso de la numeración de los individuos, en total son 24). Dado que se sospechaba que el dimorfismo sexual que presenta esta especie podía ser importante, se tuvo en cuenta el sexo, y se construyó una muestra balanceada de 12 machos y 12 hembras, con los tres genotipos representados por igual dentro de cada sexo (4 de cada genotipo por cada sexo). Cada individuo fue liofilizado, pesado, homogeneizado, y se determinó la actividad de MPI de la fracción soluble. Los datos adjuntos indican que el sexo, el genotipo, y la actividad enzimática (en unidades de $\Delta O.D./segundo/mg$ peso seco) de cada individuo. A partir de los listados R adjuntos, cuando sea necesario, responde a las siguientes cuestiones. En todo momento puedes utilizar un nivel de significación de 0.05 o de confianza de 0.95, y dar por válidos los supuestos habituales de normalidad y de homogeneidad de varianzas. **Responde en las mismas hojas del examen. Todas las preguntas valen 1 punto.**

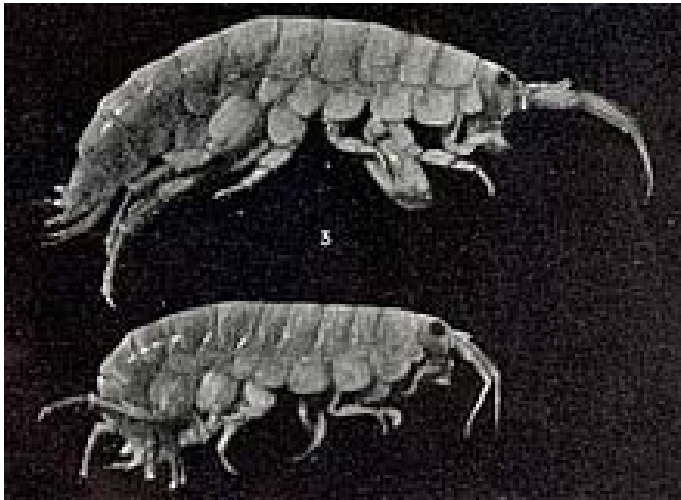


Imagen de *Platorchestia platensis* (Procedente de https://upload.wikimedia.org/wikipedia/commons/thumb/b/bf/FMIB_42836_Orchestia_agalis_Upper_figure_male%2C_lower_female_Woods_Hole.jpeg/220px-FMIB_42836_Orchestia_agalis_Upper_figure_male%2C_lower_female_Woods_Hole.jpeg)

- 1) Describe el tipo de diseño, los factores que intervienen y sus niveles, si son fijos o aleatorios, y sus relaciones (cruzado, jerárquico). Escribe el modelo lineal asociado y su parametrización. Indica razonadamente, si se trata de un estudio experimental totalmente aleatorizado o no (tienes más espacio para responder en la página siguiente).
- 2) Copia y completa aquí la tabla ANOVA adecuada y determina la significación de todos los efectos del modelo, indicando en cada caso la hipótesis nula y alternativa, el estadístico de test utilizado y su valor, los grados de libertad, el valor crítico en las tablas y la conclusión final.
- 3) ¿Tiene sentido hacer las pruebas de comparación múltiple para establecer qué genotipos tienen medias de actividad diferentes? En caso afirmativo, indica razonadamente qué sería mejor, comparar los genotipos utilizando todos los datos, sin distinción de sexos, o hacer las comparaciones por separado dentro de cada sexo. Según la opción que hayas escogido, indica qué genotipos son significativamente diferentes según la prueba HSD de Tukey.
- 4) Estima la probabilidad de que para un macho con genotipo M_{pi}^{ff} se obtenga una actividad enzimática superior a 5.
- 5) Supón que los tres genotipos estudiados no son los tres únicos posibles, sino que este locus presenta una gran variabilidad genética y que, frente a la

imposibilidad de estudiar todos los genotipos posibles, los tres genotipos de los datos sólo son una muestra aleatoria de los muchos genotipos posibles. (Es falso, pero supongamos que la situación real es ésta). Sin tener que repetir todas tus respuestas y cálculos, indica qué ha cambiado (características del diseño, hipótesis, cálculos, resultados, conclusiones...) respecto de lo que has respondido a las preguntas 1), 2) i 3).

- 6) En base a las nuevas condiciones planteadas en la pregunta anterior, estima la probabilidad de que para un macho se obtenga una actividad enzimática superior a 5. ¿Por qué razón no tiene sentido pedir esta probabilidad para un macho "con genotipo Mpiff"?
- 7) Volvamos a la situación inicial, preguntas 1), 2), 3) i 4). No ocurre en esta especie, pero supón que el locus MPI estuviera ligado al sexo, de manera que los genotipos posibles no fueran los mismos en hembras que en machos. Las hembras podrían representar los tres genotipos indicados: Mpiff, Mpifs, y Mpiss, pero en cambio los machos sólo dos genotipos: Mpif- i Mpis- por la ausencia del gen MPI en un hipotético "cromosoma Y". Indica qué cosas cambiarían respecto de lo que has respondido en la pregunta 1) tipo de diseño, los factores que intervienen y sus niveles, si son fijos o aleatorios, y sus relaciones (cruzado, jerárquico). Escribe el nuevo modelo lineal asociado y su parametrización (atención, aquí no se pide hacer el cálculo de la tabla ANOVA ni de nada de eso, sólo describir el diseño, los favores, etc.).

LISTADOS DEL PROBLEMA

```
> mpi.aov = aov(Acti vity ~ Sex * Genotype, data = mpi)
> anova(mpi.aov)
Response: Acti vity
      Df Sum Sq Mean Sq F value Pr(>F)
Sex      1  0.3022      0.3022  0.000000e+00 ***
Genotype 2  6.1236      3.0618  0.000000e+00 ***
Sex:Genotype  4  4.6300      1.1575  0.000000e+00 ***
Residuals 11 11.5982      1.0544
```

```
> model.tables(mpi.aov, type = "mean")
```

```
Tables of means
Grand mean
```

```
3.454875
```

```
Sex
Sex
female  male
3.343   3.567
```

```
Genotype
Genotype
ff      fs      ss
```

3.474 2.827 4.064

```
Sex: Genotype
      Genotype
Sex    ff    fs    ss
female 3.799 2.878 3.351
male   3.148 2.777 4.777
```

```
> TukeyHSD(mpi.aov, which = "Genotype")
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = Activity ~ Sex * Genotype, data = mpi)
```

```
$Genotype
      diff      lwr      upr      p adj
fs-ff -0.646375 -1.6706997 0.3779497 0.2669097
ss-ff  0.590500 -0.4338247 1.6148247 0.3275667
ss-fs  1.236875  0.2125503 2.2611997 0.0168342
```

```
> mpi.female.aov = aov(Activity ~ Genotype, data = subset(mpi, Sex == "female"))
> TukeyHSD(mpi.female.aov, which = "Genotype")
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = Activity ~ Genotype, data = subset(mpi, Sex == "female"))
```

```
$Genotype
      diff      lwr      upr      p adj
fs-ff -0.92125 -2.3802031 0.5377031 0.2357315
ss-ff -0.44775 -1.9067031 1.0112031 0.6791113
ss-fs  0.47350 -0.9854531 1.9324531 0.6501461
```

```
> mpi.male.aov = aov(Activity ~ Genotype, data = subset(mpi, Sex == "male"))
> TukeyHSD(mpi.male.aov, which = "Genotype")
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = Activity ~ Genotype, data = subset(mpi, Sex == "male"))
```

```
$Genotype
      diff      lwr      upr      p adj
fs-ff -0.37150 -2.07276867 1.329769 0.8185456
ss-ff  1.62875 -0.07251867 3.330019 0.0601893
ss-fs  2.00025  0.29898133 3.701519 0.0233045
```

3.3. Influencia de las citosinas en la respuesta inmunitaria de los animales de granja.

| Solución | Granja | Compt.prec |
|----------|--------|------------|
| S1 | G1 | 45 |
| S1 | G1 | 47 |
| S1 | G1 | 39 |
| S1 | G2 | 47 |
| S1 | G2 | 51 |
| S1 | G2 | 39 |
| S1 | G3 | 40 |
| S1 | G3 | 42 |

| | | |
|----|----|----|
| S1 | G3 | 35 |
| S2 | G1 | 52 |
| S2 | G1 | 52 |
| S2 | G1 | 53 |
| S2 | G2 | 54 |
| S2 | G2 | 49 |
| S2 | G2 | 49 |
| S2 | G3 | 48 |
| S2 | G3 | 51 |
| S2 | G3 | 52 |
| S3 | G1 | 66 |
| S3 | G1 | 64 |
| S3 | G1 | 62 |
| S3 | G2 | 61 |
| S3 | G2 | 58 |
| S3 | G2 | 57 |
| S3 | G3 | 56 |
| S3 | G3 | 59 |
| S3 | G3 | 59 |

Problema 1. Las interleucinas o citosinas son proteínas de bajo peso molecular esenciales en la intercomunicación celular, y en particular en la respuesta inmunitaria. Algunas de estas proteínas pueden ser inmunodepresoras y otras, en cambio, pueden ser activadoras de la respuesta inmunitaria. En un experimento realizado sobre animales de granja, se quería estudiar el efecto de tres soluciones proteínicas (S1, S2 y S2), sobre la respuesta inmunitaria. S1 contenía una citosina considerada, en principio, inmunodepresora, mientras que S3 debería de ser activadora. La respuesta a estos tratamientos se medirá en forma de un conteo de células precursoras de los linfocitos (variable "Compt.prec"). Se sospechaba que la respuesta podía variar según la procedencia de los animales, de manera que se seleccionaron al azar 3 granjas, G1, G2 y G3, de entre muchas posibles. De cada una de las granjas, se cogieron una muestra aleatoria de 9 animales (por tanto se trabajó con una muestra total de 27 animales), que se asignaron al azar y de una manera balanceada a cada uno de los tratamientos.



(Imagen de <https://navarra.elespanol.com/media/navarra/images/2017/02/01/2017020119535086795.jpg>)

Los resultados del experimento se indican en la tabla adjunta. Cada fila de la tabla corresponde a un de estos animales e indica la solución que le fue administrada, su granja de procedencia y el recuento de células precursoras observado.

Utilizando, cuando sea necesario, los listados adjuntos, responde a las siguientes cuestiones. En todo momento puedes asumir normalidad y heterogeneidad de varianzas. Utiliza un nivel de significación de 0,05 o de confianza de 0,95.

- 1) **Describe el tipo de diseño, los factores que intervienen, sus niveles y si son fijos o aleatorios. Escribe el modelo lineal asociado y su parametrización. Indica, razonadamente, si se trata de un estudio experimental totalmente aleatorizado o no (tienes más espacio para responder en la página siguiente). (1.5 puntos)**

Es un diseño de 2 factores cruzados, balanceado, con $n=3$ réplicas por condición experimental. Los factores son "Solución" (o tratamiento), fijo, con $a=3$ niveles, S_1 , S_2 y S_3 ; y "Granja", aleatorio, con $b=3$ niveles, G_1 , G_2 y G_3 .

No consideramos el "Modelo restringido" (si alguno lo hace no hay ningún problema, siempre que lo diga claramente y que lo haga coherentemente con este enfoque). El modelo lineal se podría indicar de la siguiente forma:

Comptatge per l'animal k , dels tractats amb i que procedien de la granja j

↓ efecte fix de la solució i

↓ ↓ ↙ interacció aleatòria entre solució i i procedir de la granja j

$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + e_{ijk}$$

← residu aleatori del comptatge per l'animal k tractat amb i , procedent de la granja j

↑ ↑ ↑

↑ efecte aleatori del fet de procedir de la granja j

mitjana general o constant

$i = 1, 2, 3$ (o bé $S1, S2, S3$); $a = 3$

$j = 1, 2, 3$ (o bé $G1, G2, G3$); $b = 3$

$k = 1, 2, 3$; $n = 3$

$$\sum_{i=1}^a \alpha_i = 0$$

$$\left. \begin{array}{l} B_j \sim N(0, \sigma_{\text{granja}}) \text{ iid} \\ (\alpha B)_{ij} \sim N(0, \sigma_{\text{Solució:Granja}}) \text{ iid} \\ e_{ijk} \sim N(0, \sigma) \text{ iid} \end{array} \right\} \text{ independents}$$

La parametrización, teniendo en cuenta la restricción sobre los efectos de "Solución", sería: $\mu, \alpha_1, \alpha_2, \sigma_{\text{Solució}}, \sigma_{\text{Solució:Granja}}$ y σ (o bien las correspondientes varianzas en lugar de desviaciones típicas). Un parámetro más, α_3 , si no consideráramos la restricción sobre la suma de los α_i , también buena respuesta.

Únicamente se aleatoriza por "Solución", los animales ya llevan "etiqueta" de la granja de procedencia, sin que los experimentadores la puedan asignar al azar. No confundamos "aleatorizar", que tiene relación con que asignamos los niveles de los factores de los sujetos que participan en el estudio (una vez ya obtenida la muestra), con el

hecho de que “Granja” sea un factor aleatorio, o que los animales puedan ser una muestra aleatoria de las correspondientes granjas, que habían estado escogidas al azar, son conceptos diferentes.

2) Completa la siguiente tabla ANOVA y determina la significación de todos los efectos del modelo, indicando en cada caso la hipótesis nula y alternativa, el test estadístico utilizado y su valor, los grados de libertad, el valor crítico de las tablas, y la conclusión final (tienes más espacio para responder en la página siguiente)(1.5 puntos)

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | F tablas |
|-------------------|----|---------|---------|---------|--------|-------------------------|
| Soluci on | 2 | 1370.30 | 685.15 | 45.23 | < 0.05 | $F_{0.05}(2,4) = 6.94$ |
| Granja | 2 | 81.41 | 40.70 | 2.69 | > 0.05 | $F_{0.05}(2,4) = 6.94$ |
| Soluci on: Granja | 4 | 60.59 | 15.15 | 1.48 | > 0.05 | $F_{0.05}(4,18) = 2.93$ |
| Resi duos | 18 | 184.00 | 10.22 | | | |

Dado que no hemos considerado el modelo restringido, los coeficientes R para el efecto de “Solución” y de “Granja” los hemos formado dividiendo por el cuadrado medio de interacción. De acuerdo con esto y con las hipótesis consideras, tenemos:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad \text{vs.} \quad H_1 : \alpha_i \neq \alpha_{i'} \text{ para algún } i \neq i'$$

$F = MS_{\text{Solución}} / MS_{\text{Solución:Granja}} = 45,23$. Con 2 y 4 g.d.l. el valor crítico F en las tablas para un nivel de significación 0,05 es 6,94, de manera que rechazamos H_0 , hemos obtenidos evidencias de que la solución influye en la respuesta.

$$H_0 : \sigma_{\text{Solució}}^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_{\text{Solució}}^2 > 0$$

$F = MS_{\text{Granja}} / MS_{\text{Solución:Granja}} = 2,69$. Con 2 y 4 g.d.l. el valor crítico F es 6,94, de manera que no podemos rechazar la hipótesis nula, no hemos obtenido evidencias de que haya una parte de la varianza de la respuesta atribuible a la granja de origen

$$H_0 : \sigma_{\text{Solució:Granja}}^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_{\text{Solució:Granja}}^2 > 0$$

$F = MS_{\text{Solución:Granja}} / MS_{\text{residual}} = 1,48$. Con 4 y 18 g.d.l. el valor crítico F es 2,93, de manera que no podemos rechazar la hipótesis nula, no hemos obtenido evidencias que haya una parte de la varianza de la respuesta atribuible a una posible interacción entre la solución aplicada y la granja de origen.

- 3) **Para los factores para los cuales tenga sentido, realiza un análisis “post hoc” (comparaciones múltiples) mediante la prueba de HSD de Tukey e indica cuáles son las conclusiones de esta prueba. ¿Los resultados de los apartados 2 y 3 confirman las ideas iniciales expresadas al principio del enunciado? (1 punto)**

Únicamente tiene sentido realizar las comparaciones múltiples para el factor “Solución”, que es fijo y se ha rechazado la correspondiente hipótesis nula en la prueba ANOVA. Como podemos ver en el listado:

```
> TukeyHSD(interleucines.aov, which = "Solucion")
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = Compt.prec ~ Solucion * Granja, data = interleucines)
```

```
$Solucion
      diff      lwr      upr      p adj
S2-S1  8.333333  4.486753 12.17991 8.51e-05
S3-S1 17.444444 13.597864 21.29103 0.00e+00
S3-S2  9.111111  5.264531 12.95769 2.93e-05
```

Todas las hipótesis nulas referentes a las medias para los niveles de este factor, $H_0: \mu_j = \mu_k$, se pueden rechazar al ser todos los p-valores inferiores a 0,05.

Únicamente se confirma la idea sobre el efecto de las diferentes soluciones, también corroborada por las diferencias de medias anteriores (o por el listado de medias), por el cual la respuesta media bajo S1 es inferior a la correspondiente bajo S2, que queda superada por la respuesta media bajo S3.

- 4) **Estima todos los parámetros del modelo e indica, razonadamente, cuál es la condición experimental que más favorece la producción de las células precursoras(1.5 puntos)**

$$\hat{\mu} = \bar{Y}_{...} = 51,37$$

$$\hat{\alpha}_1 = \bar{Y}_{1..} - \bar{Y}_{...} = 42,78 - 51,37 = -8,59$$

$$\hat{\alpha}_2 = \bar{Y}_{2..} - \bar{Y}_{...} = 51,11 - 51,37 = -0,26$$

$$(\hat{\alpha}_3 = -(\hat{\alpha}_1 + \hat{\alpha}_2) = 8,85)$$

$$\sigma_{\text{Granja}}^2 = \frac{MS_{\text{Granja}} - MS_{\text{Solució:Granja}}}{an} = \frac{40,70 - 15,15}{3 \times 3} = 2,84$$

$$\sigma_{\text{Solució:Granja}}^2 = \frac{MS_{\text{Solució:Granja}} - MS_{\text{residual}}}{n} = \frac{15,15 - 10,22}{3} = 1,64$$

$$\sigma_{\text{residual}}^2 = MS_{\text{residual}} = 10,22$$

Tal y como puede pasar perfectamente, a pesar de que el análisis ANOVA no ha rechazado las correspondientes hipótesis nulas, la varianza estimada de “Granja” y de interacción no son nulas, si bien son pequeñas en comparación con la residual.

(Está mejor tal y como se han resuelto los cálculos anteriores -ya que no rechazar H_0 no es una demostración de que sea cierta-, pero también sería justificable considerar nulas tanto la estimación de la varianza de “Granja” como la de interacción, es decir, optar por un modelo más simple.)

Dado que la “Granja” es un factor aleatorio, y cualquier afirmación sobre la condición óptima se ha de basar únicamente en el factor fijo “Solución”. Observando la tabla de medias, podemos atribuir esta condición a S3, y a partir de los resultados anteriores descartar que las diferencias de sean atribuibles al azar. No tienen sentido consideraciones sobre si lo óptimo sería combinar S3 con tal o cual granja, ya que estas granjas son únicamente unas representantes aleatorias de otras muchas posibles granjas.

- 5) **Calcula la probabilidad de que un animal que ha sido tractorado con S3 presente un recuento de células precursoras superior a 60. (1.5 puntos)**

La media de la respuesta de los animales tratados con S3 es $\mu + \alpha_3$, parámetro que se puede estimar directamente con la media muestral

de todas las observaciones Y_{S3} . (9 datos) de los animales que han recibido S3: 60,22. Esta variable tiene varianza igual a la suma de las varianzas de todos los efectos aleatorios, es decir, "Granja", interacción y residual. Por tanto su varianza se puede estimar como $2,84 + 1,64 + 10,22 = 14,70$, y su desviación típica como 3,83. Aproximadamente, $Y_{S3} \sim N(60,22; 3,83)$. La probabilidad que se pide es $\Pr\{Y_{S3} > 60\}$. Dado que debemos consultar la tabla de la $N(0,1)$, estandarizaremos esta variable, $Z = (Y_{S3} - 60,22) / 3,83$:

$$\Pr\{Y_{S3} > 60\} = \Pr\{Z > (60 - 60,22) / 3,83\} = \Pr\{Z > -0,057\} = \Pr\{Z < 0,057\} = 0,523, \text{ donde } Z \sim N(0,1).$$

(La varianza se ha estimado tal y como está hecho en el párrafo anterior, únicamente es justificable considerar que es 10,22 –es decir, únicamente la residual– si en la pregunta anterior se ha razonado que a la varianza estimada de Granja y de interacción se les daba el valor 0, y aquí se ha hecho servir este hecho. El resultado final es muy similar, pero lo importante es que se haya razonado correctamente).

LISTADOS

```
> model.tables(interfucines.aov, type = "means")
Tables of means
Grand mean

51.37037

Solucio
Solucio
  S1    S2    S3
42.78 51.11 60.22

Granja
Granja
  G1    G2    G3
53.33 51.67 49.11

Solucio: Granja
Granja
Solucio G1    G2    G3
  S1 43.67 45.67 39.00
  S2 52.33 50.67 50.33
  S3 64.00 58.67 58.00
> TukeyHSD(interfucines.aov, which = "Solucio")
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = Compt.prec ~ Solucio * Granja, data = interleucines)
```

```
$Solucio
```

| | diff | lwr | upr | p adj |
|-------|----------|----------|----------|----------|
| S2-S1 | 8.333333 | 4.486753 | 12.17991 | 8.51e-05 |
| S3-S1 | 17.44444 | 13.59786 | 21.29103 | 0.00e+00 |
| S3-S2 | 9.111111 | 5.264531 | 12.95769 | 2.93e-05 |

```
> TukeyHSD(interleucines.aov, which = "Granja")
```

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

```
Fit: aov(formula = Compt.prec ~ Solucio * Granja, data = interleucines)
```

```
$Granja
```

| | diff | lwr | upr | p adj |
|-------|-----------|-----------|------------|-----------|
| G2-G1 | -1.666667 | -5.513247 | 2.1799139 | 0.5228421 |
| G3-G1 | -4.222222 | -8.068803 | -0.3756416 | 0.0302169 |
| G3-G2 | -2.555556 | -6.402136 | 1.2910250 | 0.2339788 |

3.4. Estudio in vivo sobre el efecto de las cicinas en la expansión de las células progenitoras del CD34+.

| subject | cocktail | age | count |
|---------|----------|--------|-------|
| 1 | control | infant | 45 |
| 2 | control | infant | 47 |
| 3 | control | infant | 39 |
| 4 | control | infant | 47 |
| 5 | control | infant | 51 |
| 6 | control | adult | 39 |
| 7 | control | adult | 40 |
| 8 | control | adult | 42 |
| 9 | control | adult | 35 |
| 10 | control | adult | 38 |
| 11 | tcc | infant | 52 |
| 12 | tcc | infant | 52 |
| 13 | tcc | infant | 53 |
| 14 | tcc | infant | 54 |
| 15 | tcc | infant | 49 |
| 16 | tcc | adult | 49 |
| 17 | tcc | adult | 48 |
| 18 | tcc | adult | 51 |
| 19 | tcc | adult | 52 |
| 20 | tcc | adult | 51 |
| 21 | acd3s | infant | 66 |
| 22 | acd3s | infant | 64 |
| 23 | acd3s | infant | 62 |
| 24 | acd3s | infant | 61 |
| 25 | acd3s | infant | 58 |
| 26 | acd3s | adult | 57 |
| 27 | acd3s | adult | 56 |
| 28 | acd3s | adult | 59 |
| 29 | acd3s | adult | 59 |
| 30 | acd3s | adult | 60 |

Problema 1. En la Universidad de California, Davis, se diseñó un estudio in vivo sobre el efecto de las cicinas en la expansión de las células progenitoras CD34+. Con este fin, se prepararon 3 “cócteles cicina”. El primero era el “traditional cytokine cocktail” (ctt), el segundo contenía un agente anti-CD3 (acd3s) y el tercero era un placebo (control). Se seleccionó una muestra aleatoria de 30 monos *Rhesus macaque*, del mismo sexo, con la restricción que 15 fueran individuos maduros (adult) y 15 fueran individuos juveniles (infant). Tanto los individuos maduros como los juveniles fueron asignados al azar y de forma balanceada a recibir uno de los cócteles de cicina. 10 días después de la administración, a cada individuo se le extrajo una muestra de médula ósea y se determinó in vitro la expresión de las células CD34+ mediante citometría de flujo, contando el número de colonias producidas. Los resultados del experimento se muestran en la tabla adjunta.

Utilizando, cuando sea necesario, los listados correspondientes, responde a las siguientes cuestiones. En todo momento puedes asumir normalidad y homogeneidad de varianzas. Utiliza

un nivel de significación de 0,05 o de confianza de 0,95.

- 1) **Describe el tipo de diseño, los factores que intervienen, sus niveles y si son fijos o aleatorios. Escribe el modelo lineal asociado y su parametrización. Indica, razonadamente, si se trata de un estudio experimental totalmente aleatorizado o no (tienes más espacio para responder en la página siguiente). (1 punto)**

Es un diseño cruzado de dos factores, con réplicas, balanceado. Utilizando los nombres de la tabla de datos, un factor es “cocktail” con $a = 3$ niveles: “control”, “tcc” y “acd3s”. Otro factor es “age” con $b = 2$ niveles: “infant” y “adult”. El número de réplicas por casillas o condición experimental es $n = 5$. Ambos factores son fijos

ya que tanto cócteles como la condición de ser infant o adult no son una muestra aleatoria de muchos posibles niveles, sino que están fijados y preespecificados: en un futuro experimento, sin problema se pueden volver a repetir estos niveles. No se debe confundir con que los cóctels se asignen al azar, esto es "aleatorización", no significa que se trate de factores aleatorios.

No es un diseño totalmente aleatorizado, ya que los niveles del factor "age" se observan, pero no se pueden asignar al azar a los sujetos.

Sólo se puede aleatorizar el tratamiento o factor "cocktail". "age" más bien nos permite formar "bloques", para hacer comparaciones entre cócteles en condiciones más homogéneas, ante la sospecha de que la edad pueda influir en la respuesta.

El modelo lineal asociado sería:

valor observat de "count" pel subjecte (rèplica) k del "cocktail" i , "age" j
 \downarrow mitjana general o constant
 \downarrow \downarrow residu pel subjecte k del "cocktail" i , "age" j
 \downarrow \downarrow \downarrow
 $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$
 \uparrow \uparrow \uparrow
 \uparrow \uparrow interacció entre el nivell i de "cocktail" i el nivell j d'"age"
 \uparrow efecte del nivell j del factor "age"
 efecte del nivell i del factor "cocktail"

$$\sum_{i=1}^a \alpha_i = 0, \sum_{j=1}^b \beta_j = 0,$$

$$\sum_{j=1}^b (\alpha\beta)_{ij} = 0, \text{ per tot } i = 1, \dots, a,$$

$$\sum_{i=1}^a (\alpha\beta)_{ij} = 0, \text{ per tot } j = 1, \dots, b.$$

$$e_{ijk} \sim N(0, \sigma) \text{ iid}$$

Los parámetros del modelo son 13: $\mu, \alpha_{acd3s}, \alpha_{control}, \alpha_{tcc}, \beta_{adult}, \beta_{infant},$

$(\alpha\beta)_{acd\bar{3}s,adult}, (\alpha\beta)_{acd\bar{3}s,infant}, (\alpha\beta)_{control,adult}, (\alpha\beta)_{control,infant}, (\alpha\beta)_{tcc,adult}, (\alpha\beta)_{tcc,infant}$ i
 σ^2 . (O bien 7 parámetros si se consideren las restricciones.)

- 2) Copia y completa aquí la tabla ANOVA adecuada, y determina la significación de todos los efectos del modelo, indicando en cada caso la hipótesis nula y alternativa, el test estadístico utilizado y su valor, los grados de libertad, el valor crítico de las tablas, y la conclusión final. (1 punto)

Copiaremos y completaremos la tabla del LISTADO 3:

| | Df | Sum Sq | Mean Sq | F value | F taules | Pr(>F) |
|--------------|----|---------|---------|--------------------|-----------------------|--------|
| cocktail | 2 | 1602.20 | 801.10 | 801.10/7.33=109.24 | $F_{0,05}(2,24)=3.40$ | < 0.05 |
| age | 1 | 136.53 | 136.53 | 136.53/7.33=18.61 | $F_{0,05}(1,24)=4.26$ | < 0.05 |
| cocktail:age | 2 | 34.07 | 17.03 | 17.03/7.33=2.32 | $F_{0,05}(2,24)=3.40$ | > 0.05 |
| Residuals | 24 | 176.00 | 7.33 | | | |

$H_0 : \alpha_1 = \dots = \alpha_a = 0$ vs $H_1 : \alpha_i \neq 0$, per algun $i = 1, \dots, a$

El estadístico $F = 109,24$ es mayor que el estadístico 3,40 para el nivel establecido (0,05) y 2 y 24 g.d.l. Por lo tanto rechazamos H_0 , descartamos que el efecto de los diversos cócteles sea nulo e igual para todos.

$H_0 : \beta_1 = \dots = \beta_b = 0$ vs $H_1 : \beta_j \neq 0$, per algun $j = 1, \dots, b$

El estadístico $F = 18,61$ es mayor que el valor crítico 4,26 para 1 y 24 g.d.l. Rechazamos H_0 , descartamos un efecto nulo de los bloques de edad.

$H_0 : (\alpha\beta)_{11} = \dots = (\alpha\beta)_{ab} = 0$ vs $H_1 : (\alpha\beta)_{ij} \neq 0$,
per algun $i, j : i = 1, \dots, a, j = 1, \dots, b$

El estadístico $F = 2,32$ es menor que el valor crítico 3,40 para 2 y 24 g.d.l. No podemos rechazar H_0 , no tenemos evidencia de interacción entre cóctel y edad.

- 3) Los investigadores podrían haber dudado entre las dos estrategias para hacer un análisis "post hoc" (comparaciones múltiples) de los diferentes cócteles:

- a. Globalmente, para todos los datos juntos, tal y como se hace en el listado 4.
- b. Sólo para los individuos maduros y, por separado, solo para los juveniles.

Indica por qué razón o razones la estrategia “a” parece adecuada, y resume las conclusiones que se extraen. **(0.5 puntos)**

La estrategia de comparación múltiple “a” parece adecuada ya que no hemos detectado que haya interacción. A pesar de que no tenemos una total certeza de ausencia de interacción, en general se admite que comparar los niveles del factor que nos interesa a partir de todos los datos es posible, percepción reforzada paralelamente por el aspecto de las líneas en el gráfico de interacción. Si la interacción fuera más idónea para comparar por separado dentro de cada bloque o nivel de edad.

El listado resume el resultado final de haber realizado 3 comparaciones 2 a 2:

$$\begin{aligned}
 H_0 : \mu_{control} &= \mu_{acd3s} \quad VS \quad H_1 : \mu_{control} \neq \mu_{acd3s} \\
 H_0 : \mu_{tcc} &= \mu_{acd3s} \quad VS \quad H_1 : \mu_{tcc} \neq \mu_{acd3s} \\
 H_0 : \mu_{control} &= \mu_{tcc} \quad VS \quad H_1 : \mu_{control} \neq \mu_{tcc}
 \end{aligned}$$

En los tres casos el p-valor ajustado ha sido inferior al nivel 0,05 o, en otras palabras, el intercalo de confianza 95% simultáneo para las tres posibles diferencias de medias no incluye el valor 0, descarta la igualdad. Por lo tanto, todos los niveles del factor se pueden considerar diferentes en cuanto a la media (poblacional) de la variable de respuesta, con el riesgo de cometer un error de tipo I de 0,05 (aproximadamente).

- 4) Estima todos los parámetros del modelo e indica, razonadamente, cuál es la condición experimental que más favorece la expansión de las células CD34+. **(1 punto)**

Utilizaremos el listado 6.

La estimación de la constante es $\hat{\mu} = \bar{y}_{...} = 51,2$. Respecto a los efectos fijos, dado que suman cero, únicamente será necesario estimar 2 efectos de cóctel, 1 de edad y 2 de interacciones:

$$\hat{\alpha}_{acd3s} = \bar{y}_{acd3s..} - \bar{y}_{...} = 60.2 - 51.2 = 9.0$$

$$\hat{\alpha}_{control} = \bar{y}_{control..} - \bar{y}_{...} = 42.3 - 51.2 = -8.9$$

$$\hat{\alpha}_{tcc} = -0.1$$

$$\hat{\beta}_{adult} = \bar{y}_{.adult.} - \bar{y}_{...} = 49.07 - 51.2 = -2.1333$$

$$\hat{\beta}_{infant} = 2.1333$$

$$\widehat{(\alpha\beta)}_{acd3s,adult} = \bar{y}_{acd3s,adult.} - \bar{y}_{acd3s..} - \bar{y}_{.adult.} + \bar{y}_{...} = 58.2 - 60.2 - 49.07 + 51.2 = 0.1333$$

$$\widehat{(\alpha\beta)}_{control,adult} = \bar{y}_{control,adult.} - \bar{y}_{control..} - \bar{y}_{.adult.} + \bar{y}_{...} = 38.8 - 42.3 - 49.07 + 51.2 = -1.3667$$

$$\widehat{(\alpha\beta)}_{tcc,adult} = - (0.1333 - 1.3667) = 1.2333$$

$$\widehat{(\alpha\beta)}_{acd3s,infant} = -0.1333, \widehat{(\alpha\beta)}_{control,infant} = 1.3667, \widehat{(\alpha\beta)}_{tcc,infant} = -1.2333$$

(También sería justificable dar valor nulo a todas las estimaciones de los parámetros de interacción, dado el resultado de la prueba de significación. Pero recordad que la no significación de la interacción no es una prueba feaciente de que no exista).

Finalmente la varianza residual la estimaremos a partir del cuadrado medio residual: $\hat{\sigma}^2 = MS_E = 7,33$.

Estaría justificada la afirmación de que la condición que más favorece la expansión de las células CD34+ es la aplicación de acd3s en infants, a pesar de que la posible falta de interacción también justificaría hablar simplemente de acd3s, si el estudio se quiere centrar en los tratamientos. En cualquier caso, la diferencia entre niveles de tratamiento parece comprobada, no se puede atribuir al azar.

| subject | cocktail | población | count |
|---------|----------|-----------|-------|
| 1 | control | 1 | 45 |
| 2 | control | 1 | 47 |
| 3 | control | 1 | 39 |
| 4 | control | 1 | 47 |
| 5 | control | 1 | 51 |
| 6 | control | 2 | 39 |

| | | | |
|----|---------|---|----|
| 7 | control | 2 | 40 |
| 8 | control | 2 | 42 |
| 9 | control | 2 | 35 |
| 10 | control | 2 | 38 |
| 11 | tcc | 3 | 52 |
| 12 | tcc | 3 | 52 |
| 13 | tcc | 3 | 53 |
| 14 | tcc | 3 | 54 |
| 15 | tcc | 3 | 49 |
| 16 | tcc | 4 | 49 |
| 17 | tcc | 4 | 48 |
| 18 | tcc | 4 | 51 |
| 19 | tcc | 4 | 52 |
| 20 | tcc | 4 | 51 |
| 21 | acd3s | 5 | 66 |
| 22 | acd3s | 5 | 64 |
| 23 | acd3s | 5 | 62 |
| 24 | acd3s | 5 | 61 |
| 25 | acd3s | 5 | 58 |
| 26 | acd3s | 6 | 57 |
| 27 | acd3s | 6 | 56 |
| 28 | acd3s | 6 | 59 |
| 29 | acd3s | 6 | 59 |
| 30 | acd3s | 6 | 60 |

Supón ahora que los datos y el diseño anterior en realidad todos los sujetos tienen una edad similar, pero podían proceder de diferentes poblaciones. Hay muchas poblaciones posibles, pero todas son pequeñas. Se cogieron 6 poblaciones al azar. Dado que se disponían de pocos individuos de cada población, para estudiar el cóctel “control” se utilizó una muestra de 5 individuos de la población 1 y 5 individuos de la población 2, para estudiar “tcc” se utilizó una muestra de 5 individuos de la población 3 y 5 individuos de la población 4, etc. Los resultados podrían haberse expresado como los de la tabla adjunta. Las tres preguntas siguientes se refieren a esta nueva situación.

5) De forma similar al apartado “1”, describe ahora el tipo de diseño experimental, el modelo lineal asociado y su parametrización. **(1 punto)**

Ahora tenemos un diseño jerárquico. En el nivel más alto tenemos el factor “cocktail”, fijo, con las mismas características que en los apartados anteriores. Jerárquicamente por debajo de “cocktail” tenemos “población”, aleatorio, con $b = 2$ niveles para cada nivel de “cocktail”, y 6 niveles diferentes en total. Para cada combinación de los niveles de ambos factores tenemos $n=5$ réplicas.

El modelo lineal asociado sería:

valor observat de "count" pel subjecte k del "cocktail" i per individus de la població j

↓ mitjana general o constant

↓ ↓ residu pel subjecte k del "cocktail" i , "població" j

↓ ↓ ↓

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{k(j)}$$

↑ ↑

↑ efecte aleatori del nivell j del factor "població" dins el nivell i de "cocktail"
efecte fix del nivell i del factor "cocktail"

$$\sum_{i=1}^a \alpha_i = 0,$$

$$\left. \begin{array}{l} e_{k(j)} \sim N(0, \sigma) \text{ iid} \\ \beta_{j(i)} \sim N(0, \sigma_B) \text{ iid} \end{array} \right\} \text{mútuament independents.}$$

Con parametrización $\mu, \alpha_{\text{ac3s}}, \alpha_{\text{control}}, \alpha_{\text{tec}}, \sigma_B^2, \sigma^2$.

6) **De forma similar al apartado "2"**, indica el listado (tabla ANOVA) que sería adecuado ahora (substituyendo "age" por "población"), completa la tabla ANOVA correspondiente y determina cuáles serían los efectos significativos. **(1 punto)**

Utilizaremos el LISTADO 2, teniendo en cuenta que "población" es un factor aleatorio, por lo tanto obtendremos el coeficiente F de "cocktail" dividiendo por el cuadrado medio de "población (cocktail)":

| | Df | Sum Sq | Mean Sq | F value | F taules | Pr(>F) |
|--------------|----|--------|---------|--------------------|-----------------------|--------|
| cocktail | 2 | 1602.2 | 801.10 | 801.10/56.87=14.09 | $F_{0.05}(2,3)=9.55$ | < 0.05 |
| cocktail:pob | 3 | 170.6 | 56.87 | 56.87/7.33=7.75 | $F_{0.05}(3,24)=3.01$ | < 0.05 |
| Residuals | 24 | 176.0 | 7.33 | | | |

$$H_0 : \alpha_1 = \dots = \alpha_a = 0 \text{ vs } H_1 : \alpha_i \neq 0, \text{ per algun } i = 1, \dots, a$$

El estadístico $F = 14,09$ es más grande que el valor crítico $9,55$, por el nivel crítico escogido $(0,05)$ y 2 y 3 g.d.l (ahora hemos dividido por $MS_{\text{población(cocktail)}}$). Por tanto rechazamos H_0 , descartamos que el efecto de

los diversos cócteles sea nulo e igual para todos.

$$H_0 : \sigma_B^2 = 0 \text{ vs } H_1 : \sigma_B^2 > 0$$

El estadístico $F = 7,75$ es mayor que el valor crítico $3,01$, correspondiente a 3 y 24 g.d.l y un nivel de $0,05$. Por lo tanto rechazamos la hipótesis nula de que la varianza atribuible a la población es cero. En otras palabras, hay una parte de la varianza en la variable de respuesta que es atribuible a las diferencias en la población de origen.

- 7) Calcula la probabilidad de que un individuo que ha recibido el cóctel “tcc” muestre un valor de la variable “count” superior a 50. **(1 punto)**

De acuerdo con el resultado del apartado anterior, será necesario estimar la varianza debida a “población”:

$$\hat{\sigma}_B^2 = \frac{MS_{B(A)} - MS_E}{n} = \frac{56,87 - 7,33}{5} = 9,91 \quad \hat{\sigma}^2 = MS_E = 7,33$$

Por tanto: $Y_{tcc} \sim N(\bar{y}_{tcc} = 51,1, \sqrt{\hat{\sigma}_B^2 + \hat{\sigma}^2} = \sqrt{9,91 + 7,33} = 4,15)$ i

$$\Pr\{Y_{tcc} > 50\} = \Pr\left\{Z > \frac{50 - 51,1}{4,15}\right\} = \Pr\{Z > -0,26\} = \Pr\{Z \leq 0,26\} = 0,6026$$

LLI STAT 1

```
> cytoki nes. aov = aov(count ~ cocktail, data = cytoki nes)
> anova(cytoki nes. aov)
```

Analysis of Variance Table

```
Response: count
          Df Sum Sq Mean Sq F value Pr(>F)
cocktail  2 1602.2
Residuals 27  346.6
```

LLI STAT 2

```
> cytoki nes. aov = aov(count ~ cocktail / age, data = cytoki nes)
> anova(cytoki nes. aov)
```

Analysis of Variance Table

```
Response: count
          Df Sum Sq Mean Sq F value Pr(>F)
cocktail  2 1602.2
cocktail:age  3  170.6
Residuals  24  176.0
```

LLI STAT 3

```
> cytokines.aov = aov(count ~ cocktail * age, data = cytokines)
> anova(cytokines.aov)
Analysis of Variance Table
```

```
Response: count
      Df Sum Sq Mean Sq F value Pr(>F)
cocktail  2 1602.20      801.10  11.84 0.000114
age       1  136.53      136.53  1.94 0.168101
cocktail:age  2   34.07      17.03  0.24 0.784101
Residuals 24  176.00
```

LLI STAT 4

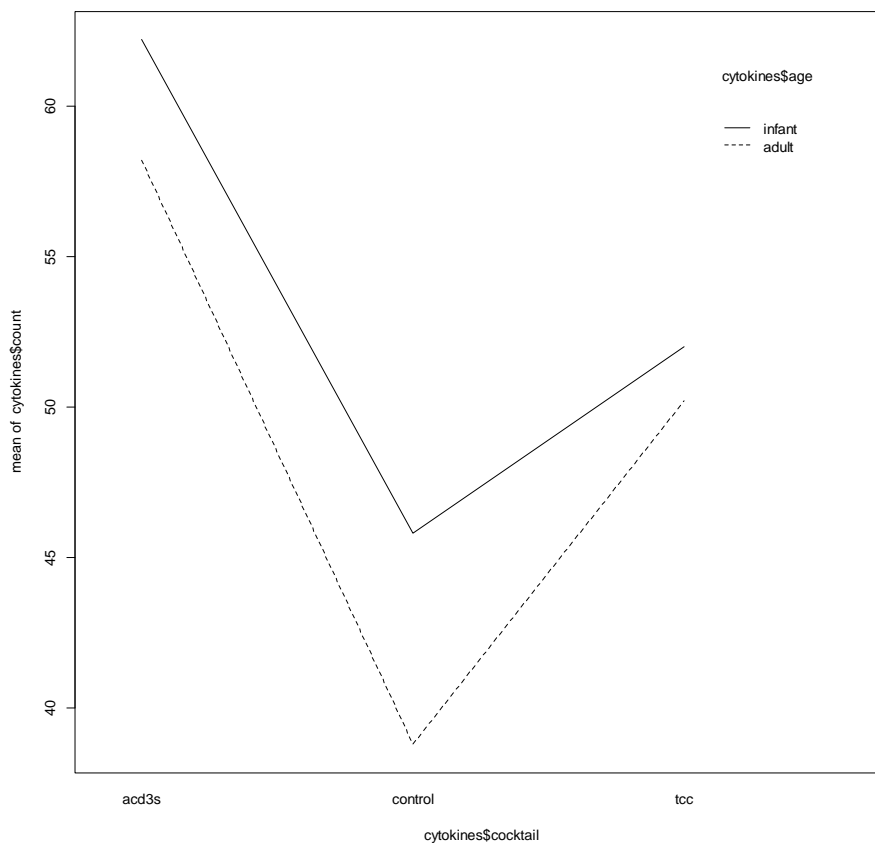
```
> TukeyHSD(cytokines.aov, which = "cocktail")
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = count ~ cocktail * age, data = cytokines)
```

```
$cocktail
      diff      lwr      upr p adj
control-acd3s -17.9 -20.924365 -14.875635 0e+00
tcc-acd3s      -9.1 -12.124365  -6.075635 3e-07
tcc-control     8.8   5.775635  11.824365 5e-07
```

LLI STAT 1 GRÁFIC 5

```
> interaction.plot(cytokines$cocktail, cytokines$age, cytokines$count)
```



LLI STAT 6

```
> model.tables(cytokines.aov, type = "mean")
Tables of means
Grand mean
```

51.2

```
cocktail  
cocktail  
acd3s control tcc  
60.2 42.3 51.1
```

```
age  
age  
adult infant  
49.07 53.33
```

```
cocktail : age  
age  
cocktail adult infant  
acd3s 58.2 62.2  
control 38.8 45.8  
tcc 50.2 52.0
```


3.5. Fiabilidad de la medida: muestra y técnico de laboratorio.



Fiabilidad de la medida (Imagen de <https://img.interempresas.net/fotos/1033132.jpeg>)

La fiabilidad es un tema constante en todos los instrumentos de medida. Su estudio trata de establecer la precisión con la que mide cualquier instrumento de medida en general y los tests en particular. Es un concepto muy asociado al error de medida. Cuanto más fiable es un test, cuanto mayor es la precisión con la que mide, menos error de medida se comete.

Problema. Se quiere medir la fiabilidad de una medida realizada en un laboratorio. Se cree que la muestra concreta de la materia orgánica sobre la que se realiza, y el técnico de laboratorio encargado de hacerla, podrían explicar la variabilidad de esta medida. Por este motivo, se seleccionan cuatro técnicos al azar y se toman dos muestras al azar de materia orgánica. Cada técnico analiza material proveniente de cada muestra, repitiendo la medida dos veces². Los resultados obtenidos quedan finalmente resumidas en la siguiente tabla de datos:

| | Técnico 1 | Técnico 2 | Técnico 3 | Técnico 4 |
|-----------|-----------|-----------|-----------|-----------|
| Muestra 1 | 7.8 | 7.9 | 7.5 | 7.4 |
| | 7.6 | 8.1 | 7.6 | 7.6 |
| Muestra 2 | 8.1 | 8.4 | 7.9 | 7.9 |

| | | | | |
|--|-----|-----|-----|-----|
| | 8.3 | 8.2 | 8.0 | 7.7 |
|--|-----|-----|-----|-----|

En todas las preguntas de este problema puedes asumir normalidad y homogeneidad de las varianzas, y puedes usar un nivel de significación de 0,05. Utilizando los listados de R oportunos, cuando sea necesario, responde a las siguientes cuestiones (cada apartado vale 1 punto).

- a. Describe el diseño, en concreto indica: factores y niveles de cada factor, carácter fijo aleatorio de los factores, descripción detallada del modelo lineal asociado, parametrización del modelo. (No te preocupes por el tema de si es un estudio experimental u observacional).

Se trata de un diseño cruzado balanceado, con $n=2$ réplicas por casilla, para los factores $A=$ ''Muestra'' con $a=2$ niveles y $B=$ ''Tècnic'' con $b=4$ niveles, ambos dos aleatorios. El modelo lineal asociado es:

Mesura per la mostra i feta pel tècnic j en la repetició k de la mesura

↓ Mitjana general

↓ ↓ Efecte aleatori de la mostra i

↓ ↓ ↓

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$$

↑ ↑ ↑

↑ ↑ Residu de l'observació ijk

↑ Efecte aleatori de la interacció entre mostra i i tècnic j

Efecte aleatori del tècnic j

On: $A_i \sim N(0, \sigma_{mostra})$ independents i idènticament distribuïdes (iid),

$B_j \sim N(0, \sigma_{tècnic})$ iid

$(AB)_{ij} \sim N(0, \sigma_{mostra:tècnic})$ iid, $e_{ijk} \sim N(0, \sigma)$ iid

i les $A_i, B_j, (AB)_{ij}$ i e_{ijk} mútuament independents.

La parametrización del modelo viene definida por 5 parámetros:

$$\mu, \sigma_{mostra}^2, \sigma_{tècnic}^2, \sigma_{mostra:tècnic}^2 \text{ i } \sigma^2$$

- b. Estima todos los parámetros del modelo lineal.

$\hat{\mu} = \bar{y}_{...} = 7.875$, la media global del listado "model.tables{...}"

En la estimación de las varianzas de los efectos aleatorios podemos usar cualquiera de las tablas ANOVA, ya que los cuadrados medios son siempre los mismos.

$$\hat{\sigma}_{\text{mostra}}^2 = \frac{MS_{\text{mostra}} - MS_{\text{mostra:tècnic}}}{b \cdot n} = \frac{0.56250 - 0.00971}{4 \cdot 2} = 0.06916625$$

$$\hat{\sigma}_{\text{tècnic}}^2 = \frac{MS_{\text{tècnic}} - MS_{\text{mostra:tècnic}}}{a \cdot n} = \frac{0.19667 - 0.00971}{2 \cdot 2} = 0.046875$$

$$\hat{\sigma}_{\text{mostra:tècnic}}^2 = \frac{MS_{\text{mostra:tècnic}} - MS_E}{n} = \frac{0.00971 - 0.01625}{2} = -0.00327 < 0$$

atès que una variància negativa no té sentit agafem $\hat{\sigma}_{\text{mostra:tècnic}}^2 = 0$ (millor que dir que l'aproximem a 0 per que és quasi 0, si fos quasi 0 però positiu, segurament aquest valor el deixariem com està)

$$\hat{\sigma}^2 = MS_E = 0.01625$$

- c. Determina la significación de todos los efectos del modelo. En este sentido, copia aquí la tabla ANOVA correcta, plantea las hipótesis nula y alternativa para cada efecto, e indica justificadamente la conclusión final.

Dado que tanto un factor como el otro son aleatorios, los coeficientes F correctos para el efecto de cada factor separadamente se han de formar dividiendo por el cuadrado medio de interacción. Por lo tanto, la tabla ANOVA adecuada es la primera:

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|----|--------|---------|---------|----------|
| Muestra | 1 | 0.5625 | 0.56250 | 61.3636 | 0.004332 |
| Tècnic | 3 | 0.5900 | 0.19667 | 21.4545 | 0.015740 |
| Muestra:Tècnic | 3 | 0.0275 | 0.00917 | 0.5641 | 0.653767 |
| Residuals | 8 | 0.1300 | 0.01625 | | |

Para el contraste $H_0 : \sigma_{\text{mostra}}^2 = 0$ vs. $H_1 : \sigma_{\text{mostra}}^2 > 0$ se debe rechazar H_0 ya que el p-valor es inferior al nivel de significación usado, 0.05. Existe evidencia de que para una parte de la variabilidad de la medida es debida a las muestras.

Similarmente, para $H_0 : \sigma_{\text{técnic}}^2 = 0$ vs. $H_1 : \sigma_{\text{técnic}}^2 > 0$ se debe rechazar H_0 . Existe evidencia de que una parte de la variabilidad de la medida se debe a los técnicos.

Sin embargo, para $H_0 : \sigma_{\text{mostra:técnic}}^2 = 0$ vs. $H_1 : \sigma_{\text{mostra:técnic}}^2 > 0$ no podemos rechazar H_0 ya que el p-valor es grande, $0.653767 > 0.05$, no hemos encontrado evidencias de interacción.

Evidentemente se obtendrían los mismos resultados a partir de los valores críticos de la tabla F, pero aquí no es necesario que dispongamos de los p-valores.

d. Des de un punto de vista estadístico, ¿tendría sentido invertir en formación para los técnicos, con el fin de conseguir mayor uniformidad en sus muestras? Justifica la respuesta.

Sí que tendría sentido ya que hay evidencias de que una parte de la variabilidad de esta medida sea debida a la heterogeneidad de los técnicos, $\sigma_{\text{técnic}}^2 > 0$, una parte relativamente importante, una estimación del 35.43% del total:

$$\begin{aligned} 100 \frac{\hat{\sigma}_{\text{técnic}}^2}{\hat{\sigma}_{\text{total}}^2} &= 100 \frac{\hat{\sigma}_{\text{técnic}}^2}{\hat{\sigma}_{\text{mostra}}^2 + \hat{\sigma}_{\text{técnic}}^2 + \hat{\sigma}_{\text{mostra:técnic}}^2 + \hat{\sigma}^2} \\ &= 100 \frac{0.046875}{0.06916625 + 0.046875 + 0 + 0.01625} \\ &= 100 \frac{0.046875}{0.13229125} = 35.43\% \end{aligned}$$

El razonamiento se ha de basar en la varianza de "técnico", en el hecho de que posiblemente con la formación la disminuiríamos. Las comparaciones múltiples con la prueba de Tukey presentados en los listados no tienen sentido para argumentar en esta situación ya que el factor es aleatorio, los cuatro técnicos considerados son solo una muestra de una población mucho más amplia, no tiene relevancia que se haya encontrado una diferencia de medias "significativa" entre tal y cual técnico, etc.

e. Calcula la probabilidad de que una medida sea superior a 9.

Podemos comprobar en los datos que ninguno de los valores observados cumplen la propiedad de ser superior a 9. Por lo tanto, una estimación

muy razonable de la probabilidad anterior sería decir que vale 0. Aunque en realidad esto es convertir lo improbable en imposible, nos puede servir para tener una indicación de si lo hemos calculado bien. Bajo las condiciones del modelo normal que hemos asumido aquí, esta probabilidad no será cero, pero previsiblemente estará muy cercana a este valor.

Dado que todos los efectos estudiados son aleatorios, la distribución de Y la podemos asimilar a una normal con media igual a la media global $\hat{\mu} = \bar{y}_{...} = 7.875$ y con varianza igual a la suma de las varianzas de todos los efectos aleatorios calculada en el apartado anterior: 0.13229125. La desviación típica estimada será $\sqrt{0.13229125} = 0.363718$. Por tanto:

$$\Pr\{Y > 9\} = \Pr\left\{Z > \frac{9 - 7.875}{0.363718}\right\} = \Pr\{Z > 3.09\} = 1 - \Pr\{Z \leq 3.09\}, \text{ para } Z$$

$\sim N(0,1)$. En las tablas de la distribución $N(0,1)$ no suele aparecer el valor de la probabilidad correspondiente a 3.09, $\Pr\{Z \leq 3.09\}$, pero se puede deducir que de ser muy cercano a 1 (correspondería a casi toda la superficie bajo la campana de Gauss, a la izquierda de un valor “muy a la derecha” como por ejemplo es 3.09) y por lo tanto la probabilidad pedida ha de estar muy cercana a 0. Si tuviéramos una calculadora científica que permitiera calcular la función de distribución $N(0,1)$ (o algún otro medio adecuado, como por ejemplo Excel o R), veríamos que sería un valor un poco superior a 0.001. Pero una respuesta correcta es decir que es casi 0.

Supón ahora que “Muestra” no significa lo que hemos dicho al principio, sino que se refiere a dos únicas posibilidades de procesamiento del material orgánico que se analizará (que siempre es muy similar). Las dos preguntas siguientes se refieren a esta nueva situación.

- f. Estima todos los parámetros del modelo (algunos parámetros continuarán teniendo el mismo significado y estimación, solo has de indicar las cosas que han cambiado respecto a la pregunta b)**

En este caso el factor “Muestra” (que debería llamarse en este contexto “Sistema de procesamiento”), será un factor fijo con $a=2$ niveles. En la especificación del modelo lineal, debemos de cambiar la variable

aleatoria A, caracterizada por su varianza σ_{mostra}^2 , por las constantes α_1 y α_2 . La especificación de los otros efectos del modelo queda igual. A partir de las medias proporcionados por la orden "model.tablea..." de los listados, las estimaciones de estos parámetros serán:

$$\hat{\alpha}_1 = \bar{Y}_{1..} - \bar{Y}_{...} = 7.687 - 7.875 = -0.188$$

$$\hat{\alpha}_2 = -\hat{\alpha}_1 = 0.188$$

g. Determina la significación de todos los efectos del modelo. Copia aquí la tabla ANOVA correcta, plantea las hipótesis nula y alternativa para cada efecto e indica justificadamente la conclusión final (únicamente indica lo para los efectos para los cuales la interpretación o los cálculos hayan cambiado respecto a la pregunta c).

Si no consideramos el modelo restringido, la tabla ANOVA adecuada sigue siendo la misma que en la pregunta c.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---------------|----|--------|---------|---------|----------|
| Mostra | 1 | 0.5625 | 0.56250 | 61.3636 | 0.004332 |
| Tecnic | 3 | 0.5900 | 0.19667 | 21.4545 | 0.015740 |
| Mostra:Tecnic | 3 | 0.0275 | 0.00917 | 0.5641 | 0.653767 |
| Residuals | 8 | 0.1300 | 0.01625 | | |

Las conclusiones serían las mismas, excepto que el contraste de hipótesis para "Muestra" sería ahora $H_0 : \alpha_1 = \alpha_2 = 0$ vs. $H_1 : \alpha_1 \neq \alpha_2$ (y por tanto $\neq 0$) y por lo tanto rechazaríamos la hipótesis nula.

(Si alguien hubiera considerado el modelo restringido, entonces debería haber escogido la segunda tabla ANOVA, y el planteamiento del contraste y la conclusión serían iguales, a pesar de que el p-valor fuera diferente, 0.002419.)

LISTADOS

Analysi s of Vari ance Table

Response: Mesura

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|----------|
| Mostra | 1 | 0.5625 | 0.56250 | 61.3636 | 0.004332 |
| Tecni c | 3 | 0.5900 | 0.19667 | 21.4545 | 0.015740 |
| Mostra: Tecni c | 3 | 0.0275 | 0.00917 | 0.5641 | 0.653767 |
| Resi dual s | 8 | 0.1300 | 0.01625 | | |

Anal ysi s of Vari ance Tabl e

Response: Mesura

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|----------|
| Mostra | 1 | 0.5625 | 0.56250 | 61.3636 | 0.004332 |
| Tecni c | 3 | 0.5900 | 0.19667 | 12.1026 | 0.002419 |
| Mostra: Tecni c | 3 | 0.0275 | 0.00917 | 0.5641 | 0.653767 |
| Resi dual s | 8 | 0.1300 | 0.01625 | | |

Anal ysi s of Vari ance Tabl e

Response: Mesura

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|-----------|
| Mostra | 1 | 0.5625 | 0.56250 | 34.6154 | 0.0003685 |
| Tecni c | 3 | 0.5900 | 0.19667 | 12.1026 | 0.0024188 |
| Mostra: Tecni c | 3 | 0.0275 | 0.00917 | 0.5641 | 0.6537675 |
| Resi dual s | 8 | 0.1300 | 0.01625 | | |

```
> model.tables(resul tats.aov, type = "mean")
```

Tables of means

Grand mean

7.875

```

Mostra
Mostra
Mostra1 Mostra2
  7.687  8.062

```

```

Tecnico
Tecnico
Tecnico1 Tecnico2 Tecnico3 Tecnico4
  7.95  8.15  7.75  7.65

```

```

Mostra: Tecnico
      Tecnico
Mostra  Tecnico1 Tecnico2 Tecnico3 Tecnico4
Mostra1 7.70  8.00  7.55  7.50
Mostra2 8.20  8.30  7.95  7.80

```

```

>
> TukeyHSD(results.aov, which = "Tecnico")

```

```

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

```

```

Fit: aov(formula = Mesura ~ Mostra * Tecnico, data = results)

```

```

$Tecnico
      diff      lwr      upr    p adj
Tecnico2-Tecnico1 0.2 -0.08865611 0.48865611 0.1977127
Tecnico3-Tecnico1 -0.2 -0.48865611 0.08865611 0.1977127
Tecnico4-Tecnico1 -0.3 -0.58865611 -0.01134389 0.0419023
Tecnico3-Tecnico2 -0.4 -0.68865611 -0.11134389 0.0093641
Tecnico4-Tecnico2 -0.5 -0.78865611 -0.21134389 0.0024190
Tecnico4-Tecnico3 -0.1 -0.38865611 0.18865611 0.6942388

```



```

>
> resul tats.Mostra1.aov = aov(Mesura ~ Tecni c,
data = subset(resul tats, Mostra == "Mostra1"))
> TukeyHSD(resul tats.Mostra1.aov, whi ch = "Tecni c")
  Tukey mul ti ple comparisons of means
    95% fami ly-wi se confi dence l evel

```

```

Fi t: aov(formul a = Mesura ~ Tecni c, data = subset(resul tats, Mostra == "Mostra1"))

```

```

$Tecni c
          di ff          l wr          upr          p adj
Tecni c2-Tecni c1  0.30 -0.2189342  0.81893423  0.2288345
Tecni c3-Tecni c1 -0.15 -0.6689342  0.36893423  0.6698642
Tecni c4-Tecni c1 -0.20 -0.7189342  0.31893423  0.4825055
Tecni c3-Tecni c2 -0.45 -0.9689342  0.06893423  0.0779124
Tecni c4-Tecni c2 -0.50 -1.0189342  0.01893423  0.0562844
Tecni c4-Tecni c3 -0.05 -0.5689342  0.46893423  0.9769156

```

```

> resul tats.Mostra2.aov = aov(Mesura ~ Tecni c,
data = subset(resul tats, Mostra == "Mostra2"))
> TukeyHSD(resul tats.Mostra2.aov, whi ch = "Tecni c")
  Tukey mul ti ple comparisons of means
    95% fami ly-wi se confi dence l evel

```

```

Fi t: aov(formul a = Mesura ~ Tecni c, data = subset(resul tats, Mostra == "Mostra2"))

```

```

$Tecni c
          di ff          l wr          upr          p adj
Tecni c2-Tecni c1  0.10 -0.4189342  0.61893423  0.8584201
Tecni c3-Tecni c1 -0.25 -0.7689342  0.26893423  0.3341607
Tecni c4-Tecni c1 -0.40 -0.9189342  0.11893423  0.1098315
Tecni c3-Tecni c2 -0.35 -0.8689342  0.16893423  0.1574844

```

Tecni c4-Tecni c2 -0.50 -1.0189342 0.01893423 0.0562844

Tecni c4-Tecni c3 -0.15 -0.6689342 0.36893423 0.6698642

3.6. Influencia del hexobarbital en enzimas metabolizadores y efecto del sexo de los animales

| sex | S | 10 | 50 | 250 |
|--------|----|----|----|-----|
| male | 21 | 33 | 32 | 36 |
| male | 29 | 36 | 33 | 38 |
| male | 32 | 40 | 35 | 40 |
| male | 37 | 42 | 37 | 44 |
| male | 38 | 47 | 40 | 45 |
| male | 41 | 49 | 41 | 48 |
| male | 42 | 53 | 44 | 48 |
| male | 45 | 55 | 49 | 49 |
| male | 51 | 58 | 53 | 54 |
| male | 56 | 67 | 44 | 60 |
| female | 23 | 19 | 17 | 20 |
| female | 40 | 23 | 22 | 24 |
| female | 43 | 24 | 26 | 28 |
| female | 44 | 27 | 26 | 29 |
| female | 47 | 32 | 32 | 31 |
| female | 50 | 34 | 32 | 33 |
| female | 62 | 38 | 35 | 35 |
| female | 66 | 40 | 35 | 37 |
| female | 67 | 40 | 36 | 45 |
| female | 67 | 43 | 51 | 52 |

Problema 1. El hexobarbital o hexobarbitona es un barbitúrico que tiene efectos hiptónicos y sedantes. En un estudio se pretendía estudiar el efecto de una droga experimental que, supuestamente, inducía la actividad de los enzimas metabolizadores del hexobarbital. El estudio se realizó sobre una muestra de 80 ratas, de las cuales 40 eran machos y 40 hembras. Los 80 animales recibieron la misma dosis de hexobarbital. Posteriormente, por separado tanto a machos como a hembras, fueron asignados al azar y de forma balanceada a 4 posibles tratamientos: un control negativo que consistía en una solución salina que no contenía la droga experimental (S) y la misma preparación pero con tres posibles concentraciones de droga (10, 50 y 250 mg/kg). El efecto de los diversos tratamientos, se media por medio de minutos de sueño inducidos por el hexobarbital; en principio se esperaba que la metabolización del hexobarbital se debería de traducir en una disminución de estos tiempos. En la tabla adjunta se indican los valores de la duración del sueño para cada

uno de estos 80 animales.

En todo momento asumir normalidad y homoscedasticidad y un nivel de significación 0.05 o de confianza 0.95.



Imagen de https://media.cnnchile.com/2018/08/imagen_principal-13040-740x400.jpg

Utilizando cuando sean necesarios los listados de R adjuntos, responde a las siguientes cuestiones:

Describe el diseño. En concreto:

- a. Indica los factores que intervienen y la variable de respuesta, el carácter fijo o aleatorio de los factores y sus niveles:

Se trata de dos factores fijos que podríamos designar como $A = \text{sexo}$ y $B = \text{tratamiento}$. El primero tiene $a = 2$ niveles: "macho" y "hembra", y el segundo $b = 4$ niveles: "S", "10", "50" i "250" –o los correspondientes nombres en inglés, tal y como aparecen en los listados. La variable de respuesta es $Y = \text{minutos de sueño}$, indicada en los listados como "sleep".

- b. Indica la relación entre los factores (diseño cruzado o jerárquico):

Se trata de un caso cruzado con réplicas balanceado. Cada nivel de uno de los factores se combina con cada nivel del otro. Por ejemplo podríamos decir que los tratamientos que reciben los machos son exactamente los mismos que reciben las hembras. Para cada situación experimental posible (o "celda" o "casilla"...) (machos que han sido tratados con S, machos que han sido tratados con 10, etc.) hay $n=10$ réplicas.

- c. ¿Para qué o cuáles factores se ha podido aleatorizar? Por tanto, ¿Qué podría decir de este estudio en referencia a su posible carácter experimental u observacional?

Tal y como se indica en el enunciado, únicamente se han podido aleatorizar por tratamiento, ya que tanto los machos como las hembras han sido asignados al azar a recibir uno de los tratamientos. El sexo es evidentemente una característica intrínseca de cada sujeto. Es decir, podríamos decir que es un estudio experimental por el factor tratamiento, en el cual posiblemente el sexo juega más bien un papel de factor bloque, considerado en el estudio para hacer comparaciones entre tratamientos en condiciones más homogéneas, para estudiar posibles interacciones, etc.

- d. Escribe el modelo lineal asociado y la parametrización del modelo.

El modelo lineal es,

Tiempo de sueño por nivel y de sexo, nivel j de tratamiento, réplica.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

The diagram illustrates the components of the linear model equation above. It consists of five vertical arrows pointing downwards from the terms in the equation to their corresponding labels:

- The first arrow (from μ) is labeled "Media general o constante".
- The second arrow (from α_i) is labeled "Efecto del nivel y del sexo".
- The third arrow (from β_j) is labeled "Efecto del nivel j de tratamiento".
- The fourth arrow (from $(\alpha\beta)_{ij}$) is labeled "Interacción entre los niveles i, j ambos factores".
- The fifth arrow (from e_{ijk}) is labeled "Residuo de la réplica k de la condición i, j ".

Donde:

$i = 1, \dots, a = 2$ ($1 \equiv \text{mascle}$, $2 \equiv \text{femella}$), $j = 1, \dots, b = 4$ ($1 \equiv S$, $2 \equiv 10$, etc.),
 $k = 1, \dots, n = 10$

$$\sum_{i=1}^a \alpha_i = 0, \quad \sum_{j=1}^b \beta_j = 0,$$

$$\sum_{j=1}^b (\alpha\beta)_{ij} = 0 \text{ per } i = 1, \dots, a \text{ i } \sum_{i=1}^a (\alpha\beta)_{ij} = 0 \text{ per } j = 1, \dots, b$$

totes les e_{ijk} independents i idènticament distribuïdes (iid) $e_{ijk} \sim N(0, \sigma)$

Parametrizaci3n:

$\mu, \alpha_{\text{mascle}}, \beta_S, \beta_{10}, \beta_{50}, (\alpha\beta)_{\text{mascle}, S}, (\alpha\beta)_{\text{mascle}, 10}, (\alpha\beta)_{\text{mascle}, 50}, \sigma^2$ (9 paràmetres desconeguts) (o bien 16 en total si no tenemos en cuenta las restricciones anteriores).

- 2) Escoge la tabla ANOVA adecuada y complétala con el fin de analizar la significaci3n de cada uno de los efectos del modelo (puedes hacerlo sobre el listado escogido). Plantea en cada caso la hip3tesis nula y alternativa, e indica, razonadamente, la conclusi3n final de cada test,

Seleccionamos la primera tabla ANOVA de los listados, la 3nica compatible con una situaci3n cruzada con r3plicas. Debemos fijarnos en los efectos considerados y los grados de libertad (los otros datos corresponden a un caso jeràrquico). En los mismos listados los hemos completado.

Dado que se trata de dos factores fijos, todas las F "observadas" o "experimentales" ("F value" en los listados), estàn obtenidas dividiendo por el cuadrado de la media de los residuos 97,16. Por ejemplo por el efecto de sexo, $F = 891,11 / 97,16 = 9,17$, etc.

Para $H_0: \alpha_{\text{mascle}} = \alpha_{\text{femella}} = 0$ vs $H_1: \alpha_{\text{mascle}} \neq \alpha_{\text{femella}}$ el valor del estadístico es $F = 9,17$. El valor crítico de la distribuci3n F para un nivel de significaci3n 0,05 con 1 y 72 grados de libertad (g.d.l), indiquémoslo como $F_{0,05}(1,72)$, no està en las tablas colgadas en el campus virtual que de 60 g.d.l en el denominador saltan a 120 g.d.l. Evidentemente el valor que estamos buscando està entre 4 y 3,92, pero no es necesario conocerlo exactamente; està claro que $9,17 > F_{0,05}(1,72)$ y por lo tanto rechazamos H_0 . En otras palabras,

el p-valor es claramente inferior a 0,05. A pesar de esto, con las condiciones propias de una prueba de hipótesis se han encontrado evidencias que existe un efecto de sexo.

Similarmente,

para $H_0 : \beta_S = \beta_{10} = \beta_{50} = \beta_{250} = 0$ vs $H_1 : \beta_j \neq \beta_{j'}$, per algun $1 \leq j \neq j' \leq 4$, $F = 2,84 > F_{0,05}(3,72)$ (un valor entre 2,76 y 2,68) y por tanto rechazamos H_0 , también bajo una probabilidad de error de tipo I de 0,05 podemos afirmar la existencia de un efecto de tratamiento.

Finalmente, para

$H_0 : (\alpha\beta)_{mascle,S} = \dots = (\alpha\beta)_{femella,250} = 0$ vs $H_1 : (\alpha\beta)_{ij} \neq (\alpha\beta)_{i'j'}$
per algun $(i, j) \neq (i', j')$

$F = 8,08 > F_{0,05}(3,72)$ y rechazamos H_0 , podemos afirmar que hay interacción entre sexo y tratamiento.

3) Análisis post-hoc:

- a. Indica razonadamente si tendría sentido hacer comparaciones múltiples para establecer posibles diferencias entre tratamientos respecto a la media de las horas de sueño.

Sí que tendría sentido, ya que el tratamiento es un factor fijo y la ANOVA anterior ha resultado significativa para este factor. Si por error en los apartados anteriores, alguno hubiera considerado que el tratamiento es un factor aleatorio, entonces debería responder que no tiene sentido y no responder las preguntas b y c.

Tiene sentido hacer comparaciones múltiples, pero no de cualquier manera (lea los apartados siguientes).

- b. En caso afirmativo, indica razonadamente si sería preferible hacerlo para los 40 machos y las 40 hembras por separado, o con los 80 datos juntos (o deja en blanco este apartado, según lo que hayas respondido en el apartado anterior).

Dado que se ha detectado evidencias de la interacción, el efecto del tratamiento puede ser diferente según el sexo —aquí el punto importante es que hay interacción, en cambio no tiene ninguna

relevancia que por separado, sexo también haya resultado significativo. Por tanto, nos debemos basar en los listados que muestran comparaciones múltiples separadamente para los 40 machos, **resaltados en amarillo**. (También sería correcto el enfoque de quien haya planteado hacer comparaciones entre las 8 celdas sexoxtratamiento, pero los listados proporcionados no permiten hacer este estudio de forma tan directa).

- c. Si según las respuestas anteriores tiene sentido, indica que tratamientos muestran una diferencia significativa, si es que hay alguna. ¿Se puede decir de forma general que la droga experimental influye en la duración media del sueño?

Como podemos ver en los **listados** de la prueba de Tukey, para los machos no se ha podido encontrar evidencia de que el tratamiento influya en la duración media (poblacional) del sueño, ya que todos los p-valores son superiores al nivel de significación previamente escogido. En otras palabras, para los datos de machos separados y utilizando la prueba de Tukey no hemos podido rechazar ninguna de las hipótesis nulas $H_0 : \mu_j = \mu_{j'},$ para $1 \leq j \neq j' \leq 4$. En cambio, para las hembras sí que se han encontrado evidencias de una diferencia de medias entre el nivel S (ausencia de droga experimental), y cada uno de los otros niveles (los diferentes grados, no nulos, de concentración de la droga), pero no se ha encontrado ninguna evidencia de diferencias entre las diferentes concentraciones. Recordad que todos estos resultados negativos no demuestran que, para los machos, las medias sean todas iguales, o que para las hembras no hayan diferencias entre concentraciones. No rechazar una hipótesis nula no la demuestra.

- 4) Calcula la probabilidad de que una hembra que ha sido tratada con S duerma más de una hora.

Un simple examen de los datos de las 40 hembras que han participado en el estudio, tan sólo 4 han dormido más de una hora, es decir, más de 60 minutos. Por tanto, una estimación muy razonable de la probabilidad que se pide sería $4 / 40 = 0,1$.

Si realmente nos podemos creer el modelo establecido en el enunciado y detallado en las respuestas anteriores, el cálculo de esta probabilidad se puede afinar de la siguiente manera:

De acuerdo con el modelo, hemos establecido la distribución de la duración del sueño de las hembras tratadas con S. Digamos que Y_{jS} de esta variable aleatoria. Fijaros que, según el modelo, los posibles valores de Y_{jS} se obtendrían sumando una serie de constantes, $\mu + \alpha_{femella} + \beta_S + (\alpha\beta)_{femella,S}$ más un único efecto aleatorio, el residuo, que tiene distribución normal con media 0 y varianza σ^2 . Por lo tanto está claro que Y_{jS} tendría distribución normal $N(\mu + \alpha_{hembra} + \beta_S + (\alpha\beta)_{hembra,S}, \sigma^2)$. Los parámetros de esta distribución son desconocidos, y se han de estimar a partir de la muestra. No hace falta estimarlos directamente, los parámetros de los efectos fijos (los α , β , etc.), ya que sabemos que $\mu + \alpha_{hembra} + \beta_S + (\alpha\beta)_{hembra,S}$ corresponde a la esperanza de la duración del sueño para las hembras tratadas con S, por tanto una estimación directa es la media muestral 50,9, valor indicado en verde en el listado "Tables of means". Estimaremos σ^2 a partir del cuadrado medio residual de la tabla ANOVA. Por lo tanto la varianza estimada de Y_{jS} será 97,16 y la desviación típica 9,86.

La probabilidad que se pide es $\Pr\{Y_{jS} > 60\}$, por tanto:

$$\begin{aligned} \Pr\{Y_{jS} > 60\} &= \Pr\left\{Z > \frac{60 - 50,9}{9,86}\right\} = \Pr\{Z > 0,92\} = 1 - \Pr\{Z \leq 0,92\} \\ &= 1 - 0,821213 = 0,18 \end{aligned}$$

para $Z \sim N(0,1)$.

LISTADOS

Analysis of Variance Table

Response: sleep

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|----|--------|---------|---------|--------|
| sex | 1 | 891.1 | 891.11 | 9.17 | < 0.05 |
| treatment | 3 | 827.2 | 275.75 | 2.84 | < 0.05 |
| sex: treatment | 3 | 2353.3 | 784.45 | 8.08 | < 0.05 |
| Residuals | 72 | 6995.7 | 97.16 | | |

Analysis of Variance Table

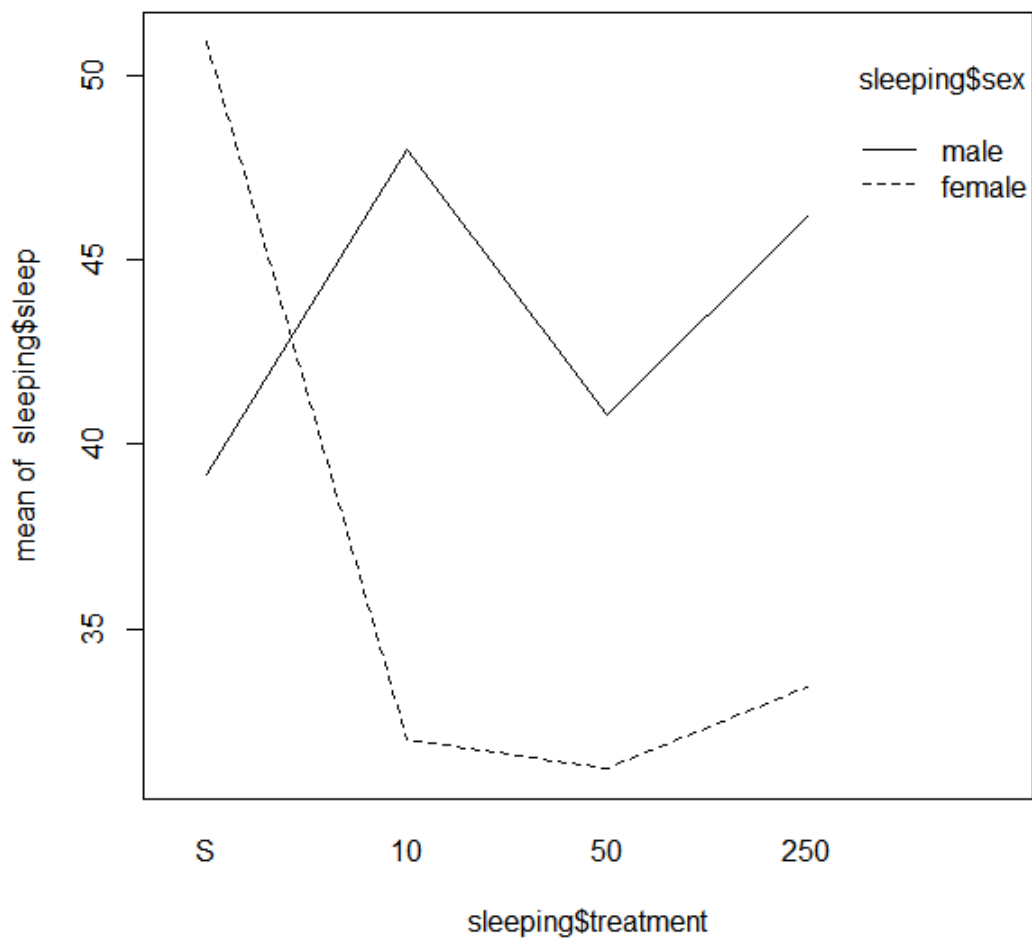
Response: sleep

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|----|--------|---------|---------|--------|
| sex | 1 | 891.1 | 891.11 | | |
| sex: treatment | 6 | 3180.6 | 530.10 | | |
| Residuals | 72 | 6995.7 | 97.16 | | |

Analysis of Variance Table

Response: sleep

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|----|--------|---------|---------|--------|
| treatment | 3 | 827.2 | 275.75 | | |
| treatment: sex | 4 | 3244.4 | 811.11 | | |
| Residuals | 72 | 6995.7 | 97.16 | | |



Tables of means

Grand mean

40.2125

sex

| | |
|--------|-------|
| female | male |
| 36.88 | 43.55 |

```

treatment
  S    10    50    250
45.05 40.00 36.00 39.80

sex: treatment
      treatment
sex    S    10    50    250
female 50.9  32.0 31.2 33.4
male   39.2 48.0 40.8 46.2

```

Tukey multiple comparisons of means
 95% family-wise confidence level
 Fit: aov(formula = sleep ~ sex * treatment, data = sleeping)

```

$treatment
      diff      lwr      upr      p adj
10-S    -5.05 -13.248138  3.148138 0.3739551
50-S    -9.05 -17.248138 -0.851862 0.0247151
250-S   -5.25 -13.448138  2.948138 0.3394249
50-10   -4.00 -12.198138  4.198138 0.5764907
250-10  -0.20  -8.398138  7.998138 0.9999043
250-50   3.80  -4.398138 11.998138 0.6168739

```

```

> sleepingMales.aov <- aov(sleep ~ treatment, data = sleeping, subset =
sex == "male")
> TukeyHSD(sleepingMales.aov, which = "treatment")
Tukey multiple comparisons of means
 95% family-wise confidence level
Fit: aov(formula = sleep ~ treatment, data = sleeping, subset = sex ==
"male")

```

```

$treatment
      diff      lwr      upr      p adj
10-S     8.8  -1.955688 19.555688 0.1416246
50-S     1.6  -9.155688 12.355688 0.9779002
250-S    7.0  -3.755688 17.755688 0.3123358
50-10   -7.2 -17.955688  3.555688 0.2886439
250-10  -1.8 -12.555688  8.955688 0.9690632
250-50   5.4  -5.355688 16.155688 0.5367527

```

```

> sleepingFemales.aov <- aov(sleep ~ treatment, data = sleeping, subset =
sex == "female")
> TukeyHSD(sleepingFemales.aov, which = "treatment")
Tukey multiple comparisons of means
 95% family-wise confidence level
Fit: aov(formula = sleep ~ treatment, data = sleeping, subset = sex ==
"female")

```

```

$treatment
      diff      lwr      upr      p adj
10-S   -18.9 -31.79269 -6.007313 0.0019079
50-S   -19.7 -32.59269 -6.807313 0.0011825
250-S  -17.5 -30.39269 -4.607313 0.0043232
50-10   -0.8 -13.69269 12.092687 0.9983083
250-10   1.4 -11.49269 14.292687 0.9911418
250-50   2.2 -10.69269 15.092687 0.9673095

```

3.7. Oxígeno disuelto en el agua (ODA) y la turbulencia en el agua de los ríos

Problema 2. El oxígeno disuelto en agua (ODA) proviene del oxígeno del aire que se ha disuelto en el agua. El nivel de ODA está muy influido por las turbulencias del río (que aumentan el ODA) o ríos sin velocidad (en los que baja el ODA). En parte, también es el resultado de la fotosíntesis de las plantas acuáticas, de manera que ríos con muchas plantas en días de sol pueden presentar sobresaturación de ODA. Otros factores como la salinidad, o la altitud (a causa del cambio de la presión) también afectan.



Imagen de <https://www.residuosprofesional.com/wp-content/uploads/2015/11/segre-300x225.jpg>

En un análisis rutinario realizado por la confederación hidrográfica, se evaluó la variabilidad en las medidas de oxígeno disuelto en una zona de un río muy cercano a determinada zona contaminada. Por este motivo, se consideran 3 transectos escogidos al azar en la zona de influencia. Para cada transecto se escogen dos días al azar y se hacen dos medidas cada día.

Los resultados obtenidos de ODA, en porcentaje de saturación (%) están indicados en la tabla adjunta.

| Transsecte | Dia | Mesura | ODA |
|------------|------|--------|-----|
| 1 | 1/7 | 1 | 42 |
| 1 | 1/7 | 2 | 44 |
| 1 | 20/7 | 1 | 39 |
| 1 | 20/7 | 2 | 40 |
| 2 | 12/7 | 1 | 53 |
| 2 | 12/7 | 2 | 54 |
| 2 | 22/7 | 1 | 58 |
| 2 | 22/7 | 2 | 57 |
| 3 | 15/7 | 1 | 86 |
| 3 | 15/7 | 2 | 85 |
| 3 | 31/7 | 1 | 84 |
| 3 | 31/7 | 2 | 85 |

Asumiendo normalidad y homogeneidad de varianzas, y considerando un nivel de significación de 0.05, y utilizando los listados apropiados cuando corresponda, se pide:

- 1) Describe el diseño y el modelo asociado. A partir de la tabla ANOVA adecuada, estudiar la significación de los efectos o de los componentes de la varianza, planteando las hipótesis correspondientes e indicando la resolución de las correspondientes pruebas de hipótesis.

Según el anunciado, cada transecto se realizará en días diferentes escogidos al azar. Por tanto se está considerando un factor aleatorio "transecto", jerárquicamente por encima de un segundo factor, también aleatorio "día". Transecto tiene, $a=3$ niveles y día tiene $b=2$ niveles, para cada nivel de transecto, los dos días escogidos al azar para realizarlo. Cada transecto realizado en un día concreto resulta finalmente en $n=2$ observaciones o réplicas de ODA, que es la variable de respuesta o dependiente.

El modelo lineal es:

valor observat d'ODA a la rèplica $k = 1, 2$ de la mesura pel transecte $i = 1, \dots, a = 3$, el dia $j = 1, \dots, b = 2$

↓ mitjana general o constant

↓ ↓ residu per la repetició k al transecte i el dia j

$$Y_{ijk} = \mu + A_i + B_{j(i)} + e_{k(ij)}$$

↑ ↑
↑ efecte aleatori del dia j pel transecte i

efecte aleatori del transecte i

$$\left. \begin{array}{l} e_{k(ij)} \sim N(0, \sigma) \text{ iid} \\ B_{j(i)} \sim N(0, \sigma_{\text{dia}}) \text{ iid} \\ A_i \sim N(0, \sigma_{\text{transecte}}) \text{ iid} \end{array} \right\} \text{m\u00fatuament independents.}$$

La tercera tabla ANOVA de los listados, marcada en amarillo, es la \u00fanica compatible con el dise\u00f1o descrito antes, fijaros especialmente en los g.d.l. $a - 1 = 3 - 1 = 2$ por transecto y $a(b - 1) = 3(2 - 1) = 3$ por d\u00eda. Dado que ambos factores son aleatorios la completamos obteniendo cada F observada o experimental dividiendo por el cuadrado medio justo por debajo jer\u00e1rquicamente. Por ejemplo, por transecto: $F = 1991,58 / 9,75 = 204,26$.

Las hip\u00f3tesis b\u00e1sicas se plantean sobre las componentes de las varianzas, dado el car\u00e1cter aleatorio de los factores estudiados:

$$H_0 : \sigma_{\text{transecte}}^2 = 0 \text{ vs } H_1 : \sigma_{\text{transecte}}^2 > 0$$

Dado que $F = 204,26 > F_{0,05}(2,3) = 9,55$, p -valor $< 0,05$, podemos rechazar la hip\u00f3tesis nula, queda establecido que hay una componente de la varianza ligada a transecto.

Similarmente:

$$H_0 : \sigma_{\text{dia}}^2 = 0 \text{ vs } H_1 : \sigma_{\text{dia}}^2 > 0$$

$F = 13 > F_{0,05}(3,6) = 4,76$ podemos rechazar la hip\u00f3tesis nula, queda establecida que hay una componente de la varianza ligada a d\u00eda.

- 2) Estimar los par\u00e1metros del modelo: efectos -si los hay- componentes de la varianza -si los hay.

$$\hat{\mu} = \bar{Y}_{...} = 60,83.$$

Estimación de los componentes de la varianza:

$$\hat{\sigma}^2 = MS_E = 0,75$$

$$\hat{\sigma}_{dia(transsecte)}^2 = \frac{MS_{dia(transsecte)} - MS_E}{n} = \frac{9,75 - 0,75}{2} = 4,5$$

$$\hat{\sigma}_{transsecte}^2 = \frac{MS_{transsecte} - MS_{dia(transsecte)}}{bn} = \frac{1991,58 - 9,75}{2 \cdot 2} = 495,46$$

LISTADOS

Analysis of Variance Table

Response: ODA

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|--------|
| Transecte | 2 | 3983.2 | | | |
| Di a | 1 | 0.1 | | | |
| Transecte: Di a | 2 | 29.2 | | | |
| Residuals | 6 | 4.5 | | | |

Analysis of Variance Table

Response: ODA

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|--------|
| Transecte | 2 | 3983.2 | | | |
| Di a | 1 | 0.1 | | | |
| Residuals | 8 | 33.7 | | | |

Analysis of Variance Table

Response: ODA

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|--------|
| Transecte | 2 | 3983.2 | 1991.58 | 204.26 | < 0,05 |
| Transecte: Di a | 3 | 29.3 | 9.75 | 13.0 | < 0,05 |
| Residuals | 6 | 4.5 | 0.75 | | |

> dummy.coef(ODA.aov)

Full coefficients are

| | | | | |
|--------------------|------------|-----------|------------|--|
| (Intercept): | 60.83333 | | | |
| Transecte: | 1 | 2 | 3 | |
| | -14.33333 | -11.33333 | 25.66667 | |
| Di a_1: | -0.1666667 | | | |
| Transecte: Di a_1: | 1 | 2 | 3 | |
| | -3.3333333 | 4.1666667 | -0.8333333 | |

3.8. Microbiología de los quesos

El queso se elabora mediante fermentación bacteriana de leche pasteurizada. Las variedades de queso dependen, en gran medida, de las bacterias que se añadan controladamente a los cultivos. Pero, juntamente con estas bacterias, llamadas “iniciales”, en el producto final se suelen encontrar otras bacterias, llamadas “no iniciales” o “salvajes”. En un estudio se investigó la influencia, en las propiedades nutritivas de determinada clase de queso, del hecho de añadir intencionadamente en el proceso de fermentación dos de las cepas bacterianas “salvajes” detectadas más comúnmente: R50#10 i R21#2. En un experimento se elaboraron 12 quesos asignando al azar 3 muestras de leche en cada una de las siguientes condiciones experimentales:

- a) No se añade ninguna cepa bacteriana “salvaje”,
- b) Se añade solamente R21#2,
- c) Se añade solamente R50#10,
- d) Se añade una combinación de R50#10 y R21#2.

La tabla siguiente muestra el contenido en aminoácidos libres, una medida de la actividad bacteriana, relacionada con la calidad nutritiva de un queso. Los datos se refieren a cada una de las unidades elaboradas:

| | | R21#2 | | | | | |
|--------|---------|--------|---------|-------|-------|-------|-------|
| | | Absent | Present | | | | |
| R50#10 | Absent | 1.697 | 1.601 | 1.830 | 2.211 | 1.673 | 1.973 |
| | Present | 2.032 | 2.017 | 2.409 | 2.091 | 2.255 | 2.987 |

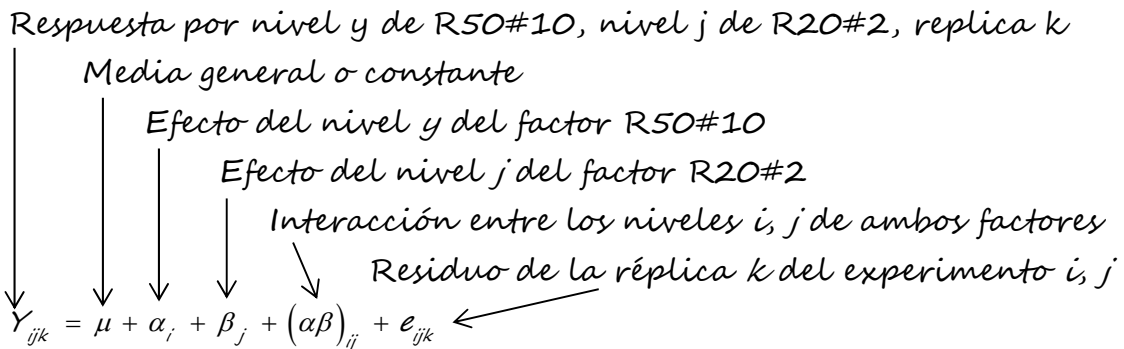


Imagen de https://www.inapi.cl/images/default-source/default-album/queso-noticia-2.jpg?sfvrsn=d393c0f_0

- 1) Indica el tipo de diseño experimental, los factores que intervienen, su carácter (fijo o aleatorio), los niveles y la variable de respuesta (1 punto).

Diseño completamente aleatorizado, cruzado, balanceado. Con dos factores fijos: R50#10, con dos niveles, Absent/Present y R20#2 con dos niveles con los mismos nombres y significados. La variable de respuesta es el contenido de aminoácidos libres.

- 2) Indica el modelo lineal asociado y la parametrización (1 punto).



On:

$i = 1, \dots, a = 2, j = 1, \dots, b = 2$ (1 = Absent, 2 = Present) i k = 1, \dots, n = 3
 $\sum_{i=1}^a \alpha_i = 0, \sum_{j=1}^b \beta_j = 0,$
 $\sum_{j=1}^b (\alpha\beta)_{ij} = 0$ per $i = 1, \dots, a$ i $\sum_{i=1}^a (\alpha\beta)_{ij} = 0$ per $j = 1, \dots, b$
 totes les e_{ijk} independents i idènticament distribuïdes $e_{ijk} \sim N(0, \sigma)$

Parametrizació: $\mu, \alpha_{Absent}, \beta_{Absent}, (\alpha\beta)_{Absent, Absent}, \sigma^2$ (5 paràmetres desconeguts) (o 10 en total)

- 3) Completa la tabla ANOVA y estudia la significación de los efectos del modelo. Indica claramente las hipótesis planteadas y las conclusiones a las cuales se llega bajo un nivel de significación de 0.05. Puedes asumir que se verifican las condiciones habituales de normalidad y homogeneidad de varianzas (2 puntos).

Analysis of Variance Table

Response: Amioacids

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | F Tables |
|-----------------|----|---------|---------|---------|--------|------------------------|
| R50. 10 | 1 | 0.65614 | 0.65614 | 7.2335 | < 0.05 | $F_{0.05}(1,8) = 5.32$ |
| R21. 2 | 1 | 0.21440 | 0.21440 | 2.3636 | > 0.05 | 5.32 |
| R50. 10: R21. 2 | 1 | 0.00178 | 0.00178 | 0.0196 | > 0.05 | 5.32 |
| Residuals | 8 | 0.72566 | 0.09071 | | | |

(para obtener las F experimentales dividimos todos los MS por MS residual).

$$H_0 : \alpha_{\text{Absent}} = \alpha_{\text{Present}} = 0 \text{ vs. } H_1 : \alpha_{\text{Absent}} \neq \alpha_{\text{Present}} (\neq 0)$$

El valor F experimental es de 7.2335 que es superior al valor en las tablas para un nivel de significación 0.05 y 1 y 8 g.d.l., 5.32. Por tanto, el p -valor seguro que es inferior a 0.05 y se puede rechazar H_0 por el efecto de R50.10.

Para los otros efectos, el valor de las tablas es el mismo 5.32.

$$H_0 : \beta_{\text{Absent}} = \beta_{\text{Present}} = 0 \text{ vs. } H_1 : \beta_{\text{Absent}} \neq \beta_{\text{Present}} (\neq 0)$$

$F = 2.3636 < 5.32$, no tenemos evidencias para rechazar H_0 .

$$H_0 : (\alpha\beta)_{\text{Absent},\text{Absent}} = (\alpha\beta)_{\text{Absent},\text{Present}} = (\alpha\beta)_{\text{Present},\text{Absent}} = (\alpha\beta)_{\text{Present},\text{Present}} = 0$$

$$H_1 : \text{alguna} \neq 0$$

$F = 0.0196 < 5.32$, no tenemos suficiente evidencia para afirmar que haya interacción.

- 4) Estima todos los parámetros del modelo. Puede ser de utilidad la siguiente tabla de medias (1 punto).

| |
|-------------------------------|
| Tables of means Grand mean |
|-------------------------------|

| |
|-----------------------|
| 2.064667 |
| R50.10 |
| R50.10 |
| Absent Present |
| 1.8308 2.2985 |
| R21.2 |
| R21.2 |
| Absent Present |
| 1.9310 2.1983 |
| R50.10: R21.2 |
| R21.2 |
| R50.10 Absent Present |
| Absent 1.7093 1.9523 |
| Present 2.1527 2.4443 |

$$\begin{aligned}
\hat{\mu} &= \bar{Y}_{\dots} = 2.064667 \\
\hat{\alpha}_{Ab\ se\ nt} &= \bar{Y}_{Ab\ se\ nt..} - \bar{Y}_{\dots} = \\
& 1.8308 - 2.064667 = -0.23383 \\
\hat{\beta}_{Ab\ se\ nt} &= \bar{Y}_{.Ab\ se\ nt.} - \bar{Y}_{\dots} = \\
& 1.9310 - 2.064667 = -0.13367 \\
(\widehat{\alpha\beta})_{Ab\ se\ nt., Ab\ se\ nt.} &= \bar{Y}_{Ab\ se\ nt., Ab\ se\ nt.} - \bar{Y}_{Ab\ se\ nt..} - \bar{Y}_{.Ab\ se\ nt.} + \bar{Y}_{\dots} = \\
& 1.7093 - 1.8308 - 1.9310 + 2.064667 = \\
& 0.012167 \\
\hat{\sigma}^2 &= MS_E = 0.09071
\end{aligned}$$

Por el hecho de sumar 0, los restantes coeficientes del modelo serían los mismos que los anteriores, salvo por el signo.

También sería correcto considerar que las $\hat{\beta}_j$ i $(\widehat{\alpha\beta})_{ij}$ son 0 por el hecho de no haber rechazado la correspondiente H_0 . No podemos estar seguros de que sean realmente 0 (pueden haber sólo error de tipo II), pero es una posibilidad justificable que conduce a un modelo más simple.

- 5) Estima la probabilidad de que un queso elaborado añadiendo tanto R50#10 como R21#2 tenga un contenido en aminoácidos libres superior a 2.441 (1 punto).

Bajo esta condición la media estimada de aminoácidos libres es de $\hat{\mu}_{present,present} = 2.4443$, muy cercana del valor 2.441. Dado que estamos asumiendo que esta variable aleatoria sigue una distribución normal, y en la normal la mediana y la media coinciden, es de esperar que el resultado sea ligeramente más grande que 0.5, como podemos intuir haciendo un simple dibujo de la función de densidad normal. Esto nos puede servir de pista para ver si lo hemos hecho bien.

De forma más precisa, estimamos que la distribución de la variable es:

$$N\left(\hat{\mu}_{\text{present,present}} = 2.4443, \hat{\sigma} = \sqrt{MS_E} = \sqrt{0.09071} = 0.301\right)$$

Por tanto,

$$\Pr\{Y > 2.441\} = \Pr\left\{Z > \frac{2.441 - 2.4443}{0.301}\right\} = \Pr\{Z > -0.01\} = \Pr\{Z \leq 0.01\} = 0.504$$

per $Z \sim N(0, 1)$

3.8. Fertilizantes orgánicos y nitrógeno

Problema 1. Los fertilizantes orgánicos utilizados en agricultura, al descomponerse, liberan nitrógeno, fósforo y potasio, que promueven el crecimiento del fitoplancton que sirve de alimento a los peces. Una empresa de agricultura compra ingredientes orgánicos para la fertilización de depósitos de peces a través de tres proveedores diferentes, siempre los mismos (los únicos proveedores posibles). Dado que ha habido problemas con algunos clientes que han observado diferencias de la calidad de una compra a otra, la empresa de agricultura quiere estudiar los factores que pueden influir en las propiedades de los ingredientes. Por ello hace un estudio consistente en seleccionar al azar cuatro lotes de ingrediente orgánico de cada uno de los proveedores (por tanto, los lotes son diferentes de un proveedor a otro). Para cada lote, el laboratorio de la empresa hace tres determinaciones independientes del nivel de nitrógeno (ésta es la única medida de calidad del agua en la que nos fijaremos). El aparato medidor está calibrado para dar un valor entre -10 y 10. Los datos obtenidos son:

| Proveedor 1 | | | | | Proveedor 2 | | | | | Proveedor 3 | | | | |
|-------------|----|----|----|---|-------------|----|---|----|---|-------------|---|----|----|---|
| lote | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 |
| | 1 | -2 | -2 | 1 | | 1 | 0 | -1 | 0 | | 2 | -2 | 1 | 3 |
| | -1 | -3 | 0 | 4 | | -2 | 4 | 0 | 3 | | 4 | 0 | -1 | 2 |
| | 0 | -4 | 1 | 0 | | -3 | 2 | -2 | 2 | | 0 | 2 | 2 | 1 |

A pesar de que estas medidas parecen altamente discretas, en todo el problema daremos por válidas los supuestos habituales de normalidad y de homoscedasticidad. Utilizaremos siempre un nivel de significación de 0.05.



Imagen de https://fotos02.laopinioncoruna.es/2008/04/14/318x200/2008-04-21_IMG_2008-04-14_10:29:23_58f1.jpg

Utilizando los listados adecuados cuando sean necesarios, responde a las siguientes cuestiones:

- 1) Describe el diseño experimental y el modelo lineal asociado, procurando ajustar tu respuesta a los siguientes espacios:

| | | | |
|---|--|---|--|
| Nombre del diseño | <i>Diseño jerárquico a dos factores, balanceado. (Sería un estudio observacional ya que no hay asignación aleatoria de los niveles de los factores, no hacía falta decirlo, pues no se pregunta explícitamente).</i> | | |
| Nombre de cada factor: Proveedor Lote (Proveedor) | Carácter fijo o aleatorio: Fijo Aleatorio | Nombre de los niveles: 1, 2, 3 1, 2, 3, 4 (pero diferentes para cada nivel de "Proveedor") | |
| Número de replicas | $n = 3$ | | |
| Variable de respuesta: | $Y = \text{nivel de nitrógeno}$ | | |
| Modelo lineal: | | | |

Nivell de nitrògen per la rèplica k del lot j del proveïdor i

↓ efecte del proveïdor i

↓ ↓

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{ijk} \leftarrow \text{residu aleatori de la determinació o}$$

↑ ↑ rèplica k del lot j del proveïdor i

↑ efecte aleatori del lot j provinent del proveïdor i

mitjana general o constant

$$i = 1, 2, 3; a = 3$$

$$j = 1, 2, 3, 4; b = 4$$

$$k = 1, 2, 3; n = 3$$

$$\sum_{i=1}^a \alpha_i = 0$$

$$\left. \begin{array}{l} \beta_{j(i)} \sim N(0, \sigma_B) \text{ iid} \\ e_{ijk} \sim N(0, \sigma) \text{ iid} \end{array} \right\} \text{ independents}$$

Parametrizació del model:

$$\mu, \alpha_1, \alpha_2, \alpha_3, \sigma_B^2, \sigma^2$$

- 2) Estudia la significació de los efectos del modelo. En concreto: copia aquí y completa la tabla ANOVA correcta para esta situación. Por cada test de la tabla ANOVA, indica las hipótesis nula y la alternativa, y la conclusión final.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|----|--------|---------|--------------------------|---------------------|
| Proveedor | 2 | 15.056 | 7.5278 | 0.9690 = 7.5278 / 7.7685 | > 0.05 ⁷ |
| Proveedor:Lote | 9 | 69.917 | 7.7685 | 2.9439 = 7.7685 / 2.6389 | < 0.05 ⁸ |
| Residuals | 24 | 63.333 | 2.6389 | | |

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0 \text{ vs } H_1: \alpha_i \neq \alpha_{i'}, \text{ per algun } i \neq i', i, i' = 1, 2, 3$$

No podem rebutjar H_0 , no hem pogut demostrar que hi hagi efecte de proveïdor.

$$H_0: \sigma_B^2 = 0 \text{ vs } H_1: \sigma_B^2 > 0$$

Podem rebutjar H_0 , sota un nivell de significació 0.05, podem afirmar que una part de la variabilitat del nivell de nitrògen és atribuïble al lot

- 3) Calcula la probabilidad de que una determinación de nitrógeno hecha sobre material del proveedor 2 sea superior a 4.

⁷ $0.9690 < F_{0.05}(2, 9) = 4.26$ en las tablas

⁸ $2.9439 > F_{0.05}(9, 24) = 2.30$ en las tablas

Aquí hay dos posturas razonables y que se pueden dar por válidas en la solución de esta pregunta,, siempre que se justifiquen. La primera sería considerar que, dado que no se ha podido demostrar que hay diferencias entre proveedores, la estimación más razonable del nivel de nitrógeno medio es la media general, 0.3611111 según los listados. La segunda postura es considerar que a pesar de todo, se puede haber cometido un error de tipo TT (que en realidad sí que sean diferentes y no se hayan podido detectar), y que la mejor estimación es la media particular del proveedor 2, 0.3333 según los listados⁹. Haremos los cálculos según esta segunda posibilidad. En la otra únicamente sería necesario cambiar el valor de la media. En cualquier caso, lo que sí parece claro es que la varianza del nivel de nitrógeno tiene una componente residual y una atribuible al lote.

$$\hat{\sigma}_{\text{Nitrogen}}^2 = \hat{\sigma}_{\text{Lot}}^2 + \hat{\sigma}^2 = 1.709877 + 2.638889.$$

$$\hat{\sigma}_{\text{Lot}}^2 = \frac{MS_{\text{Lot}} - MS_{\text{Residual}}}{n} = \frac{7.7685 - 2.6389}{3} = 1.709877$$

$$\hat{\sigma}^2 = MS_{\text{Residual}} = 2.6389$$

$$P(Y_2 > 4) = P\left\{Z > \frac{4 - 0.3333}{\sqrt{1.709877 + 2.638889}}\right\} = 1 - P(Z \leq 1.76) = 0.039$$

$$Z \sim N(0,1)$$

(Continuación del problema.) Supongamos ahora que no hay 3 proveedores diferentes, sino que todo el material procede de un único proveedor. Se cogen al azar 4 lotes de materia prima proporcionados por este único proveedor (ahora tenemos únicamente 4 lotes en total), y una parte de cada lote se lleva a un laboratorio externo, en total 3 laboratorios (que son los únicos posibles, no una muestra aleatoria). De cada lote, cada laboratorio hace tres determinaciones independientes de nitrógeno, en las mismas condiciones que antes. Por comodidad supondremos que los valores numéricos de los valores obtenidos son los mismos, pero con una interpretación diferente.

⁹ Una tercera posibilidad sería utilizar un estimador "sintético" basado en una combinación lineal de los dos anteriores, pero este tema no se ha tratado

$$\left. \begin{array}{l} B_{ij} \sim N(0, \sigma_B) \quad iid \\ (\alpha B)_{ij} \sim N(0, \sigma_{AB}) \quad iid \\ e_{ijk} \sim N(0, \sigma) \quad iid \end{array} \right\} \text{ independents}$$

(D'acord amb allò fet habitualment a classe, no considerarem que hi hagi cap restricció addicional als efectes d'interacció, no farem servir el model restringit. Si algú l'utilitza cal que ho digui.)

Parametrización del modelo:

$$\mu, \alpha_1, \alpha_2, \alpha_3, \sigma_A^2, \sigma_{AB}^2, \sigma^2$$

5) Estudia la significación de los efectos del modelo. En concreto copia aquí y completa la tabla ANOVA correcta para esta situación. Para cada test de la tabla ANOVA, indica las hipótesis nula y la alternativa, y la conclusión final.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) ¹⁰ |
|-----------------|----|--------|---------|--------------------------|----------------------|
| Laboratorio | 2 | 15.056 | 7.5278 | 1.0201 = 7.5278 / 7.3796 | > 0.05 |
| Lote | 3 | 25.639 | 8.5463 | 1.1581 = 8.5463 / 7.3796 | > 0.05 |
| Laboratori:Lote | 6 | 44.278 | 7.3796 | 2.7965 = 7.3796 / 2.6389 | < 0.05 |
| Residuals | 24 | 63.333 | 2.6389 | | |

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad \text{vs} \quad H_1 : \alpha_i \neq \alpha_{i'}, \text{ per algun } i \neq i', i, i' = 1, 2, 3$$

No podem rebutjar H_0 , no hem pogut demostrar que hi hagi efecte de laboratori.

$$H_0 : \sigma_B^2 = 0 \quad \text{vs} \quad H_1 : \sigma_B^2 > 0$$

No podem rebutjar H_0 , no hem pogut demostrar que una part de la variabilitat del nivell de nitrògen sigui atribuïble al lot per si sol.

$$H_0 : \sigma_{AB}^2 = 0 \quad \text{vs} \quad H_1 : \sigma_{AB}^2 > 0$$

Podem rebutjar H_0 , sembla que una part de la variabilitat del nitrogen es pot atribuir a una interacció entre lot i laboratori.

6) Estima todos los parámetros del modelo.

¹⁰ Tal como se ha comentado en el apartado 4, seguimos el criterio de considerar el modelo no restringido. En caso contrario la tabla ANOVA correcta sería:

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|--------|
| Laboratorio | 2 | 15.056 | 7.5278 | 1.0201 | > 0.05 |
| Lote | 3 | 25.639 | 8.5463 | 3.2386 | < 0.05 |
| Laboratori:Lote | 6 | 44.278 | 7.3796 | 2.7965 | < 0.05 |
| Residuals | 24 | 63.333 | 2.6389 | | |

A partir del listado de medias (último listado del problema):

$$\begin{aligned}\hat{\mu} &= \bar{Y}_{\dots} = 0.3611 \\ \hat{\alpha}_1 &= \bar{Y}_{1..} - \bar{Y}_{\dots} = -0.4167 - 0.3611 = -0.7778 \\ \hat{\alpha}_2 &= \bar{Y}_{2..} - \bar{Y}_{\dots} = 0.3333 - 0.3611 = -0.0278 \\ \hat{\alpha}_3 &= -(-0.7778 - 0.0278) = 0.8056\end{aligned}$$

(o proponer el valor 0 para las $\hat{\alpha}_i$, dados los resultados del apartado 4)

$$\begin{aligned}\hat{\sigma}^2 &= MS_{\text{residual}} = 2.6389 \\ \sigma_{AB}^2 &= \hat{\sigma}_{\text{Laboratori, Lot}}^2 = \frac{MS_{\text{Laboratori, Lot}} - MS_{\text{Residual}}}{n} = \frac{7.3796 - 2.6389}{3} = 1.5802 \\ \hat{\sigma}_{\text{Lot}}^2 &= \frac{MS_{\text{Lot}} - MS_{\text{Laboratori, Lot}}}{an} = \frac{8.5463 - 7.3796}{3 \cdot 3} = 0.1296\end{aligned}$$

(o proponer el valor 0 para la estimación de la varianzade lote, dados los resultados del apartado anterior.)

LISTADOS

Response: Ni trogen

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|--------|
| Provei dor | 2 | 15.056 | 7.5278 | | |
| Lot | 3 | 25.639 | 8.5463 | | |
| Provei dor: Lot | 6 | 44.278 | 7.3796 | | |
| Resi dual s | 24 | 63.333 | 2.6389 | | |

Response: Ni trogen

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|--------|
| Provei dor | 2 | 15.056 | 7.5278 | | |
| Provei dor: Lot | 9 | 69.917 | 7.7685 | | |
| Resi dual s | 24 | 63.333 | 2.6389 | | |

Response: Ni trogen

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|------------------|----|--------|---------|---------|--------|
| Laboratori | 2 | 15.056 | 7.5278 | | |
| Lot | 3 | 25.639 | 8.5463 | | |
| Laboratori : Lot | 6 | 44.278 | 7.3796 | | |
| Resi dual s | 24 | 63.333 | 2.6389 | | |

Response: Ni trogen

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|------------------|----|--------|---------|---------|--------|
| Laboratori | 2 | 15.056 | 7.5278 | | |
| Laboratori : Lot | 9 | 69.917 | 7.7685 | | |
| Resi dual s | 24 | 63.333 | 2.6389 | | |

Tables of means
Grand mean

0.3611111

Provei dor
Provei dor

| | 1 | 2 | 3 |
|--|---------|--------|--------|
| | -0.4167 | 0.3333 | 1.1667 |

Provei dor: Lot
Lot

| Provei dor | 1 | 2 | 3 | 4 |
|------------|---------|---------|---------|--------|
| 1 | 0.0000 | -3.0000 | -0.3333 | 1.6667 |
| 2 | -1.3333 | 2.0000 | -1.0000 | 1.6667 |
| 3 | 2.0000 | 0.0000 | 0.6667 | 2.0000 |

Tables of means
Grand mean

0.3611111

Laboratori
Laboratori

| | 1 | 2 | 3 |
|--|---------|--------|--------|
| | -0.4167 | 0.3333 | 1.1667 |

Lot
Lot

| | 1 | 2 | 3 | 4 |
|--|--------|---------|---------|--------|
| | 0.2222 | -0.3333 | -0.2222 | 1.7778 |

Laboratori : Lot
Lot

| Laboratori | 1 | 2 | 3 | 4 |
|------------|---------|---------|---------|--------|
| 1 | 0.0000 | -3.0000 | -0.3333 | 1.6667 |
| 2 | -1.3333 | 2.0000 | -1.0000 | 1.6667 |
| 3 | 2.0000 | 0.0000 | 0.6667 | 2.0000 |

4.Cuestiones conceptuales (tipo test)

4.1. Primer test

Debes de marcar la única opción correcta

(La siguiente cuestión tiene relación con el enunciado del problema de los quesos)

1) Supongamos que los datos del problema anterior no se hubieran obtenido asignando al azar las muestras de leche a las cuatro posibles condiciones experimentales, sino que se hubiera cogido una muestra aleatoria de 12 quesos ya elaborados y se hubieran clasificado en cuatro grupos, según si no estaban “contaminados” ni por R50#10 ni por R21#2, únicamente por R21#2, únicamente por R50#10, o por ambas cepas “salvajes” (el objetivo del estudio continua siendo estudiar la presencia o no de R50#10 y R21#2 concretamente, por ejemplo no haciendo caso, ignorando, la presencia de cualquier otro bacteria). Supongamos que casualmente hubiera 3 quesos de cada clase y que los contenidos en aminoácidos libres fueran los indicados en la tabla de datos del problema.

- a. En este caso tendríamos un estudio observacional, no experimental (*en un estudio observacional el experimentador no controla la asignación de los niveles del factor a estudiar a los sujetos*)
- b. Los factores aleatorios (*no tiene ninguna relación*)
- c. La variable de respuesta no podría ser normal (*no tiene ninguna relación*)
- d. Ninguna de las tres afirmaciones anteriores es correcta.

*(A partir de aquí, las preguntas **no tienen ninguna relación con el problema.**)*

- 2) En una ANOVA de un factor, el p-valor al contrastar la significación del factor ha resultado ser de 0.035.
- a. Si el nivel de significación α se ha fijado en 0.05, aceptamos H_0 (*en este caso se necesitaría rechazar H_0 –aparte que no esté bien dicho que “se acepta” H_0*)

- b. Si el nivel de significación α se ha fijado de 0.01, rechazamos H_0 (si el nivel de significación es 0.01, para rechazar H_0 sería necesario que el p-valor fuera inferior a 0.01)
 - c. La probabilidad que H_0 sea cierta es 0.035 (el p-valor no es la probabilidad que H_0 sea cierta, sino la probabilidad que el estadístico de test –en este caso F– tome valores tan o más extremos que los que ha presentado, si H_0 es cierta)
 - d. Ninguna de las anteriores es cierta.
- 3) Si se aumenta el número de réplicas manteniendo el mismo nivel de significación α , se consigue:
- a. Disminuir la probabilidad de error de tipo I (en una prueba exacta como el ANOVA, la probabilidad de error de tipo I coincide con el nivel de significación previamente fijado, independientemente del tamaño muestral)
 - b. Disminuir la probabilidad de error de tipo II (aumenta la potencia)
 - c. Aumentar la probabilidad de error de tipo I (vale el mismo comentario que en la primera opción)
 - d. Aumentar la probabilidad de error de tipo II (disminuye la potencia).
- 4) Indica cuál de las siguientes afirmaciones es cierta (si alguno no lo ve claro, que repase el esquema sobre las decisiones que se pueden tomar en un contraste de hipótesis y los errores asociados)
- a. El nivel de significación se define como: Prob (Aceptar H_0 | si H_0 es cierta)
 - b. La potencia de test define como: Prob (Rechazar H_0 | si H_1 es cierta)
 - c. El error de tipo II es rechazar la hipótesis nula cuando es cierta.
 - d. La probabilidad de error de tipo II corresponde al nivel de significación.
- 5) En el típico modelo lineal asociado a un diseño ANOVA de un único factor aleatorio, (con la distribución normal de los efectos aleatorios, homogeneidad de varianzas, etc.) la variable de respuesta Y_{ij} (donde el subíndice i corresponde a los niveles del factor y el subíndice j en las réplicas)
- a. Se distribuye según una normal, con media que puede variar según los niveles del factor.

- b. Se distribuye según una normal, y su varianza depende de los niveles del factor.
- c. Tiene la misma varianza que el residuo
- d. Se distribuye según una normal, con distribución que no depende de los niveles del factor.

Recordad que el modelo lineal asociado a un diseño de un solo factor aleatorio $Y_{ij} = \mu + A_i + e_{ij}$, donde las A_i y las e_{ij} son normales independientes, con media 0 y con varianzas σ_A^2 y σ^2 respectivamente. Como consecuencia la variable de respuesta será también normal, con media $E(Y_{ij}) = \mu + 0 + 0$ y varianza $\text{var}(Y_{ij}) = 0 + \sigma_A^2 + \sigma^2$. Esta media y esta varianza son siempre las mismas, independientemente de "i", es decir, del nivel del factor.

4.2 Test 2

Cuestiones conceptuales. Selecciona la única opción correcta y justifica brevemente la respuesta.

Pregunta 1.- Sea una ANOVA de 2 factores cruzados con 3 niveles en cada factor y con 3 réplicas por condición experimental. En la tabla ANOVA, el p-valor para la interacción es 0.012. Se verificará que:

- a) El valor del estadístico F (F_{exp}) es inferior a 0.05
- b) Se rechaza la hipótesis nula de ausencia de interacción
- c) El valor del estadístico F (F_{exp}) es inferior a 4.58
- d) Ninguna de las anteriores

Justificación:

Para 4 y 18 g.d.l., que serían los que corresponderían a la F de interacción, en la tabla de la F para una probabilidad en la cola derecha de 0.01 encontramos el valor 4.58. Un valor de F experimental, para tener asociado un p-valor (es decir una área de probabilidad en la cola derecha a partir de él), más grande que 0.01, como por ejemplo 0.012, debe ser más pequeño que 4.58, debe estar a su izquierda. Esto se ve fácilmente a partir de un esquema gráfico (aunque no es necesario justificarlo, fijémonos que la opción "a" es absurda –puede ser una confusión entre F y p-valor–, y respecto a la opción "b", recordad que nadie nos ha dicho qué nivel de significación se está considerando.)

Pregunta 2.- Si en un análisis de la varianza se está realizando de forma correcta (cumplimiento de las condiciones, etc.) se aumenta el número de réplicas manteniendo el mismo nivel de significación α , se consigue:

- a) *Disminuir la probabilidad de error de tipo I*
- b) *Disminuir el valor de la suma de cuadrados residual en la tabla ANOVA*
- c) *Aumentar la probabilidad de error de tipo I*
- d) **Ninguna de las anteriores**

Justificación:

No validez de las otras opciones:

- a) *En una prueba paramétrica exacta con el ANOVA, la probabilidad de error de tipo I coincide con el nivel de significación prefijado y no varía con el tamaño muestra. No lo confundáis con la probabilidad de error de tipo II y la potencia que sí que varía con el tamaño muestral.*
- b) *Probablemente la suma de cuadrados residual aumentará ya que hay más sumandos, pero tampoco podemos decir una cosa segura, ya que también las medias muestrales dentro de cada casilla serán diferentes. Pero ciertamente no podemos asegurar que disminuirá.*
- c) *El mismo comentario que en "a".*

Pregunta 3.- Tenemos un experimento que comparan 4 dosis de un fármaco. Se considera que el hecho de ser fumador puede también tener influencia en la variable de respuesta. Así pues, se toman a 5 individuos fumadores, y 5 no fumadores diferentes para cada una de las dosis (por tanto, 40 individuos en total), y para cada individuo se observa (una única vez) el valor de la variable de respuesta:

- a) *"Individuo" es un factor, cruzado con el factor "dosis"*
- b) *"Ser fumador" es un factor aleatorio*
- c) **Los individuos definen las réplicas, en este caso no son estudiados como un factor.**
- d) *Ninguna de las anteriores*

Justificación:

Cada individuo nos proporciona en un único valor de la variable de respuesta, por cada combinación de ser fumador o no y de dosis tenemos valores replicados gracias a tener más de un individuo –si de alguna

manera se hubiera repetido la medida dentro de cada individuo, entonces que "individuo" sería estudiable como un factor.

A parte de no ser ningún factor en este caso, "individuo" no está cruzado con "dosis", ya que para cada dosis, tenemos individuos diferentes, no pasa que cada individuo se somete a todas las dosis "Ser fumador" es un factor fijo, son las únicas opciones (niveles) posibles en un planteamiento como este.

4.2. Segundo test

Preguntas conceptuales

Cada pregunta vale 1 punto. Únicamente se debe justificar la respuesta si es pide explícitamente.

- 1) En una prueba ANOVA de un solo factor fijo con 5 niveles y 7 réplicas para cada nivel, por un **nivel de significación 0.01**. **NO** se ha podido rechazar la hipótesis nula. Si suponemos que el procedimiento se ha aplicado correctamente, indica cuáles de las siguientes afirmaciones **seguro que son falsas** (puede haber más de una):
 - a. El p-valor obtenido es 0.023
 - b. **Estamos cometiendo el error de tipo I**
 - c. El cuadrado medio residual es muy pequeño, y vale 0,00135.
 - d. El valor del estadístico F observado, calculado sobre los datos, es inferior a 3.
 - e. **El valor del cuadrado medio residuales el indicado en "c" y el valor del cuadrado medio para el factor es 0,03.**

A pesar de que no era necesario justificar la respuesta, comentamos un poco las opciones:

"a" un p-valor como este es perfectamente compatible con no rechazar la hipótesis nula si el nivel de significación es 0.01 tal y como se indica al principio.

"b" el error de tipo I sólo se puede cometer si rechazamos la hipótesis nula (por error), por tanto aquí no es aplicable.

“c” no es suficiente el cuadrado medio residual para tomar la decisión, se necesita también el cuadrado medio asociado al factor para formar el coeficiente F y, en función de lo que valga, acabar de decidir

“d” según la tabla de la distribución F por nivel de significación $0,01$ y para 4 y 30 g.d.l., el valor crítico es $3,25$. Por lo tanto, un valor observado de F inferior a 3 sería perfectamente compatible con no rechazar la hipótesis nula.

“e” con estos cuadrados medios, el coeficiente F sería $22,22$, que para un valor crítico $3,25$ conduciría a rechazar la hipótesis nula.

2) Estamos analizando dos factores cruzados, mediante el análisis de la varianza y utilizando un programa estadístico como por ejemplo R. Son datos obtenidos en un diseño totalmente aleatorizado, con más de una réplica por casilla. Se ha visto que se cumplen perfectamente las condiciones de validez de la prueba ANOVA, excepto que los residuos eran claramente no normales. Esta falta de normalidad de los residuos puede provocar que (una única opción es cierta):

- a. La distribución del estadístico F ya no siga una F de Fisher-Snedecor bajo hipótesis nula cierta
- b. No sea posible estimar la varianza residual
- c. Los cuadrados medios de la tabla ANOVA están mal calculados
- d. No es posible calcular el cuadrado medio de interacción
- e.

Justificación de la opción escogida como cierta:

La normalidad de los residuos es una de las condiciones para garantizar que la distribución del estadístico F sea la indicada bajo el supuesto que H_0 sea cierta, recordad que para establecer los otros resultados que tienen relación con las opciones “b” o “d” no amentaron para nada la normalidad, son resultados puramente algebraicos.

- 3) Selecciona cuáles de las siguientes afirmaciones son ciertas
- a. El error de tipo I corresponde a rechazar la hipótesis nulas H_0 , cuando es falsa

- b. La potencia de test se define como la probabilidad de rechazar H_0 si la hipótesis alternativa es cierta
- c. El error de tipo II es no rechazar la hipótesis nula cuando es falsa
- d. La probabilidad de error de tipo II se denomina p-valor.

Si no lo veis claro, repasad estos conceptos básicos, y si conviene discutidlo con los compañeros o con los profesores.

5. Bibliografía

5. Bibliografia

- ALCAÑIZ M., GUILLEM M., PONS E. i RUFINO H. Problemes d'estadística descriptiva. Publicacions i Edicions UB. ISBN 978-84-475-3816-4.
- Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26: 32-46.
- Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2017. *Lme4: Linear Mixed-Effects Models Using 'Eigen' and S4*. <https://CRAN.R-project.org/package=lme4>.
- BESALÚ M i ROVIRA C. 2013. Probabilitats i estadística. Publicacions i Edicions UB. ISBN: 978-84-475-3653-5.
- BONFILL, X.; BARNADAS, A.; CARREÑO, A.; ROSER, M.; MONLEÓN, T.; MORAL, I.; HERNÁNDEZ-BOLUDA, J. C.; PÉREZ, P.; LLOMBART, A. (Coordinació Roset, M. / Monleón, T.) (2004). Metodología de investigación y estadística en oncología y hematología. EDIMAC.
- Cortés J. Cobo E, Rufino H, Vilaró M y González JA. 2014. Control del riesgo alfa en Diseños adaptativos. Estadística para no estadísticos. UPC. https://ocw.upc.edu/sites/all/modules/ocw/estadistiques/download.php?file=715001/2013/1/54994/t14_control_del_riesgo_alfa-5228.pdf
- Crawley, M.J. 2002. *Statistical Computing: An Introduction to Data Analysis Using S-PLUS*
- CUADRAS, C. M. [et al.]. Fundamentos de estadística. Aplicación a las ciencias humanas. Barcelona : EUB, 1996.
- CUADRAS, C. M. Problemas de probabilidades y estadística. Vol. 1: Probabilidades. ed. corregida i ampliada Barcelona : EUB, 1999-2000.
- CUADRAS, C. M. Problemas de probabilidades y estadística. Vol. 2: Inferencia estadística. ed. corregida i ampliada. Barcelona : EUB, 1999-2000.
- DAVIS, J. C. *Statistics and data analysis in geology*. 3a ed. New York [etc.] : Wiley, cop. 2002. JEKHOWSKY, B. *DE Éléments de statistique à l'usage des géologues*. París : Technip, 1977.
- de Mendiburu, Felipe. 2017. *Agricolae: Statistical Procedures for Agricultural Research*. <https://CRAN.R-project.org/package=agricolae>.
- Excoffier, L., P.E. Smouse, and J.M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, 131:479-491.
- Efron, B. (1979). "Bootstrap methods: Another look at the jackknife." *Annals of Statistics*. Vol. 7, pp.1-26.
- F. Konietzschke, L.A. Hothorn, E. Brunner: Rank-Based Multiple Test Procedures and Simultaneous Confidence Intervals. *Electronic Journal of Statistics*, Vol.0 (2011) 1-8.
- Gao, X. et al. (2008). Nonparametric Multiple Comparison Procedures for Unbalanced One-Way Factorial Designs. *JSPI* 138, 2574 - 2591.

- Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2016. *Multcomp: Simultaneous Inference in General Parametric Models*. <https://CRAN.R-project.org/package=multcomp>.
- JULIÀ DE FERRAN O., MÁRQUEZ-CARRERAS D., ROVIRA ESCOFET C. i SARRÀ ROVIRA M. 2003. Probabilitats: problemes i més problemes. Publicacions i Edicions UB. ISBN 978-84-475-2906-3.
- JULIÀ O. i MÁRQUEZ-CARRERAS D. Un primer curs d'estadística. 2011. Publicacions i Edicions UB. ISBN: 978-84-475-3483-8.
- Konietzschke, F. (2009). Simultane Konfidenzintervalle fuer nichtparametrische relative Kontrastefekte. PhD-thesis, University of Goettingen.
- Konietzschke, F., Brunner, E., Hothorn, L.A. (2008). Nonparametric Relative Contrast Effects: Asymptotic Theory and Small Sample Approximations, Research report.
- Konietzschke, F., Pauly, M. (2012). A Studentized Permutation Test for the Nonparametric Behrens- Fisher Problem in Paired Data. *Electronic Journal of Statistic*, Vol 6, 1358-1372
- Konietzschke, F., Placzek, M., Schaarschmidt, S., Hothorn, L.A. (2014). nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. *Journal of Statistical Software*, 61(10), 1-17
- Kuehl, R.O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*. Duxbury/Thomson Learning.
- Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2016. *LmerTest: Tests in Linear Mixed Effects Models*. <https://CRAN.R-project.org/package=lmerTest>.
- Legendre, P. and M.J. Anderson. 1999. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69:1-24.
- Matthews RA. Significance levels for the assessment of anomalous phenomena. *Journal of Scientific Exploration*. 1999;13(1):1-7
- McArdle, B.H. and M.J. Anderson. 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82: 290-297.
- Meier L. 2018. ANOVA: A Short Intro Using R. 2018 (Ver en: <https://stat.ethz.ch/~meier/teaching/anova/>)
- MILLER, R. L.; KAHN, H. N. *Statistical analysis in the geological sciences*. New York : Wiley, 1962. SÁNCHEZ ALGARRA, P. (coord.) *Métodos estadísticos aplicados*. Barcelona. Publicacions i Edicions de la Universitat de Barcelona, DL 2006.
- MONLEON T, RODRIGUEZ C. Probabilitat i estadística per a ciències I. PPU. BARCELONA. 2015.
- MONLEON T, RODRIGUEZ C. Probabilitat i estadística per a ciències II. PPU. BARCELONA. 2017.
- MONLEÓN, A.; BARNADAS, A.; ROSET, M.; HERNÁNDEZ-BOLUDA, J. C.; PÉREZ, P.; LOMBART, A. (coordinació Monleón, T.) (2005). *Metodología de investigación y estadística en oncología y hematología (Casos Prácticos)*. EDIMAC.

- Monleón, T. (2009). Manual de introducción a la simulación de ensayos clínicos. (pp. 1 - 182). Publicacions i Edicions UB. ISBN: 97978-84-475-3181.
- MONLEÓN, T. (2010). El tratamiento numérico de la realidad. Reflexiones sobre la importancia actual de la estadística en la Sociedad de la Información. Arbor, 186(743), 489-497. ISSN: 0210-1963.
- MONLEÓN, T.; BARNADAS, A.; ROSET, M. (2008). «Diseños secuenciales y análisis intermedio en la investigación clínica: tamaño versus dificultad». Medicina Clínica, 11(132), 437-442. http://www.clinicasdenorteamerica.com/clinicas/ctl_servlet?_f=3&pident=13134609 ISSN: 0025-7753
- MONLEÓN, T.; COLLADO, M. (2008). «Técnicas de control estadístico de procesos y software. Calidad industrial del trigo y la harina». Alimentación, Equipos y Tecnología, 238, 50-54. ISSN: 0212-1689
- Monleón-Getino T. 2017. Diseño y planificación de estudios científicos: Calidad de datos (data management) y principios de diseño experimental). Lulu Press, Inc. ISBN: 9781326918194
- Montgomery, D.C. 2012. *Design and Analysis of Experiments*. John Wiley & Sons, Incorporated.
- Munzel, U., Brunner, E. (2002). An Exact Paired Rank Test. Biometrical Journal 44, 584-593.
- Munzel. U., Hothorn, L.A. (2001). A unified Approach to Simultaneous Rank Tests Procedures in the Unbalanced One-way Layout. Biometric Journal, 43, 553-569.
- Oehlert, Gary. 2010. *A First Course in Design and Analysis of Experiments*. <http://users.stat.umn.edu/~gary/book/fcdae.pdf>.
- Quinn, Gerry P., and Michael J. Keough. 2010. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/> (<http://www.R-project.org/>)
- Rios D, Monleon-Getino T, Cubedo M, Rios M. 2017. A Graphical Classification of European Countries According to Physical Activity Level of Its Citizens. Open Access Library Journal 03(12):1-11
- Rodriguez CI, Monleon-Getino T. 2016. A new R library for discriminating groups based on abundance profile and biodiversity in microbiome metagenomic matrices. Article in International Journal of Scientific and Engineering Research 7(10):243-253
- Rodriguez-Casado, Clara; Monleon-Getino, Toni; Cubedo, Marta; Rios-Alcolea, Martin (Journal of Mathematics, 2017). A priori groups based on Bhattacharyya distance and partitioning around medoids(PAM) with applications to metagenomics.
- Ryu, E. (2009): Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. Statistics In Medicine, 28(25), 3179-3188.

- SÁNCHEZ ALGARRA, P. (coord.) Métodos estadísticos aplicados. Barcelona. Publicacions i Edicions de la Universitat de Barcelona, DL 2006.
- Scheipl, Fabian, and Ben Bolker. 2016. *RLRsim: Exact (Restricted) Likelihood Ratio Tests for Mixed and Additive Models*. <https://CRAN.R-project.org/package=RLRsim>.
- Speed, M, R Hocking, and P Hackney. 1978. "Methods of Analysis of Linear Models with Unbalanced Data." *Journal of the American Statistical Association* 73 (361): 105–12. doi:10.1080/01621459.1978.10480012.
- STIGLER, S. M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press/Harvard University Press. ISBN 0-674-40341-X.
- Wickham, Hadley, and Winston Chang. 2016. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. <https://CRAN.R-project.org/package=ggplot2>.
- Wu, C.F.J. (1986). "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis." *Annals of Statistics*. Vol. 14, No. 4, pp.1261 - 1295.
- Zapala, M.A. and N.J. Schork. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences, USA*, 103:19430-19435.