ANOVA course (Session 2)

S.Civit

21 de enero de 2020

1 Objetivos:

-**Objetivo 5: ANOVA 2-way crossed and nested designs**. Emplearemos en lo posible **tidyverse** y **dplyr** para realizar nuestros análisis de datos.

2 What is ANOVA?

La técnica de técnicas denominada Análisis de la varianza (ANOVA), del acrónimo Analysis of variance: ANalysis Of VAriance, tiene como objetivo básico la comparación de las medias de más de dos poblaciones.

En el ANOVA se comparan siempre las medias de varias poblaciones y se hace a través de un contraste de hipótesis donde se analiza la varianza, es cierto; pero no sólo eso, porque también se analizan las diferencias de medias que hay entre las muestras, y también, por supuesto, como siempre en Estadística, se analiza el tamaño de muestra.

3 Vocabulario básico:

3.1 Definir objetivo del experimento a analizar

3.2 Variable respuesta: Cuantitativa continua

3.3 Factor.

Un factor en ANOVA es una **variable cualitativa que genera o que contempla una serie de poblaciones a comparar**. Por ejemplo, se ensayan tres tipos de fertilizantes en unos campos de cultivo para evaluar la productividad, se ensayan cuatro medicamentos distintos para ver si aumentan los niveles de hemoglobina en pacientes con anemia. En estos casos tenemos, en primer lugar el factor tipo de fertilizante. En el segundo, el factor fármaco.

3.4 Niveles del factor

Los niveles de un factor son los **grupos o poblaciones que genera** un factor. En el primer ejemplo anterior tenemos tres niveles. En el segundo tenemos cuatro niveles.

3.5 Tipo de factor

Un factor es fijo si los niveles que tenemos de él en el estudio son realmente todos los que nos interesa comparar. Un factor es aleatorio si los niveles que tenemos en nuestro estudio es una muestra de niveles tomados de una población de niveles que son los que, en realidad, queremos comparar. Los dos ejemplos anteriores si los tres fertilizantes o los cuatro fármacos son nuestro objeto de comparación, estamos ante factores fijos. Pero, observemos lo siguiente: si en otro ejemplo, estoy comparando si hay diferencias en la calidad de un producto fabricado por 100 operarios trabajando en una industria y, para hacerlo, elijo al azar a 5 de esos 100 operarios y analizo 3 productos elaborados por cada uno de ellos, pero lo que me interesa es ver si hay diferencias entre los 100, no entre esos 5, estoy ante el factor operario con 5 niveles, pero ese factor es, ahora, no fijo, sino aleatorio.

3.6 Réplicas (balanceado no balanceado)

Número de experimentos realizados por condición experimental. Si todos los grupos (niveles del factor) tienen el mismo número de réplicas se denomina **diseño balanceado**, de gran trascendecia en el uso de funciones de R.

3.7 Comparaciones múltiples/Componentes de la varianza

Si tenemos un factor fijo y detectamos que hay diferencias entre esas poblaciones, nos interesará decir cuáles son esas diferencias concretas. Las comparaciones múltiples hacen esa labor, comparan, dos a dos, de una forma muy especial, todas las poblaciones para dibujar un mapa de las diferencias. Si tenemos un factor aleatorio, el planteamiento es ahora muy diferente: debemos pasar de la muestra de muestras de poblaciones que tenemos a una población de poblaciones y eso lo haremos estimando la varianza, la dispersión que debe haber dentro de esa población de poblaciones.

3.8 Factores cruzados/Factores anidados

Cuando hay más de un factor en un estudio, los factores, dos a dos, pueden estar cruzados o anidados. Tenemos factores cruzados cuando todos los niveles de un factor están combinados con todos los niveles del otro factor. Tenemos factores anidados cuando los niveles de un factor están jerarquizados entre los niveles del otro factor.

3.9 Interacción entre factores

Cuando los factores están cruzados podemos estudiar algo muy importante en ANOVA: la interacción entre esos factores. Hay interacción cuando la respuesta, el efecto conseguido con la presencia de un nivel de un factor, depende de con qué nivel del otro factor esté combinado.

3.10 Randomización

Este es un concepto clave “en el diseño de cualquier experimento des de la vertiente estadística y por tanto elemento clave en el Diseño experimental que no es otro que la”asignación aleatoria" de los unidades experimentales a los grupos (niveles del factor). Se denomina **completely randomized design (CRD)**.

Si queremos **randomizar** factor tratamiento a 4 niveles (treatment factor, 4 levels) A,B,C and D en un total de 20 unidades experimentales (a balanced design with 5 replicates), obtendriamos con el siguiente codigo como “realizar el experimento”

This means that the first experimental unit will get treatment xxx, the second xxx and so on.

```
treat.ord <- rep(c("A", "B", "C", "D"), each = 5) ## could also use LETTERS[1:4]
treat.ord
```

```
##  [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C" "D" "D" "D" "
D"
## [20] "D"
```

```
sample(treat.ord) ## random permutation
```

```
##  [1] "D" "C" "A" "B" "D" "C" "B" "B" "B" "C" "C" "A" "A" "A" "A" "C" "B" "D" "
D"
## [20] "D"
```

3.11 Ejemplo “conceptual”

Se toma una muestra de 30 alumnos durante toda la ESO.

Se dividen en tres clases distintas (definimos tres “líneas distintas de aprendizaje”). Cada una va a seguir, durante los cuatro años (duración ESO), un plan distinto de enseñanza del inglés.

Se sabe el nivel escrito y el nivel oral de esos alumnos al final de la primaria. Se han diferenciado dos niveles dentro de cada grupo, según el promedio de notas globales de esos alumnos ha sido alto o bajo, en el global de las materias.

Durante los cuatro cursos de la ESO se ha hecho un seguimiento, alumno por alumno, del nivel de inglés oral de esos alumnos.

```
-La variable estudiada es el nivel de inglés oral.
-Hay dos factores fijos: Grupo y Nivel.
-Grupo a tres niveles y Nivel a dos niveles.
-Los dos factores son fijos y están cruzados.
```

Fuera del scope del curso

```
-Hay tercer factor: el factor ESO, con cuatro niveles fijos.
-Grupo y Nivel son intersujetos.
-El factor ESO es intrasujetos.
Las variables InglésEscrito e InglésOral a finales de primaria podría tratarse com
o covariable.
```

4 Fixed Effects Models: two-Way ANOVA

The two-way ANOVA model with interaction is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

We assume a general setup with a factor A with a levels, a factor B with b levels and n replicates for every combination of A and B (a balanced design). Hence, we have a total of $N = a \cdot b \cdot n$ observations. In the next example we had $a = 2, b = 2, n = 3$ and therefore $N = 12$.

The effects can be interpreted as follows:

Think of **main effects** as **average effects** on the expected value of the response when changing the level of a factor (keeping the **other factor fixed**).

The interaction effect can be thought of as a correction factor to the main effects model. The interaction effect **tells us how much an effect of a certain factor changes, when we “switch” the level of the other factor**. Strictly speaking, interpretation depends on the side-constraint that we apply.

A model without interaction term is additive. This means that the effect of A does not depend on the level of B (and vice versa), **it is always the same, no matter what the level of the other factor**.

An experiment was performed how intentionally added bacteria affect cheese quality. Two strains of bacteria, “R50” and “R21” were investigated. None (control treatment), both or one of the strains were being added (R50 with levels “yes” and “no” and R21 with levels “yes” and “no”). Total free amino acids was measured as response (acids).

-Questions to solve:

With a factorial treatment structure we typically have questions about both factors and/or their possible interplay:

```
"Does the effect of adding R50 depend on whether we have added R21 or not (or vice versa)?"

-If the effects do not depend on the other factor, we could also ask:

-What is the effect of adding R50 on the expected value of the total free amino acids?”

"What is the effect of adding R21#10 on the expected value of the total free amino acids?"
```

4.1 ANOVA table

```
# Create data (skip if not interested) ####
acids <- c(1.697, 1.601, 1.830,
           2.032, 2.017, 2.409,
           2.211, 1.673, 1.973,
           2.091, 2.255, 2.987)
R50 <- rep(c("no", "yes", "no", "yes"), each = 3)
R21 <- rep(c("no", "no", "yes", "yes"), each = 3)
cheddar <- data.frame(R50, R21, acids)
summary(cheddar)
boxplot(acids ~ R50, data=cheddar)
boxplot(acids ~ R21, data=cheddar)
options(contrasts = c("contr.sum", "contr.poly"))
fit.cheddar <- aov(acids ~ R50 * R21, data = cheddar)
coef(fit.cheddar)
summary(fit.cheddar)
#TukeyHSD(fit.cheddar, which="R50")
#plot(TukeyHSD(fit.cheddar, conf.level = 0.99),las=1, col = "red")
#write.table(cheddar, "cheddar.txt")
```

4.2 Interaction plot

We can visualize such kind of data with a so called **interaction plot** using the **function interaction.plot**.

For every combination of R50 (“yes” / “no”) and R21 (“yes” / “no”) we calculate the average value of the response.

We use R50 on the x-axis (the first argument, also called x.factor). In addition, settings corresponding to the same level of R21 are connected with lines (argument trace.factor). The role of R50 and R21 can of course also be interchanged.

```
#standard way
interaction.plot(x.factor = cheddar$R50, trace.factor = cheddar$R21,response = che
ddar$acids)
## elegant way, using the function "with"
with(cheddar, interaction.plot(x.factor = R50, trace.factor = R21, response = acid
s))

##ggplot way
library(ggplot2)
ggplot(cheddar, aes(x = R50, y = acids, color = R21)) + geom_point() +
  stat_summary(fun.y = mean, geom = "line", aes(group = R21), size = 1) + theme_bw
()
```

Here, there is no evidence of an interaction between R50 and R21. Only the main effect of R50 is significant. This finding is consistent with what we have observed in the interaction plot. There, the two lines were (almost) parallel. Adding R50 (significantly) increases the free amino acids. For R21 the interaction plot suggests a larger amount of free amino acids if we add it. However, the effect is not significant. Hence it could very well have happened just by chance.

5 Checking Model Assumptions

5.1 Normality

```
options(contrasts = c("contr.sum", "contr.poly"))
fit.cheddar <- aov(acids ~ R50 * R21, data = cheddar)

#First graph (Residuals vs fitted)
plot(fit.cheddar, which = 1, add.smooth = FALSE)

#Second graph (qqplot)
plot(fit.cheddar, which = 2, add.smooth = FALSE)
#Hypothesis test
shapiro.test(fit.cheddar$residuals)
```

5.2 Homocedasticity

```
library(car)
leveneTest(fit.cheddar)
```

6 Fixed Effects Models: two-Way ANOVA. Unbalanced data

We have a look at some (simulated) data about a sports experiment using a factorial design with factor gender and factor energy drink (having two levels). Response was running time in seconds for a specific track. Clearly, this is an unbalanced data set.

```
library(xlsx)
running <- read.table("http://stat.ethz.ch/~meier/teaching/data/running.dat", head
er = TRUE)
str(running)
```

```
## 'data.frame':    70 obs. of  3 variables:
##  $ gender: Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ drink : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
##  $ y     : num  40.6 49.7 42.1 42.2 39 44.2 44.1 43.1 44.7 46.3 ...
```

```
summary(running)
```

```
##     gender   drink        y
##  female:50   A:20   Min.   :39.00
##  male  :20   B:50   1st Qu.:52.25
##                     Median :57.05
##                     Mean   :56.01
##                     3rd Qu.:61.15
##                     Max.   :66.20
```

```
write.xlsx(running, "running.xlsx")
```

We use contr.sum, otherwise type III sum of squares will be wrong (technical issue).

```
options(contrasts = c("contr.sum", "contr.poly"))
## Type I sum of squares
fit0 <- aov(y ~ gender * drink, data = running)
summary(fit0)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## gender        1 2024.0  2024.0 263.719  < 2e-16 ***
## drink         1  455.2   455.2  59.316 9.05e-11 ***
## gender:drink  1   29.1    29.1   3.791   0.0558 .
## Residuals    66  506.5     7.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we change the order of the factors in the model formula.

```
options(contrasts = c("contr.sum", "contr.poly"))
## Type I sum of squares
fit2 <- aov(y ~ drink * gender, data = running)
summary(fit2) ## sum of squares change!
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## drink         1 1145.9  1145.9 149.299 <2e-16 ***
## gender        1 1333.4  1333.4 173.737 <2e-16 ***
## drink:gender  1   29.1    29.1   3.791 0.0558 .
## Residuals    66  506.5     7.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the sum of squares depend on the ordering of the model terms in the model formula if we use a type I approach. Hence, we also get different F-ratios and different p-values. However, the p-values of the main effects stay very small here, no matter what “ordering” we use.

6.1 ANOVA Table

We could also use the Anova function for type III sum of squares, but drop1 will do the job too.

```
library(car)
```

```
## Loading required package: carData
```

```
fit <- aov(y ~ gender * drink, data = running)
Anova(fit, type = "II", data = running)
```

	Sum Sq <dbl>	Df <dbl>	F value <dbl>	Pr(>F) <dbl>
gender	1333.41197	1	173.736516	3.730542e-20
drink	455.24807	1	59.316412	9.052786e-11
gender:drink	29.09423	1	3.790824	5.579162e-02
Residuals	506.54400	66	*NA*	*NA*

4 rows

```
# Type III sum of squares
drop1(fit, scope = ~., test = "F", data = running)
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	506.5440	146.5381	*NA*	*NA*
gender	1	1352.40377	1858.9478	235.5489	176.211047	2.653334e-20
drink	1	484.23700	990.7810	191.4999	63.093516	3.346862e-11
gender:drink	1	29.09423	535.6382	148.4475	3.790824	5.579162e-02

4 rows

```
## or:
Anova(fit, type = "III", data = running)
```

	Sum Sq <dbl>	Df <dbl>	F value <dbl>	Pr(>F) <dbl>
(Intercept)	136968.21608	1	17846.233024	4.811825e-82
gender	1352.40377	1	176.211047	2.653334e-20
drink	484.23700	1	63.093516	3.346862e-11
gender:drink	29.09423	1	3.790824	5.579162e-02
Residuals	506.54400	66	*NA*	*NA*

5 rows

7 Checking Model Assumptions

7.1 Normality

```
#First graph (Residuals vs fitted)
plot(fit, which = 1, add.smooth = FALSE)

#Second graph (qqplot)
plot(fit, which = 2, add.smooth = FALSE)
#Hypothesis test
shapiro.test(fit$residuals)
```

7.2 Homocedasticity

```
library(car)
leveneTest(fit)
```

8 Random Effects Models: two-Way ANOVA.

We can extend this to the two-way ANOVA situation. For this reason we consider Example 7.1 in Kuehl (2000). A manufacturer was developing a new spectrophotometer for medical labs. A critical issue is consistency of measurements from day to day among different machines. 4 machines were randomly selected from the production process and tested on 4 randomly selected days. Per day 8 serum samples were randomly assigned to the 4 machines (2 samples per machine). Response is the triglyceride level [mg/dl] of a sample.

We denote the triglyceride level of the kth sample on day i and machine j by Y_{ijk}. We use the model

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}, \quad e_{ijk} \approx N(0, \sigma)$$

Let us fit this in R. We want to have a random effect per day (= (1 | day)), a random effect per machine (= (1

| machine)) and a random effect per combination of machine and day (= (1 | machine:day)).

8.1 Variance components

```
library(lmerTest)
# Create data-set ####
##     machine 1     machine 2     machine 3     machine 4
y <- c(142.3, 144.0, 148.6, 146.9, 142.9, 147.4, 133.8, 133.2, ## day 1
       134.9, 146.3, 145.2, 146.3, 125.9, 127.6, 108.9, 107.5, ## day 2
       148.6, 156.5, 148.6, 153.1, 135.5, 138.9, 132.1, 149.7, ## day 3
       152.0, 151.4, 149.7, 152.0, 142.9, 142.3, 141.7, 141.2) ## day 4

trigly <- data.frame(y = y, day = factor(rep(1:4, each = 8)),
                     machine = factor(rep(rep(1:4, each = 2), 2)))
str(trigly)
summary(trigly)

fit.trigly <- lmer(y ~ (1 | day) + (1 | machine) + (1 | machine:day), data = trigl
y)

summary(fit.trigly)

write.xlsx(trigly, "randomfactor2.xlsx")
#write.table(trigly, "randomfactor2.txt")
```

From the output we get $\hat{\sigma}^2_A = 44.7$ (day), $\hat{\sigma}^2_B$=57.7 (machine),$\hat{\sigma}^2_{AB} = 34.7$ (interaction of day and machine) and $\hat{\sigma}^2$=17.9

Hence, total variance is $44.7 + 57.7 + 34.7 + 17.9 = 155$. We see that the largest contribution to the variance is variability between different machines $(57.34 / (44.7 + 57.7 + 34.7 + 17.9) x 100) = 37.22\%$.

8.2 ANOVA table

```
 # generem l'objecte data.aov, un ANOVA 2F complet
data.aov <- aov(y ~ day*machine, data=trigly)
auxAnova       <- summary(data.aov)[[1]]

# calcula la F fent el quocient MSA/MSAB
auxAnova[1,4]  <- auxAnova$Mean[1]/auxAnova$Mean[3]

# calcula la F fent el quocient MSB/MSAB
auxAnova[2,4]  <- auxAnova$Mean[2]/auxAnova$Mean[3]

# calcul del p-valor MSA/MSAB
auxAnova[1,5]  <- 1-pf(auxAnova[1,4], auxAnova$Df[1],auxAnova$Df[3])

# calcul del p-valor MSB/MSAB
auxAnova[2,5]  <- 1-pf(auxAnova[2,4], auxAnova$Df[2],auxAnova$Df[3])
auxAnova
str(auxAnova)
```

9 Checking Model Assumptions

9.1 Normality

```
#First graph (Residuals vs fitted)
plot(fit.trigly, which = 1, add.smooth = FALSE)

#Second graph (qqplot)
plot(fit.trigy, which = 2, add.smooth = FALSE)
#Hypothesis test
?shapiro.test

shapiro.test(fit.trigly$residuals)
```

9.2 Homocedasticity

```
library(car)
leveneTest(fit.trigly)
```

10 Mixed Effects Models

In practice we often encounter models which contain both random and fixed effects. We call them mixed effects models.

We start with the data set Machines from the package nlme. As stated in the help file: “Data on an experiment to compare three brands of machines used in an industrial process are presented in Milliken and Johnson (p. 285, 1992). Six workers were chosen randomly among the employees of a factory to operate each machine three times. The response is an overall productivity score taking into account the number and quality of components produced.”

```
data("Machines", package = "nlme")
## technical detail for nicer output:
Machines[, "Worker"] <- factor(Machines[, "Worker"], levels = 1:6, ordered = FALS
E)
str(Machines, give.attr = FALSE) ## give.attr in order to shorten output
```

Let us now model this data. We assume that there is a population machine effect (think of an average “profile”), but each worker is allowed to have its own (random) deviation.

With Y_{ijk} we denote the kth productivity score of worker j on machine i. We use the model

$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \epsilon_{ijk}$$

, where α_i is the fixed effect of machine i (with the usual side constraint), B_j is the random effect of worker j and $(\alpha B)_{ij}$ is the corresponding (random) $interaction$.

We could use the lme4 package to fit such a model. We want to have

```
a random effect $B_j$ per worker: (1 | Worker)

a random effect $(\apha B)_{ij}$ per combination of worker and machine: (1 | Worke
r:Machine)
```

Hence, the lmer call would look like this

10.1 Variance components

```
library(car)
options(contrasts = c("contr.treatment", "contr.poly"))
library(lmerTest)
fit <- lmer(score ~ Machine + (1 | Worker) + (1 | Worker:Machine), data = Machine
s)
anova(fit)
#Fixed effects
fixef(fit)
#Random effects
confint(fit, oldNames = FALSE)

#Interaction
#ggplot(Machines, aes(x = Machine, y = score, group = Worker, col = Worker)) +
#  geom_point() + stat_summary(fun.y = mean, geom = "line")

## classical interaction plot would be
#with(Machines, interaction.plot(x.factor = Machine, trace.factor = #Worker, respo
nse = score))
```

As **lme4 does not calculate p-values for the fixed effects**, we use the package **lmerTest** instead. Technically speaking, it is nothing else than a wrapper for the same function in package lme4 but with modified outputs which include p-values. There are still many open issues regarding statistical inference in mixed effects model,

The (fixed) effect of machine is significant. If we closely inspect the output, we see that an F-distrubtion with 2 and 10 degrees of freedom is being used. Also we observe that productivity is largest on machine C, followed by B and A. Most workers show a similar “profile”, with the exception of worker 6 who performs badly on machine B.

A confidence interval for the expected value of the difference between machine A and B is given by [3.74,12.2].

10.2 ANOVA table

```
# generem l'objecte data.aov, un ANOVA 2F complet
data.aov <- aov(score ~ Machine*Worker, data=Machines)
# el segon factor és l'aleatori, sino girar
auxAnova       <- summary(data.aov)[[1]]
# calcula la F fent el quocient MSA/MSAB (model restringit)
auxAnova[1,4]  <- auxAnova$Mean[1]/auxAnova$Mean[3]
#calcul del p-valor
auxAnova[1,5]  <- 1-pf(auxAnova[1,4], auxAnova$Df[1],auxAnova$Df[3])
auxAnova
```

11 Checking Model Assumptions

11.1 Normality

```
## QQ-plots:
par(mfrow = c(1, 3))
qqnorm(ranef(fit)$Worker[, 1], main = "Random effects of worker")
qqnorm(ranef(fit)$'Worker:Machine'[, 1], main = "Random interaction")
qqnorm(resid(fit), main = "Residuals")
shapiro.test(data.aov$residuals)
```

11.2 Homocedasticity

```
plot(fit)
library(car)
leveneTest(score ~ Machine* Worker, data = Machines)
```

12 Wrong Model:

Again, in order to better understand the mixed effects model we check what happens if we would fit a purely fixed effects model here.

```
fit.fixed <- aov(score ~ Machine * Worker, data = Machines)
summary(fit.fixed)
```