Anova
Parte 1: Aplicaciones

Natalie Julian[1]

[1] Estadística de la Pontificia Universidad Católica de Chile, Docente de Jornada parcial Facultad de Matemáticas UC

www.nataliejulian.com

Concepto

Muchas veces tenemos una población que queremos estudiar en términos de características o *factores* cuya relación con la variable respuesta Y no es clara. La población puede separarse en *subpoblaciones*, una por cada nivel i del factor de interés (factor A). Defina como a la cantidad de niveles del factor A a estudiar.

La hipótesis que interesa estudiar es la siguiente:

$$H_0 : \text{Las medias de todas las subpoblaciones son iguales} \iff \mu_1 = \mu_2 = \ldots = \mu_a$$

$$H_1 : \text{Las medias de las subpoblaciones no son todas iguales} \iff \exists(i, i') \quad \mu_i \neq \mu_{i'}$$

Parafraseando lo anterior, se busca determinar evidencia (o no evidencia) suficiente para establecer si el factor induce/*explica*/sugiere diferencias en términos de la variable respuesta.

Algunos ejemplos:

a) Se desea determinar si el monto de crédito gastado de los clientes se vería afectado por si la persona tiene hijos o no tiene hijos. En este caso la variable respuesta es el monto de crédito y la variable factor es si tiene o no tiene hijos. Es decir, interesaría contrastar si las medias de monto de crédito para el grupo tiene hijos y para el grupo no tiene hijos, son iguales.

b) Resulta de interés evaluar si las pastillas anticonceptivas logran generar diferencias en términos del IMC de las personas. La variable respuesta es el IMC y el factor de interés es personas con consumo de pastillas anticonceptivas y personas sin pastillas anticonceptivas. Es decir, interesaría contrastar si el IMC medio para el grupo con pastillas anticonceptivas difere del grupo sin pastillas anticonceptivas.

- n: cantidad total de unidades experimentales (individuos)
- a: número de grupos o niveles del factor de interés
- i: indica el nivel del factor, con $i = 1, ..., a$
- $\bar{Y}$: media global de la variable respuesta (variable de interés)
- n_i: número de individuos en el grupo o nivel i del factor. Notar que $\sum_{i=1}^{a} n_i = n$
- Y_{ij}: observación (o individuo) j dentro del nivel i del factor
- $\bar{Y}_i$: media de la variable respuesta en el grupo o nivel i del factor

Modelo de efectos fijos

Un modelo anova de un factor de efectos fijos es:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

Donde i indica el nivel del factor A, con $i = 1, ..., a$ y a es la cantidad de niveles del factor A. j corresponde al subíndice que recorre las observaciones dentro de cada nivel i, es decir, para $i, j = 1, ..., n_i$. Cuando $n_i = n_{i'}$, es decir, se tiene la misma cantidad de observaciones o *réplicas* por nivel i del factor, denominamos a éste un caso *balanceado*.

Cuando el efecto se asume fijo, uno de los contrastes que puede utilizarse es el contraste suma:

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

Matemáticamente, el contraste suma implica que:

$$\alpha_1 + \alpha_2 + \ldots + \alpha_a = 0$$

Equivalentemente:

$$\alpha_a = -(\alpha_1 + \alpha_2 + \ldots + \alpha_{a-1})$$

Esta restricción permite que el modelo sea identificable.

i. Las unidades experimentales son seleccionadas por muestreo aleatorio simple.
ii. Para cada subpoblación, la variable respuesta posee una distribución normal.
iii. La variabilidad dentro de cada subpoblación, se asume igual.

Estos supuestos **deben** analizarse al plantear un modelo.

Defina como $\bar{Y}$ la media global de la variable respuesta Y e $\bar{Y}_i$ es la media de la variable respuesta por nivel i del factor.

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{a} n_i(\bar{Y}_i-\bar{Y})^2}_{SCFactor} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2}_{SCE}$$

La suma cuadrática del factor (también conocida como variación *between*) indica cuánta es la variabilidad explicada por el factor utilizado.

La suma cuadrática del error (también conocida como variación *within*) es la variabilidad asociada al error de ajuste del modelo planteado.

Las sumas cuadráticas del factor (tratamiento) y del error se definen como sigue:

$$MCA = \frac{SCA}{a-1}$$

$$MCE = \frac{SCE}{n-a}$$

El test F para medir significancia del factor A, se define:

$$F_A = \frac{MCA}{MCE} = \frac{Between}{Within}$$

Es decir, el estadístico es el cuociente entre la variabilidad entre grupos (niveles del factor) respecto de la variabilidad dentro de los grupos. Se espera que si el factor logra explicar diferencias, entonces la variabilidad entre grupos sea alta, por lo que, a grandes valores del estadístico F_A, se espera gran significancia del factor A.

La tabla anova resume toda la información como sigue:

Fuente	Suma cuadrática	G.l	Media cuadrática	Test F
Factor A	$\sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2$	a-1	$MCA = \frac{SCA}{a-1}$	$F_A = \frac{MCA}{MCE}$
Error	$\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	n-a	$MCE = \frac{SCE}{n-a}$	
Total	$\sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	n-1		

Caso aplicado unifactorial de efecto fijo

Se realizó un experimento en el cual se midió el tiempo de demora en una prueba de memoria que rindieron pacientes de un centro médico (seleccionados aleatoriamente). Algunos de estos participantes tenían un tratamiento de Alprazolam o Triazolam y otros no presentaban ningún tratamiento. Se busca medir si el consumo de estas drogas afecta la memoria de las personas respecto al tiempo que tardan en resolver la prueba.

Intuiciones

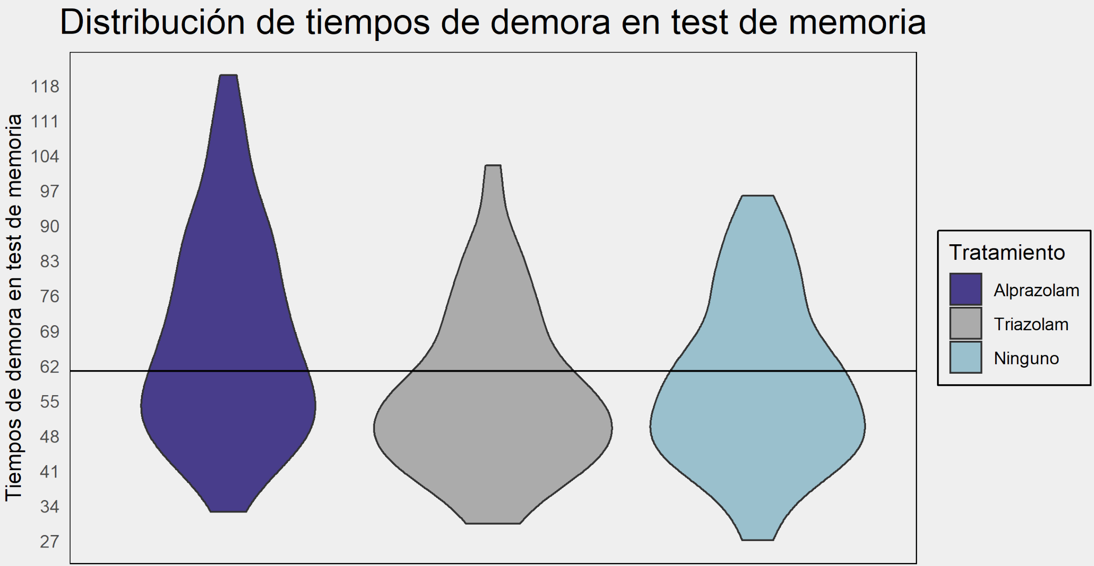
Nuestro factor de interés es la droga, la cual toma los valores *alprazolam*, *triazolam* y *ninguno*. Es decir, la variable factor droga posee a=3 niveles. Considere i el subíndice para indicar los niveles del factor.

Tratamiento	n_i	Time memory test
Alprazolam	67	61.2 40.7 55.1 51.2 47.1 58.1 56.0 74.8 45.0 75.9 102.0 63.7 40.7 84.3 32.8 56.3 44.6 72.5 65.4 49.2 ...
Triazolam	65	46.9 51.4 56.8 42.2 102.0 66.8 50.4 40.5 44.1 41.8 37.9 41.1 74.0 39.0 61.5 65.8 37.8 57.3 52.7 56.8 ...
Ninguno	66	73.3 90.0 64.2 53.6 56.7 61.4 59.0 48.5 50.9 44.1 61.5 81.4 41.7 47.6 45.6 59.2 90.0 62.9 52.1 49.4 ...

Cuando $n_i = n_{i'}$ se denomina caso balanceado, en este caso la cantidad de observaciones por nivel del factor no son iguales, por ende, nos encontraríamos en un caso desbalanceado.

Tratamiento	n_i	Mean time memory test
Alprazolam	67	$\frac{\sum tiempos_{Alprazolam}}{67}$ =67.7
Triazolam	65	$\frac{\sum tiempos_{Triazolam}}{65}$ =56.6
Ninguno	66	$\frac{\sum tiempos_{Ninguno}}{66}$ =58.3

La media global de tiempos de demora en el test de memoria es de 60.92.



Distribución de tiempos de demora en test de memoria
Tiempos de demora en test de memoria
118
111
104
97
90
83
76
69
62
55
48
41
34
27
Tratamiento
Alprazolam
Triazolam
Ninguno

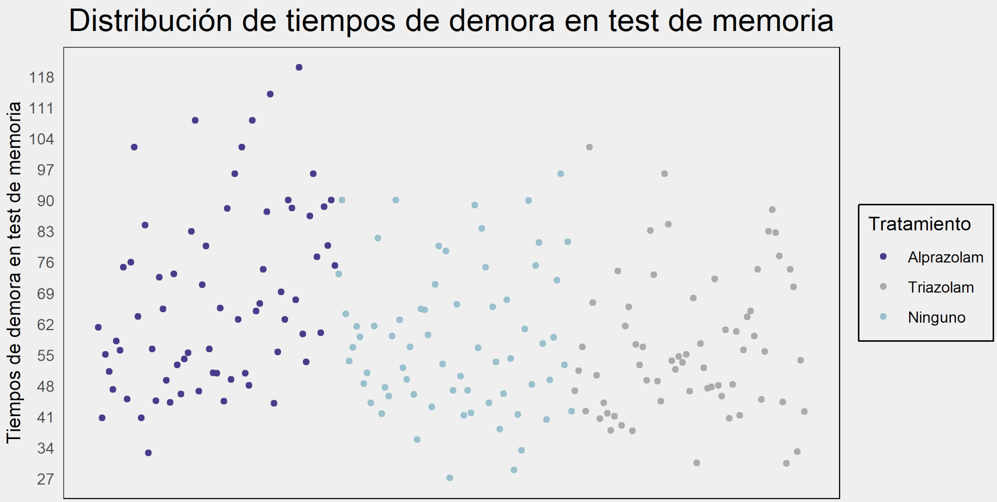
¿Por qué Anova?

Queremos resolver preguntas de investigación tales como:

a) ¿Se observa efecto de uno u otro tratamiento en mi variable respuesta?

b) ¿Existe algún tratamiento más efectivo que otro en términos de tiempo de resolución del test de memoria?

c) La variable factor, ¿es significativa para explicar diferencias en los resultados de la variable de interés?


Distribución de tiempos de demora en test de memoria
Tiempos de demora en test de memoria
118
111
104
97
90
83
76
69
62
55
48
41
34
27
Tratamiento
Alprazolam
Triazolam
Ninguno

Tratamiento	n	Mean	Diferencia con media global
Alprazolam	67	67.7	Mayor que la media global
Triazolam	65	56.6	Menor que la media global
Ninguno	66	58.3	Menor que la media global

El modelo utiliza la información de la media global y de la media en cada grupo o tratamiento.

- a corresponde a la cantidad de grupos o niveles del factor droga, es decir, $a = 3$ (pues tenemos tres tratamientos)
- n corresponde a la cantidad de pacientes en el estudio, con $n = \sum_{i=1}^{a} n_i = 67 + 65 + 66 = 198$
- i indica el grupo o nivel del factor (pudiendo ser Alprazolam, Triazolam y ninguno)
- $\bar{Y}$ media global de la variable respuesta tiempos de demora en el test de memoria
- n_i número de individuos en el grupo o nivel i del factor (con $n_1 = 67, n_2 = 65, n_3 = 66$
- Y_{ij} corresponde al paciente j que posee un tratamiento i (alprazolam, triazolam o ninguno)
- $\bar{Y}_i$ media de la variable respuesta tiempos de demora en el test de memoria por grupo o droga

El modelo teóricamente es:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

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

Con $i = 1, 2, 3,\ \ j = 1, ..., n_i$ y $n_i = 67, 65, 66.$

```
Time<-drugs$‘Memory score‘
Drug<-factor(drugs$Drug)   #Se debe definir como variable de tipo factor

contrasts(Drug)<-contr.sum #contraste suma

model<-aov(Time~Drug)
```

El modelo anova bajo el contraste suma puede entenderse de la siguiente forma:

$$\underbrace{\begin{bmatrix} Y_{1\,1} \\ Y_{1\,2} \\ \vdots \\ Y_{1\,67} \\ Y_{2\,1} \\ Y_{2\,2} \\ \vdots \\ Y_{2\,65} \\ Y_{3\,1} \\ Y_{3\,2} \\ \vdots \\ Y_{3\,66} \end{bmatrix}}_{Variable\ respuesta} = \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}}_{Matriz\ de\ diseno} \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}}_{Vector\ parametros} + \underbrace{\begin{bmatrix} \epsilon_{1\,1} \\ \epsilon_{1\,2} \\ \vdots \\ \epsilon_{1\,67} \\ \epsilon_{2\,1} \\ \epsilon_{2\,2} \\ \vdots \\ \epsilon_{2\,65} \\ \epsilon_{3\,1} \\ \epsilon_{3\,2} \\ \vdots \\ \epsilon_{3\,66} \end{bmatrix}}_{Errores\ asociados}$$

Note que al considerar el contraste suma, no se especifica α_3 pues este queda escrito en término de los demás coeficientes α_1 y α_2. Así, la matriz de diseño se reduce a una columna asociada al intercepto y a las columnas asociadas al producto con α_1 y α_2.

```
model.matrix(model) #Matriz de diseño
    (Intercept) Drug1 Drug2
1             1     1     0
2             1     1     0
3             1     1     0
4             1     1     0
5             1     1     0
6             1     1     0
7             1     1     0
8             1     1     0
9             1     1     0
10            1     1     0
11            1     1     0
12            1     1     0
13            1     1     0
14            1     1     0
15            1     1     0
16            1     1     0
17            1     1     0
18            1     1     0
19            1     1     0
20            1     1     0
```

Efectos estimados

```
coef(model)  #coeficientes estimados, bajo contraste suma entrega alpha1 y alpha2
(Intercept)       Drug1       Drug2
  60.866268    6.815822   -4.263191

levels(Drug)
[1] "A" "T" "N"

#alpha1 se asocia al primer nivel del factor Droga (Drug1), es decir Alprazolam
#alpha2 se asocia al segundo (Drug2), es decir, Triazolam

#Bajo el contraste suma, alpha3=-(alpha1+alpha2)

#alpha3 (Drug3: No droga)

-sum(coef(model)[2:3])
[1] -2.552631
```

Los coeficientes estimados $\hat{\alpha}_i$ obtenidos son:

Tratamiento	α_i	Diferencia con media global
Alprazolam	6.81	67.7-60.92=6.78
Triazolam	-4.26	56.6-60.92=-4.32
Ninguno	-2.55	58.3-60.92=-2.62

Note que estos efectos estimados se relacionan estrechamente con la diferencia de la media por grupo respecto a la media global.

α_i es el efecto asociado al nivel i del factor tratamiento.

Si $\alpha_i > 0$ se observa un efecto positivo, es decir, en ese grupo se observa en general, tiempos de demora en el test de memoria mayores a la media global.

Si $\alpha_i < 0$ se observa un efecto negativo, es decir, en ese grupo, se observan en general, tiempos de demora en el test de memoria menores a la media global.

Testear que el factor es significativo es equivalente a testear si existe algún efecto distinto de cero:

$$H_0 : \alpha_i = 0 \quad \forall\, i$$

$$H_1 : \exists \ \alpha_i \neq 0$$

El estadístico para realizar el test es:

$$F_A = \frac{MCTrat}{MCE} \sim F^{0.95}_{(a-1),\,(n-a)}$$

Se rechaza que los efectos sean iguales a cero si $F_A > F^{0.95}_{(a-1),\,(n-a)}$ o si el valor-p asociado a $P(F_A \leq F^{0.95}_{(a-1),\,(n-a)})$ es menor a 0.05 (significancia estadística usual).

Note que en el ejemplo práctico, el factor A corresponde al factor droga, posee 3 niveles, es decir, $a = 3$ y además el total de observaciones es 198, por lo tanto, el cuantil teórico de la F sería:

$$F^{0.95}_{(3-1),\ (198-3)}$$

El test F de modelos anidados es útil para comparar cuál es el cambio de la suma cuadrática asociada al error al considerar un modelo más pequeño (o restringido) versus el modelo más grande (o completo).

$$H_0 \ : \text{Modelo1: solo con intercepto es correcto}$$

$$H_1 \ : \text{Modelo2: con intercepto y factor es correcto}$$

El estadístico asociado a este test es:

$$F = \frac{(SCE_{\text{Modelo 1}} - SCE_{\text{Modelo 2}}/(glE_{\text{Modelo 1}} - glE_{\text{Modelo 2}})}{SCE_{\text{Modelo 2}}/glE_{\text{Modelo 2}}}$$

Regla de decisión:

$$\text{Si} \ \ F > F_{(}glE_{\text{Modelo 1}} - glE_{\text{Modelo 1}}, glE_{\text{Modelo 2}}) \ \ \text{ o si el valor-p es menor a 0.05}$$

Se rechaza la hipótesis nula.

```
#Manualmente

SCEm1<-anova(modelo1)[1,2]  #Suma cuadrática error modelo 1
glEm1<-anova(modelo1)[1,1]  #Grados de libertad asociados al error
SCEm2<-anova(modelo2)[2,2]  #Suma cuadrática error modelo 2
glEm2<-anova(modelo2)[2,1]  #Grados de libertad asociados al error

#Plantear el Estadistico F:

(F=((SCEm1-SCEm2)/(glEm1-glEm2))/(SCEm2/glEm2))
[1] 7.668025

#Regla de decisión

F>qf(0.95,glEm1-glEm2,glEm2)  #El estadístico es mayor al cuantil
[1] TRUE

#Por ende, se rechaza la hipótesis nula. Existe evidencia para rechazar que el modelo restringido
# (sólo con intercepto) fuera el correcto en términos de bondad de ajuste.

#Opción rápida:

anova(modelo1,modelo2) #Realiza test F de modelos anidados
Analysis of Variance Table

Model 1: Time ~ 1
Model 2: Time ~ Drug
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    197 64781
2    195 60057  2    4723.3 7.668 0.0006227 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla Anova

```
anova(model)
Analysis of Variance Table

Response: Time
           Df Sum Sq Mean Sq F value    Pr(>F)
Drug        2   4723 2361.65   7.668 0.0006227 ***
Residuals 195  60057  307.99
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Suma cuadrática	Expresión	Valor de tabla anova
Tratamiento	SCTrat	4723
Residuos	SCE	60057
Total	SCT	4723+60057= 64780

Note que el tratamiento logra explicar $\frac{4723}{64780} = 0.072$ de la variabilidad total. Valor bastante mejorable.

El test de Tukey compara la diferencia de las medias de la variable respuesta por niveles del factor de interés. Se dice test de comparaciones múltiples pues se testea de a pares, si las medias de la variable respuesta por niveles son iguales:

$$H_0 : \mu_i = \mu_{i'} \text{ con } i \neq i' \quad H_1 : \mu_i \neq \mu_{i'}$$

En R:

```
TukeyHSD(modelo2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Time ~ Drug)

$Drug
          diff        lwr       upr     p adj
T-A -11.079013 -18.294955 -3.863071 0.0010657
N-A  -9.368453 -16.556594 -2.180312 0.0067073
N-T   1.710559  -5.532251  8.953370 0.8425971
```

Es posible obtener un intervalo de confianza de manera sencilla en R:

```
confint.lm(model)
                  2.5 %      97.5 %
(Intercept) 58.406360 63.3261752
Drug1        3.350124 10.2815200
Drug2       -7.755245 -0.7711362
```

Caso aplicado multifactorial de efectos fijos

Un tema de interés para una escuela fue analizar la participación de sus estudiantes. Para ésto se midió la cantidad de veces que los alumnos hicieron preguntas en el salón de clases en un periodo de tiempo fijo. Además, se cree que la participación en clases pudiera verse influenciada por el sexo del alumno, el apoderado del alumno (padre o madre) y la satisfacción del apoderado respecto a la escuela.

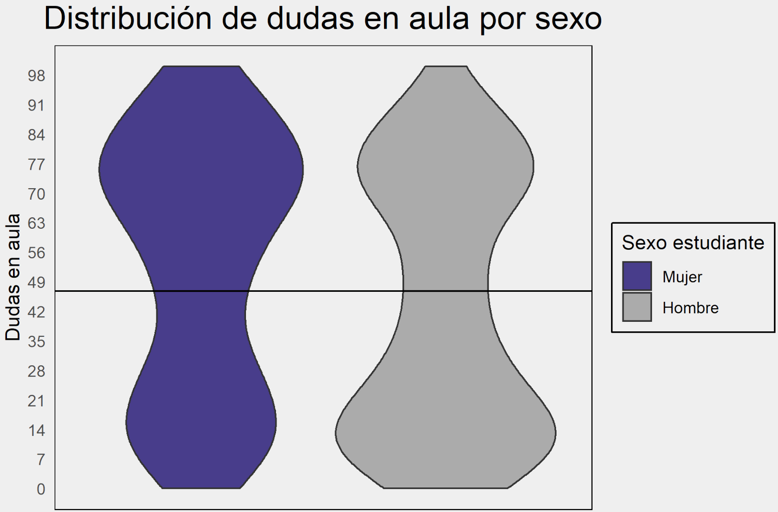
Intuiciones previas

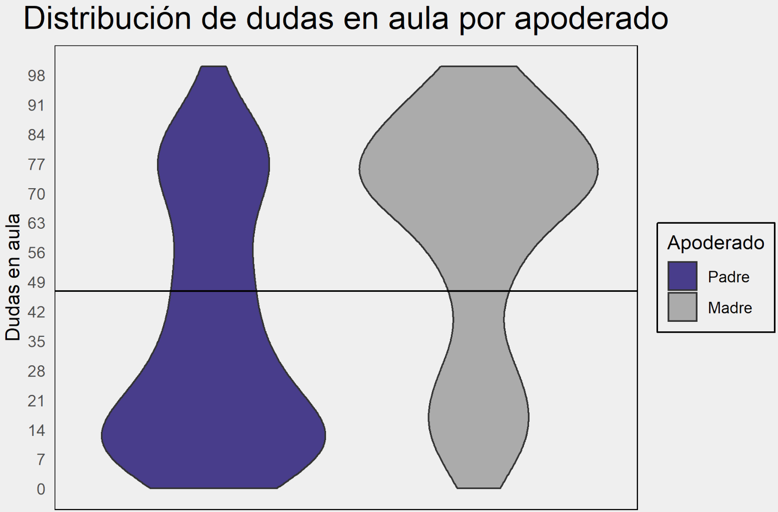
El promedio global de la cantidad de preguntas es 46.775.

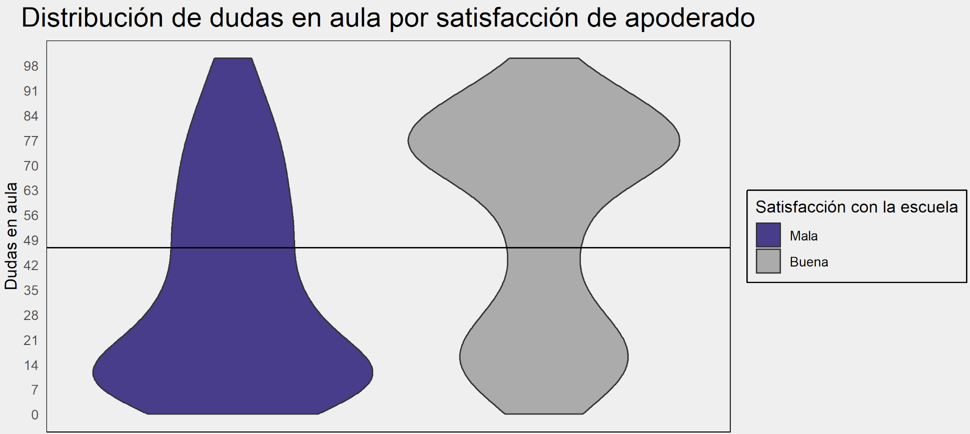
Género	n	Promedio n° preguntas
Femenino	175	52.9
Masculino	305	43.3

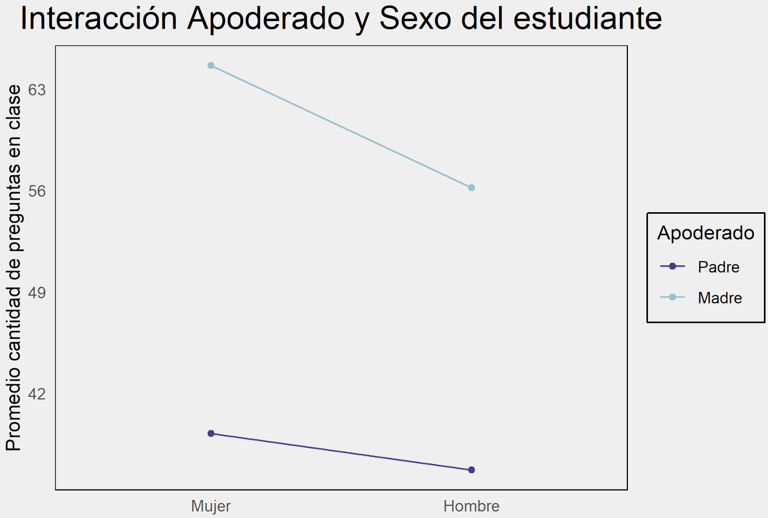
Apoderado	n	Promedio n° preguntas
Madre	283	60.2
Padre	197	37.4

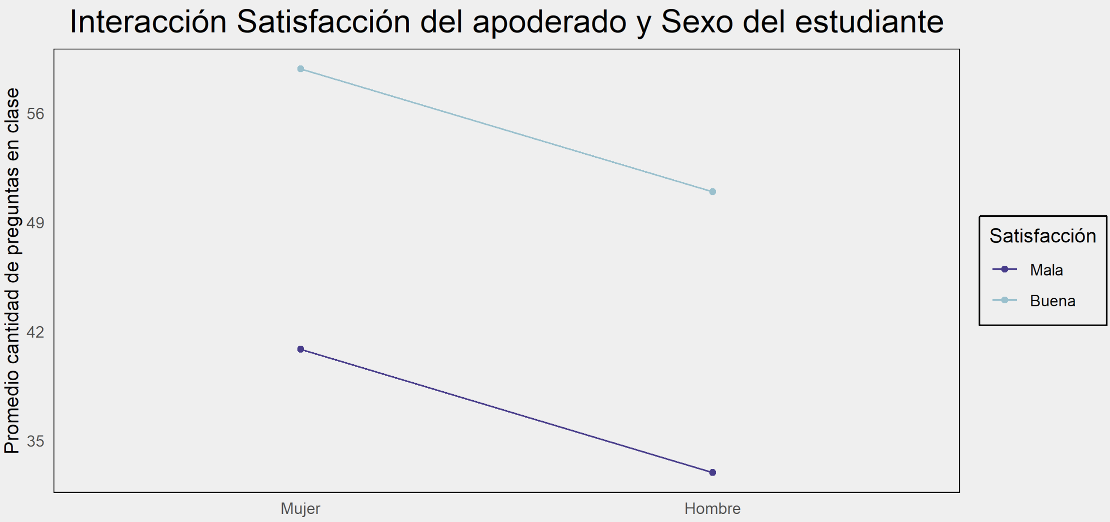
Satisfacción	n	Promedio n° preguntas
Buena	292	54.1
Mala	188	35.4

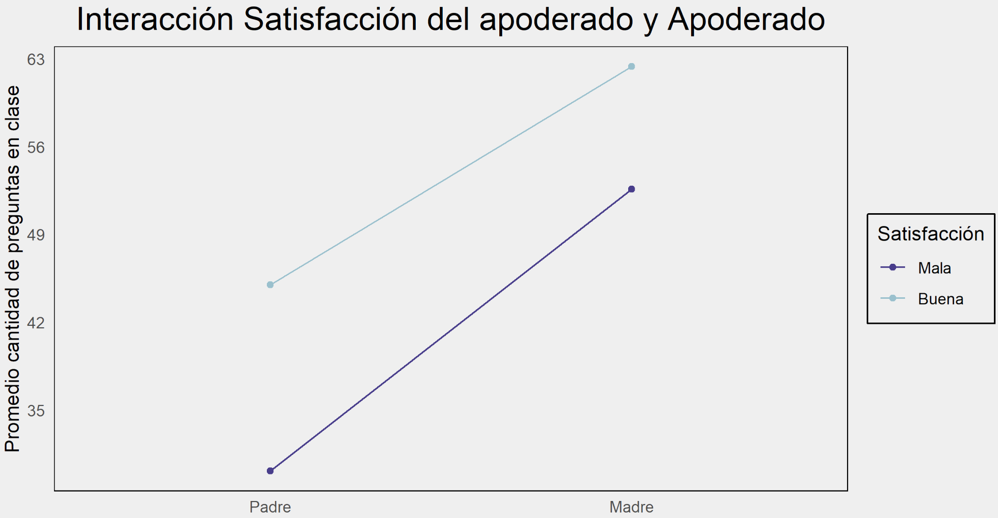

Distribución de dudas en aula por sexo
Dudas en aula
98
91
84
77
70
63
56
49
42
35
28
21
14
7
0
Sexo estudiante
Mujer
Hombre


Distribución de dudas en aula por apoderado
Dudas en aula
98
91
84
77
70
63
56
49
42
35
28
21
14
7
0
Apoderado
Padre
Madre


Distribución de dudas en aula por satisfacción de apoderado
Dudas en aula
98
91
84
77
70
63
56
49
42
35
28
21
14
7
0
Satisfacción con la escuela
Mala
Buena


Interacción Apoderado y Sexo del estudiante
Promedio cantidad de preguntas en clase
63
56
49
42
Mujer
Hombre
Apoderado
Padre
Madre


Interacción Satisfacción del apoderado y Sexo del estudiante
Promedio cantidad de preguntas en clase
56
49
42
35
Mujer
Hombre
Satisfacción
Mala
Buena


Interacción Satisfacción del apoderado y Apoderado
Promedio cantidad de preguntas en clase
63
56
49
42
35
Padre
Madre
Satisfacción
Mala
Buena

Un modelo anova de tres factores con interacciones de segundo orden es:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijkl}$$

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

i corresponde al nivel del factor sexo del alumno, con $i = 1, 2$.
j corresponde al nivel del factor apoderado del alumno, con $j = 1, 2$
k corresponde al nivel del factor satisfacción de apoderado, con $k = 1, 2$
l corresponde al índice que recorre la observación por combinación de factores, es decir $l = 1, ..., n_{ijk}$.

Y se utilizan los contrastes usuales:

$$\sum_{i=1}^{2} \alpha_i = 0$$

$$\sum_{j=1}^{2} \beta_i = 0$$

$$\sum_{l=1}^{2} \gamma_l = 0$$

Las interacciones por niveles i, j y k también suman cero.

```
#Las variables de tipo factor se definen como factores:

sexo<-factor(school$gender)
apoderado<-factor(school$parent)
satisfaccion<-factor(school$satisfaction)

participacion<-school$raisedhands

contrasts(sexo)<-contr.sum     #contraste suma para cada variable factor
contrasts(apoderado)<-contr.sum
contrasts(satisfaccion)<-contr.sum


#Modelo con interacciones dobles (pudieran incluirse triples)
model<-aov(participacion~(sexo+apoderado+satisfaccion)^2)
```

```
anova(model)
Analysis of Variance Table

Response: participacion
                       Df Sum Sq Mean Sq F value    Pr(>F)    
sexo                    1  10207   10207 13.0239 0.0003403 ***
apoderado               1  52933   52933 67.5390 1.999e-15 ***
satisfaccion            1  17731   17731 22.6234 2.621e-06 ***
sexo:apoderado          1    315     315  0.4017 0.5265353    
sexo:satisfaccion       1   1021    1021  1.3034 0.2541756    
apoderado:satisfaccion  1    873     873  1.1141 0.2917208    
Residuals             473 370706     784                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al realizar los tests F de significancia, es posible observar que ninguna interacción resulta significativa. Se debe ajustar un modelo aditivo.

El modelo aditivo, resulta como sigue:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl}$$

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

Con las restricciones usuales para efectos fijos:

$$\sum_{i=1}^{2} \alpha_i = 0 \qquad \sum_{j=1}^{2} \beta_i = 0$$

$$\sum_{l=1}^{2} \gamma_l = 0$$

```
aditivo<-aov(participacion~sexo+apoderado+satisfaccion)

anova(aditivo)   #Suma cuadrática secuencial SS(A), SS(B|A), SS(C|A,B)
Analysis of Variance Table

Response: participacion
              Df Sum Sq Mean Sq F value    Pr(>F)
sexo           1  10207   10207  13.029 0.0003392 ***
apoderado      1  52933   52933  67.565 1.951e-15 ***
satisfaccion   1  17731   17731  22.632 2.605e-06 ***
Residuals    476 372915     783
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La tabla anova usual o Tipo I cuantifica el aporte de cada variable dado que están las anteriores ya en el modelo. Se denomina tabla de sumas cuadráticas secuenciales. El orden de ingreso de las variables sí importa.

Tabla Anova tipo II o jerárquica

```
library(car)

Anova(aditivo, type="II") #Suma cuadrática jerárquica SS(A|B,C),SS(B|A,C),SS(C|A,B)
Anova Table (Type II tests)

Response: participacion
             Sum Sq  Df F value    Pr(>F)
sexo           2387   1  3.0468   0.08154 .
apoderado     34001   1 43.3999 1.184e-10 ***
satisfaccion  17731   1 22.6320 2.605e-06 ***
Residuals    372915 476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La tabla anova Tipo II cuantifica el aporte de cada variable dado que están las demás en el modelo. Se denomina tabla de sumas cuadráticas jerárquicas. El orden de ingreso de las variables no importa en esta tabla.

Los coeficientes obtenidos en el modelo son:

Coeficiente	Efecto
Intercepto	47.6
Sexo Femenino	2.3
Sexo Masculino	-2.3
Apoderado padre	-9.0
Apoderado madre	9.0
Satisfacción mala	-6.5
Satisfacción buena	6.5

$$\underbrace{\sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y})^2}_{SCT}$$

$$\underbrace{\sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y_i})^2}_{SCsexo} \quad \underbrace{\sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y_j})^2}_{SCapoderado}$$

$$\underbrace{\sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y_k})^2}_{SCsatisfaccion} \quad \underbrace{\sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y_{ijk}})^2}_{SCE}$$

$$SCT = SCSexo + SCapoderado + SCsatisfaccion + SCE$$

```
anova(aditivo)[,1:3]
               Df Sum Sq Mean Sq
sexo            1  10207   10207
apoderado       1  52933   52933
satisfaccion    1  17731   17731
Residuals     476 372915     783
```

Modelo anova efectos aleatorios, factores *within*

Asumimos efectos fijos cuando creemos que no existe ninguna fuente de variabilidad adicional a la variabilidad ya asumida entre las unidades experimentales.

¿A qué se refiere con *fuente de variabilidad adicional*?

1. Alguna característica (no medida) intra grupo de la que se sopecha, pudiera afectar los resultados del estudio (por ejemplo, factores sociales, factores contextuales, factores culturales, factores demográficos), es decir, presencia de influencia(s) que aumenten la variabilidad de los resultados
2. Diferencias en la selección de unidades intra grupo (distintos métodos de muestreo por nivel de los factores)
3. Niveles del factor corresponden a una selección pequeña respecto del macro de posibilidades (elegir representantes aleatorios para los factores, países, ciudades, para tratar de inferir sobre continentes, regiones)

Considere que el factor A tiene a niveles, el factor B tiene b niveles, luego, $i = 1, ..., a$, $j = 1, ..., b$ y $k = 1, ..., n_{ij}$ donde n_{ij} corresponde a la cantidad de observaciones en la combinación (i, j) de factores. Luego, el modelo es el siguiente:

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

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

Con los contrastes usuales:

$$\sum_{i=1}^{2} \alpha_i = 0$$

$$\sum_{j=1}^{2} \beta_i = 0$$

$$\sum_{i=1}^{2} (\alpha\beta)_{ij} = 0 \quad \sum_{j=1}^{2} (\alpha\beta)_{ij} = 0$$

Cuando tenemos efectos fijos, resulta razonable testear si los efectos son nulos o no:

Factor A:

$$H_0 : \alpha_i = 0 \ \ \forall\, i \qquad H_1 : \exists \ \alpha_i \neq 0$$

$$F_A = \frac{MCA}{MCE}$$

Factor B:

$$H_0 : \beta_j = 0 \ \ \forall\, j \qquad H_1 : \exists \ \beta_j \neq 0$$

$$F_B = \frac{MCB}{MCE}$$

Factor interacción:

$$H_0 : (\alpha\beta)_{ij} = 0 \ \ \forall\, (i,j) \qquad H_1 : \exists \ (\alpha\beta)_{ij} \neq 0$$

$$F_{AB} \frac{MCAB}{MCE}$$

En R, la tabla Anova tipo I entrega los valores-p asociados a estos tests.

El modelo es el siguiente:

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

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

Y dada la naturaleza aleatoria de los factores, se tiene que:

$$\alpha_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\alpha)$$

$$\beta_j \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\beta)$$

$$(\alpha\beta)_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\alpha\beta})$$

Cuando tenemos efectos aleatorios, es necesario incorporar la definición del efecto aleatorio: que añade variabilidad. Así, lo que resulta natural testear es lo siguiente:

Factor A:

$$H_0 : \sigma^2_\alpha = 0 \quad H_1 : \sigma^2_\alpha > 0$$

$$F_A = \frac{MCA}{MCAB}$$

Factor B:

$$H_0 : \sigma^2_\beta = 0 \quad H_1 : \sigma^2_\beta > 0$$

$$F_B = \frac{MCB}{MCAB}$$

Factor interacción:

$$H_0 : \sigma^2_{\alpha\beta} = 0 \quad H_1 : \sigma^2_{\alpha\beta} > 0$$

$$F_{AB} = \frac{MCAB}{MCE}$$

En un modelo de efectos aleatorios, los efectos estimados $\hat{\alpha}_i$ y $\hat{\beta}_j$ y $(\widehat{\alpha\beta})_{ij}$ no son de interés, pues estos coeficientes son aleatorios y por ende, en cada realización del experimento se obtendrán distintas estimaciones. Lo que resulta de interés estudiar son las fuentes de variabilidad asociadas a cada componente.

Caso aplicado efectos aleatorios

La base de dados `suicide` contiene información sobre la cantidad de suicidios registrados en el año 2015. Los investigadores buscan determinar si por continente se observan diferencias en la cantidad de suicidios. Para ello, por continente se eligieron 4 países aleatorios como representantes. Además, otro tema de interés es determinar si existen diferencias en la cantidad de suicidios por edad, considere que para cada rango etario se tomó una muestra aleatoria (del mismo tamaño) de los fallecidos el 2015 y se realizó un conteo de quiénes tuvieron como causa de muerte asociada al sucidio.

El factor A corresponde a continente, este factor es aleatorio pues para cada continente se seleccionan aleatoriamente 4 países representantes, por lo tanto, las mediciones se verán vastamente influenciadas por estos 4 países elegidos.

El factor B corresponde a rango etario, este factor es aleatorio pues **dentro** de cada grupo se realiza un muestreo aleatorio (de tamaño no especificado) de los fallecidos. Es decir, en cada realización de este experimento las mediciones se verán influenciadas por las elecciones aleatorias de los sujetos dentro de cada rango etario.

Por lo tanto, planteamos el modelo de efectos aleatorios:

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

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

Con

$$\alpha_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\alpha)$$

$$\beta_j \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\beta)$$

$$(\alpha\beta)_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\alpha\beta})$$

```
Continente<-factor(suicide$Continente)
Etario<-factor(suicide$Etario)
Suicidios<-suicide$Suicidios

modelo<-aov(Suicidios~Continente*Etario)

anova(modelo)
Analysis of Variance Table

Response: Suicidios
                  Df    Sum Sq  Mean Sq F value  Pr(>F)
Continente         3  58605347 19535116  3.8974 0.01223 *
Etario             5  26703880  5340776  1.0655 0.38668
Continente:Etario 15  25789409  1719294  0.3430 0.98808
Residuals         72 360888350  5012338
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

No olvidar: En un modelo de efectos aleatorios con interacción, los estadísticos son diferentes a un caso fijo.

```
anova(modelo)
Analysis of Variance Table

Response: Suicidios
                  Df    Sum Sq  Mean Sq F value   Pr(>F)
Continente         3  58605347 19535116  3.8974 0.01723 *
Etario             5  26703880  5340776  1.0655 0.38668
Continente:Etario 15  25789409  1719294  0.3430 0.98808
Residuals         72 360888350  5012338
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Factor A: Continente

$$H_0 : \sigma_\alpha^2 = 0 \quad H_1 : \sigma_\alpha^2 > 0$$

$$F_A = \frac{MCA}{MCAB} = \frac{19535116}{1719294} = 11.362 \quad \text{F teórico} = 3.28$$

Factor B: Rango etario

$$H_0 : \sigma_\beta^2 = 0 \quad H_1 : \sigma_\beta^2 > 0$$

$$F_B = \frac{MCB}{MCAB} = \frac{5340776}{1719294} = 3.1 \quad \text{F teórico} = 2.9$$

Factor interacción entre continente y rango etario:

$$H_0 : \sigma_{\alpha\beta}^2 = 0 \quad H_1 : \sigma_{\alpha\beta}^2 > 0$$

$$F_{AB} = \frac{MCAB}{MCE} = \frac{1719294}{5012338} = 0.3430124 \quad \text{F teórico} = 3.28$$

Es posible estimar cada una de las fuentes de variabilidad de nuestro modelo:

$$\hat{\sigma}^2 = MCE = 5012338$$

$$\hat{\sigma}^2_\alpha = \frac{MCA - MCAB}{nb} = \frac{19535116 - 1719294}{4 \cdot 6} = 742325.9$$

$$\hat{\sigma}^2_\beta = \frac{MCB - MCAB}{na} = \frac{5340776 - 1719294}{4 \cdot 4} = 3621482$$

$$\hat{\sigma}^2_{\alpha\beta} = \frac{MCAB - MCE}{n} = \frac{1719294 - 5012338}{4} = -823261$$

Si una de las variabilidades resulta ser negativa, significa que el modelo considerado no es el adecuado. En este caso, la estimación de la variabilidad asociada a la interacción resulta ser negativa, es decir, la componente interacción de los factores continente y rango etario posee un aporte minúsculo en términos de variabilidad.

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

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

Con

$$\alpha_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\alpha)$$

$$\beta_j \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\beta)$$

```
aditivo<-aov(Suicidios~Continente+Etario)

anova(aditivo)
Analysis of Variance Table

Response: Suicidios
           Df    Sum Sq  Mean Sq F value   Pr(>F)
Continente  3  58605347 19535116  4.3953 0.006285 **
Etario      5  26703880  5340776  1.2016 0.315197
Residuals  87 386677759  4444572
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\hat{\sigma^2} = MCE = 4444572$$

$$\hat{\sigma}_\alpha^2 = \frac{MCA - MCE}{nb} = \frac{19535116 - 4444572}{4 \cdot 6} = 628772.7$$

$$\hat{\sigma}_\beta^2 = \frac{MCB - MCE}{na} = \frac{5340776 - 4444572}{4 \cdot 4} = 56012.755$$

Luego, la variabilidad total incorporada en el experimento es:

$$\hat{\sigma^2} + \hat{\sigma}_\alpha^2 + \hat{\sigma}_\beta^2 = 4444572 + 628772.7 + 56012.755 = 5129357$$

Y puede realizarse una tabla como sigue:

Fuente	Variabilidad	Porcentaje
Factor Continente	628772.7	12%
Factor Rango etario	56012.755	1.09%
Residuos	4444572	86.6%

Factor rango etario no resulta ser significativa dada la presencia de Continente, esto va estrechamente legado con que resulta ser una variable con tan poco porcentaje de variabilidad explicada.

```
library(lme4)
m1<-lmer(Suicidios~(1|Continente)+(1|Etario))

summary(m1)
Linear mixed model fit by REML ['lmerMod']
Formula: Suicidios ~ (1 | Continente) + (1 | Etario)

REML criterion at convergence: 1733.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.9265 -0.2398 -0.1309 -0.0442  6.5144

Random effects:
 Groups     Name        Variance Std.Dev.
 Etario     (Intercept)   56006   236.7
 Continente (Intercept)  628755   792.9
 Residual               4444580  2108.2
Number of obs: 96, groups:  Etario, 6; Continente, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)    803.7      461.3   1.742
```

Caso aplicado efectos mixtos

La base de datos performance contiene mediciones sobre las puntuaciones en un test de educación física de distintos participantes. Se busca evaluar el efecto del sexo y del nivel de estrés del participante al realizar el test. Además, cada individuo repitió el test continuamente dos veces, al inicio t_1 y al final t_2, por lo cual resulta relevante analizar si los resultados de las puntuaciones varían por tiempo/realización del test. Plantee un modelo adecuado que permita analizar este caso.

Una manera de distinguir si un factor es de efectos fijos o aleatorios es realizándose la siguiente pregunta:

¿Tiene sentido estudiar el efecto del factor en sí?
¿o tiene más sentido estudiar la variabilidad asociada a dicho factor?

Por ejemplo, para el factor género, tiene sentido estudiar el efecto de cada género por sobre los resultados de las puntuaciones. También suena razonable estudiar el efecto de cada nivel de estrés por sobre los resultados de las puntuaciones, pero, no tiene mucho sentido estudiar el *efecto* del tiempo en las puntuaciones del test. Más interesante sería analizar la *variabilidad* de las puntuaciones en el tiempo.

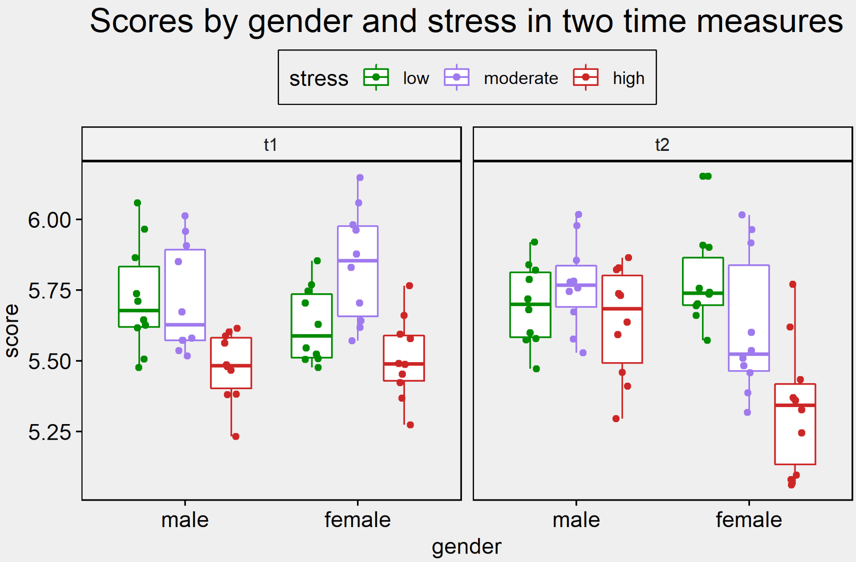
Note que el género y el estrés es una característica que varía *entre* participantes.
Mientras que el tiempo es una característica que varía *intra* participantes. Es decir, existiría variabilidad asociada de la performance de los participantes por tiempo.

Si yo me paro en el grupo de género femenino, espero obtener similitudes en las observaciones, no esperaría mayor variabilidad.
Si yo me paro en el grupo de estrés bajo, espero obtener similitudes en las observaciones, no esperaría mayor variabilidad.
Si yo me paro en en el grupo de observaciones del tiempo t_2 es decir, la realización final del test, podría esperar mayor variabilidad, pues el cambio al t_2 variará intrasujeto.

```
library(datarium)
data("performance", package = "datarium")  #Carga la base de datos performance

performance  #información agrupada por id
# A tibble: 60 x 5
      id gender stress    t1    t2
   <int> <fct>  <fct>  <dbl> <dbl>
 1     1 male   low     5.96  5.58
 2     2 male   low     5.51  5.82
 3     3 male   low     5.63  5.47
 4     4 male   low     5.71  5.79
 5     5 male   low     5.74  5.72
 6     6 male   low     5.62  5.68

library(tidyverse)
library(ggpubr)
library(rstatix)

performance <- performance %>%
  gather(key = "time", value = "score", t1, t2) %>%
  convert_as_factor(id)

performance #información a lo largo, extendida
# A tibble: 120 x 5
   id    gender stress time  score
   <fct> <fct>  <fct>  <fct> <dbl>
 1 1     male   low    t1     5.96
 2 2     male   low    t1     5.51
 3 3     male   low    t1     5.63
 4 4     male   low    t1     5.71
 5 5     male   low    t1     5.74
 6 6     male   low    t1     5.62
```

```
performance%>%
  group_by(gender,stress,time)%>%
  summarise(n=n())  #Caso balanceado
# A tibble: 12 x 4
# Groups:   gender, stress [6]
   gender stress   time      n
   <fct>  <fct>    <fct> <int>
 1 male   low      t1       10
 2 male   low      t2       10
 3 male   moderate t1       10
 4 male   moderate t2       10
 5 male   high     t1       10
 6 male   high     t2       10
 7 female low      t1       10
 8 female low      t2       10
```

Nos encontramos en un caso balanceado. Esto es relevante pues, en estos casos, las tablas anova tipo I y tipo II coinciden.


Scores by gender and stress in two time measures
stress
low
moderate
high
t1
t2
6.00
5.75
5.50
5.25
score
male
female
male
female
gender

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

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

Con las restricciones usuales para efectos fijos:

$$\sum_{i=1}^{2} \alpha_i = 0 \qquad \sum_{j=1}^{3} \beta_j = 0$$

$$\sum_{i} (\alpha\beta)_{ij} = \sum_{j} (\alpha\beta)_{ij} = 0$$

Y considerando el efecto aleatorio de γ:

$$(\alpha\beta\gamma)_{ijk} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\alpha\beta\gamma})$$

$$(\alpha\gamma)_{ik} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\alpha\gamma})$$

$$(\beta\gamma)_{jk} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\beta\gamma})$$

$$\gamma_k \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\gamma})$$

Interesa realizar test F de las siguientes hipótesis:

Género

$$H_0 : \alpha_i = 0 \ \ \forall\, i \qquad H_1 : \exists \ \alpha_i \neq 0$$

Estrés

$$H_0 : \beta_j = 0 \ \ \forall\, j \qquad H_1 : \exists \ \beta_j \neq 0$$

Tiempos

$$H_0 : \sigma^2_{\gamma} = 0 \qquad H_1 : \sigma^2_{\gamma} > 0$$

Interacción sexo, estrés y tiempos

$$H_0 : \sigma^2_{\alpha\beta\gamma} = 0 \qquad H_1 : \sigma^2_{\alpha\beta\gamma} > 0$$

Interacción sexo y tiempos

$$H_0 : \sigma^2_{\alpha\gamma} = 0 \qquad H_1 : \sigma^2_{\alpha\gamma} > 0$$

Interacción estrés y tiempos

$$H_0 : \sigma^2_{\beta\gamma} = 0 \qquad H_1 : \sigma^2_{\beta\gamma} > 0$$

Interacción sexo y estrés

$$H_0 : (\alpha\beta)_{ij}) \ \forall\, (i, j) \qquad H_1 : \exists \ (\alpha\beta)_{ij} \neq 0$$

Test F en R

```
performance$gender<-factor(performance$gender)
performance$stress<-factor(performance$stress)
performance$time<-factor(performance$time)

levels(performance$gender)
[1] "male"   "female"
levels(performance$stress)
[1] "low"      "moderate" "high"
levels(performance$time)
[1] "t1" "t2"

contrasts(performance$gender)<-contr.sum
contrasts(performance$stress)<-contr.sum

Type1<- anova_test(data = performance, #base de datos
                   dv = score,  #variable dependiente
                   between = c(gender, stress), #variables efecto fijo
                   wid = id, #id del individuo
                   within = time, type=1) #variable efecto aleatorio

get_anova_table(Type1)
ANOVA Table (type I tests)

              Effect DFn DFd      F        p p<.05      ges
1             gender   1  54  2.406 1.27e-01       0.023000
2             stress   2  54 21.166 1.63e-07     * 0.288000
3               time   1  54  0.063 8.03e-01       0.000564
4      gender:stress   2  54  1.554 2.21e-01       0.029000
5        gender:time   1  54  4.730 3.40e-02     * 0.041000
6        stress:time   2  54  1.821 1.72e-01       0.032000
7 gender:stress:time   2  54  6.101 4.00e-03     * 0.098000
```

¿Para cada tiempo/realización del test se tiene la misma significancia de los factores sexo y estrés?

```
performance %>%
  group_by(time) %>%
  anova_test(dv = score, wid = id, between = c(gender, stress), type=1)
  time  Effect          DFn   DFd      F          p `p<.05`   ges
  <fct> <chr>         <dbl> <dbl>  <dbl>      <dbl> <chr>   <dbl>
1 t1    gender            1    54  0.186 0.668      ""      0.003
2 t1    stress            2    54 14.9   0.00000723 "*"     0.355
3 t1    gender:stress     2    54  2.12  0.131      ""      0.073
4 t2    gender            1    54  5.97  0.018      "*"     0.1
5 t2    stress            2    54  9.60  0.000271   "*"     0.262
6 t2    gender:stress     2    54  4.95  0.011      "*"     0.155
```

En la segunda realización del test, los factores adquieren mayor significancia, en particular hay diferencias considerables para `gender`.

Test de comparaciones múltiples

```
performance %>%
  group_by(time, gender) %>%
  pairwise_t_test(score ~ stress, p.adjust.method = "bonferroni") %>%
  select(-p, -p.signif)
  # A tibble: 12 x 9
   gender time  .y.   group1   group2      n1    n2    p.adj p.adj.signif
 * <fct>  <fct> <chr> <chr>    <chr>    <int> <int>    <dbl> <chr>
 1 male   t1    score low      moderate    10    10 1        ns
 2 male   t1    score low      high        10    10 0.012    *
 3 male   t1    score moderate high        10    10 0.0131   *
 4 female t1    score low      moderate    10    10 0.0196   *
 5 female t1    score low      high        10    10 0.357    ns
 6 female t1    score moderate high        10    10 0.000301 ***
 7 male   t2    score low      moderate    10    10 1        ns
 8 male   t2    score low      high        10    10 1        ns
 9 male   t2    score moderate high        10    10 0.265    ns
10 female t2    score low      moderate    10    10 0.323    ns
11 female t2    score low      high        10    10 0.000318 ***
12 female t2    score moderate high        10    10 0.0235   *
```

Supuestos

En un modelo anova se deben estudiar las observaciones outliers dentro de cada combinación de factores:

```
performance %>%
  group_by(gender, stress, time) %>%
  identify_outliers(score)
# A tibble: 1 x 7
  gender stress time  id    score is.outlier is.extreme
  <fct>  <fct>  <fct> <fct> <dbl> <lgl>      <lgl>
1 female low    t2    36     6.15 TRUE       FALSE
```

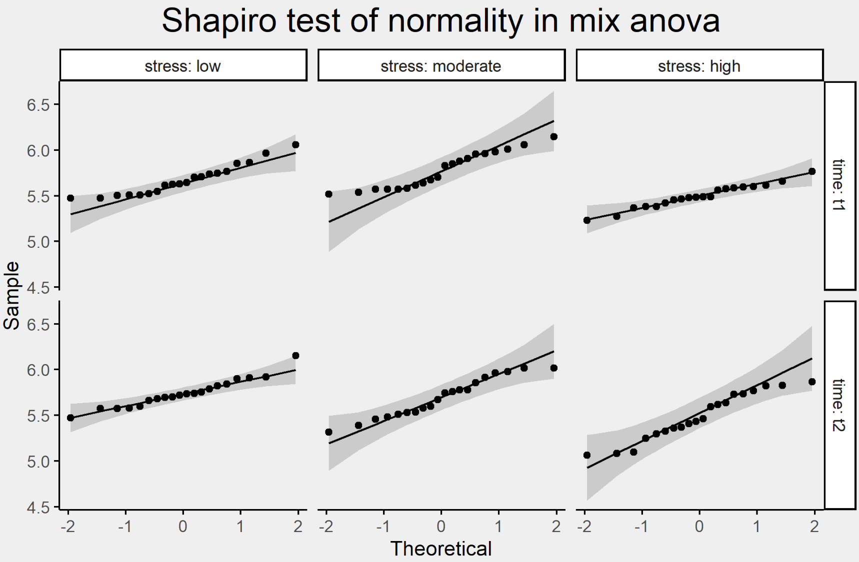
Sólo se detecta un outlier asociado al id 36, no es un outlier extremo.

Normalidad

La hipótesis nula del test de Shapiro es la normalidad.

```
performance %>%
  group_by(gender, stress, time ) %>%
  shapiro_test(score)
# A tibble: 12 x 6
   gender stress   time  variable statistic      p
   <fct>  <fct>    <fct> <chr>        <dbl>  <dbl>
 1 male   low      t1    score        0.942 0.574
 2 male   low      t2    score        0.966 0.849
 3 male   moderate t1    score        0.848 0.0547
 4 male   moderate t2    score        0.958 0.761
 5 male   high     t1    score        0.915 0.319
 6 male   high     t2    score        0.925 0.403
 7 female low      t1    score        0.898 0.207
 8 female low      t2    score        0.886 0.154
 9 female moderate t1    score        0.946 0.626
10 female moderate t2    score        0.865 0.0880
11 female high     t1    score        0.989 0.996
12 female high     t2    score        0.930 0.452
```

Para ninguna de las combinaciones de factores se rechaza el supuesto de normalidad.


Shapiro test of normality in mix anova
stress: low
stress: moderate
stress: high
time: t1
time: t2
Sample
Theoretical
6.5
6.0
5.5
5.0
4.5
-2
-1
0
1
2

```
performance %>%
  group_by(time) %>%
  levene_test(score~gender*stress)  #dentro de cada combinación de factores fijos
    time    df1   df2 statistic     p
  <fct> <int> <int>     <dbl> <dbl>
1 t1        5    54     0.974 0.442
2 t2        5    54     0.722 0.610
```

Parámetros estimados de efectos fijos

```
completo<-lmer(score~gender*stress*(1|time), data=performance)

fixef(completo)   #Estimaciones de componentes fijas
    (Intercept)          gender1          stress1          stress2
     5.64425612       0.02607720       0.06227302       0.09150249
gender1:stress1 gender1:stress2
    -0.02315229      -0.01868816

fitted(completo)  #Valores ajustados

residuals(completo)  #residuos

summary(residuals(completo))  #estadísticas de los errores
      Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
-0.4114546 -0.1512498  0.0002265  0.0000000  0.1506693  0.4474113
```

```
fijos<-aov(score~gender*stress, data=performance)

anova(fijos)
Analysis of Variance Table

Response: score
               Df Sum Sq Mean Sq F value    Pr(>F)
gender          1 0.0816 0.08160  2.2157    0.1394
stress          2 1.4359 0.71795 19.4943 5.224e-08 ***
gender:stress   2 0.1054 0.05272  1.4314    0.2432
Residuals     114 4.1985 0.03683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

coef(fijos)
    (Intercept)         gender1         stress1         stress2
     5.64425612      0.02607720      0.06227302      0.09150249
gender1:stress1 gender1:stress2
    -0.02315229     -0.01868816

summary(residuals(fijos))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.4114546 -0.1512498  0.0002265  0.0000000  0.1506693  0.4474113
```

Ancova

¿Por qué ancova?

Pasamos de un modelo anova a un modelo ancova, cuando incorporamos variables cuantitativas:

i. Muchas veces, las variables de tipo factor no logran explicar el porcentaje de variabilidad deseado
ii. Si existe información correlacionada con la variable respuesta, ¿por qué no incorporarla?

Caso aplicado modelo ANCOVA

La base de datos breast contiene información de células de masas mamarias. La información se obtuvo a través de aspirado de células con aguja fina (FNA) de la masa mamaria. Los registros describen las características de los núcleos celulares presentes en la imagen.

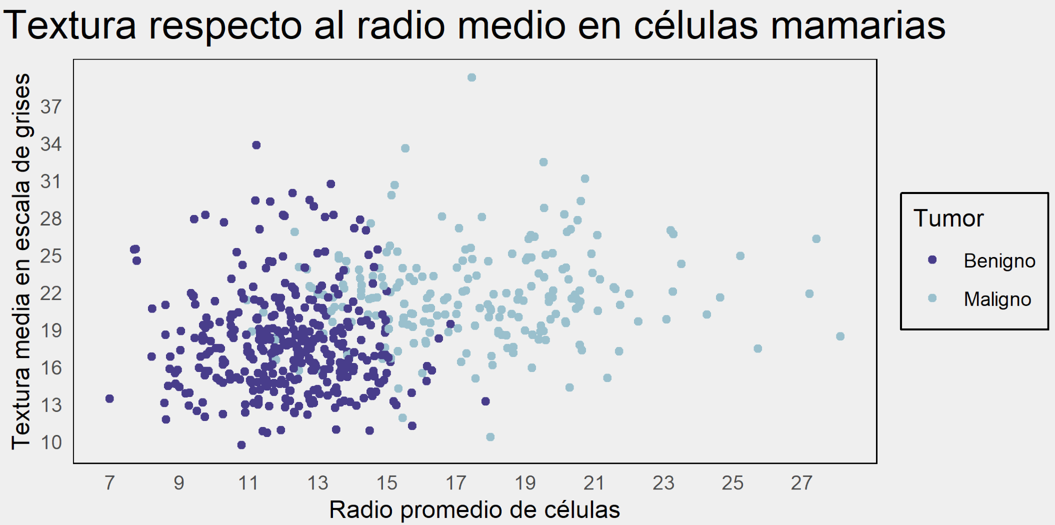
Interesa estudiar la textura media (valor en escala de grises) y determinar si existiría algún efecto en la naturaleza del tumor (benigno o maligno) en la textura del tumor. Además, se cuenta con información de disversas variables cuantitativas en la base de datos, determine cuál incluiría y plantee el modelo correspondiente.

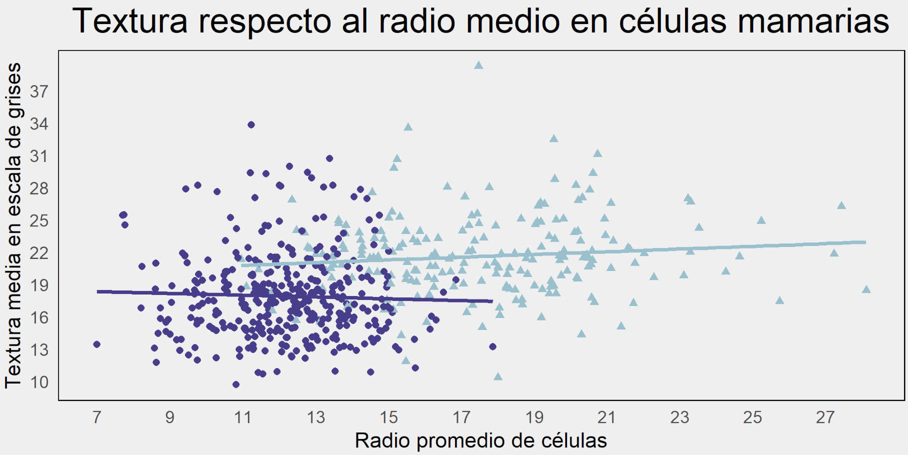
```
dim(breast)  #569 registros y 33 columnas
[1] 569  33

names(breast) #primera columna es id
[1] "id"                          "diagnosis"                  "radius_mean"
 [4] "texture_mean"                "perimeter_mean"             "area_mean"
 [7] "smoothness_mean"             "compactness_mean"           "concavity_mean"
[10] "concave points_mean"         "symmetry_mean"              "fractal_dimension_mean"
[13] "radius_se"                   "texture_se"                 "perimeter_se"
[16] "area_se"                     "smoothness_se"              "compactness_se"
[19] "concavity_se"                "concave points_se"          "symmetry_se"
[22] "fractal_dimension_se"        "radius_worst"               "texture_worst"
[25] "perimeter_worst"             "area_worst"                 "smoothness_worst"
[28] "compactness_worst"           "concavity_worst"            "concave points_worst"
[31] "symmetry_worst"              "fractal_dimension_worst" "X33"

table(table(breast$id))  #no existen registros con mismo id

  1
569

table(breast$diagnosis)  #357 tumores benignos, 212 tumores malignos

  B   M
357 212
```

Un supuesto del modelo ancova es que las covariables deben estar correlacionadas con la variable respuesta.
¿Cuáles son las variables más correlacionadas con la variable textura media?

```
cor(breast[,3:32])[order(abs(cor(breast[,3:32])[,2]), decreasing=TRUE),2]
           texture_mean           texture_worst              texture_se
            1.000000000             0.912044589             0.386357623
        perimeter_worst            radius_worst              area_worst
            0.358039575             0.352572947             0.343545947
         perimeter_mean             radius_mean               area_mean
            0.329533059             0.323781891             0.321085696
         concavity_mean         concavity_worst    concave points_worst
            0.302417828             0.301025224             0.295315843
    concave points_mean            perimeter_se       compactness_worst
            0.293464051             0.281673115             0.277829592
              radius_se                 area_se        compactness_mean
            0.275868676             0.259844987             0.236702222
         compactness_se       concave points_se            concavity_se
            0.191974611             0.163851025             0.143293077
fractal_dimension_worst          symmetry_worst        smoothness_worst
            0.119205351             0.105007910             0.077503359
 fractal_dimension_mean           symmetry_mean    fractal_dimension_se
           -0.076437183             0.071400980             0.054457520
        smoothness_mean             symmetry_se           smoothness_se
           -0.023388516             0.009127168             0.006613777
```

Utilizaremos `radius_mean` como covariable.


Textura respecto al radio medio en células mamarias
Textura media en escala de grises
37
34
31
28
25
22
19
16
13
10
7
9
11
13
15
17
19
21
23
25
27
Radio promedio de células
Tumor
Benigno
Maligno



La idea es determinar la recta que pase para cada una de las dos subpoblaciones: tumor benigno y tumor maligno.

Un modelo que podría plantearse es el siguiente:

$$Y_{ij} = \mu + \tau_i + \gamma X_{ij} + \epsilon_{ij}$$

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

$$\sum_{i=1}^{r} \tau_i = 0$$

El problema de este modelo:

$$\bar{Y}_{..} = \mu + \bar{\gamma}\bar{X}_{..} + \bar{\epsilon}_{..} \iff \mu = \bar{Y}_{..} - \bar{\gamma}\bar{X}_{..} + \bar{\epsilon}_{..}$$

μ depende de $\bar{X}$. Una manera de arreglar esto es incorporar la información de X centrado en 0 (es decir, restándole su media):

$$Y_{ij} = \mu + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}$$

De este modo, $\mu \approx \bar{Y}_{..}$

Modelo en R

```
breast$diagnosis<-factor(breast$diagnosis)
contrasts(breast$diagnosis)<-contr.sum

radius_meancent<-breast$radius_mean-mean(breast$radius_mean)
model<-lm(texture_mean~diagnosis*radius_meancent, data=breast)
summary(model)  #interaccion no significativa
Call:
lm(formula = texture_mean ~ diagnosis * radius_meancent, data = breast)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2911  -2.5845  -0.4656   1.9250  17.6754

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                19.46772    0.24854  78.330  < 2e-16 ***
diagnosis1                 -1.71807    0.24854  -6.913 1.29e-11 ***
radius_meancent             0.02115    0.07184   0.294    0.769
diagnosis1:radius_meancent -0.10450    0.07184  -1.455    0.146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.914 on 565 degrees of freedom
Multiple R-squared:  0.1764,Adjusted R-squared:  0.172
F-statistic: 40.33 on 3 and 565 DF,  p-value: < 2.2e-16
```

```
aditivo<-lm(texture_mean~diagnosis+radius_meancent, data=breast)

summary(aditivo)
Call:
lm(formula = texture_mean ~ diagnosis + radius_meancent, data = breast)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2534  -2.6290  -0.3982   1.9583  17.6752

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     19.72322    0.17602 112.051  < 2e-16 ***
diagnosis1      -1.70139    0.24852  -6.846 1.98e-11 ***
radius_meancent  0.05405    0.06825   0.792    0.429
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.918 on 566 degrees of freedom
Multiple R-squared:  0.1733,Adjusted R-squared:  0.1704
F-statistic: 59.32 on 2 and 566 DF,  p-value: < 2.2e-16
```

i. El intercepto obtenido, es decir $\hat{\mu}$ es 19.72, es un valor bastante similar a la media de la variable `texture_mean` (19.28).

ii. `diagnosis1` corresponde al efecto asociado al diagnóstico benigno del tumor por sobre la media de la textura en escala de grises. Notar que $\mu + \tau_i$ corresponde al intercepto de la recta asociada al grupo i. Por lo cual:

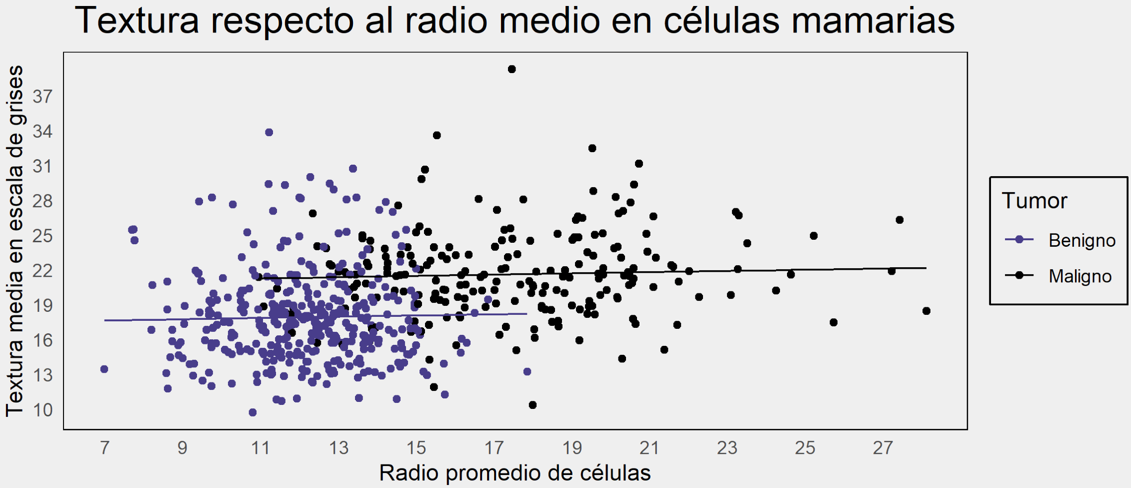
$$\text{El intercepto para tumores benignos es } \hat{\mu} + \hat{\tau}_1 = 19.72 - 1.7 = 18.02$$

$$\text{El intercepto para tumores malignos es } \hat{\mu} - \hat{\tau}_1 = 19.72 + 1.7 = 21.42$$

iii. `radius_meancent` corresponde a la pendiente asociada a la variable radio medio centrada. Es decir, las rectas para ambas poblaciones son:

$$\text{Recta para tumores benignos es } 18.02 + 0.054(X - \bar{X})$$

$$\text{Recta para tumores malignos es } 21.42 + 0.054(X - \bar{X})$$



Textura respecto al radio medio en células mamarias
Textura media en escala de grises
37
34
31
28
25
22
19
16
13
10
7
9
11
13
15
17
19
21
23
25
27
Radio promedio de células
Tumor
Benigno
Maligno

Análisis del modelo

```
#### ¿Pendientes iguales?
confint(lm(texture_mean~radius_mean,data=subset(breast, diagnosis=="M")))
                  2.5 %     97.5 %
(Intercept) 16.57812935 22.2433096
radius_mean -0.03390703  0.2852049

confint(lm(texture_mean~radius_mean, data=subset(breast, diagnosis=="B")))
                 2.5 %     97.5 %
(Intercept) 16.0541097 21.8004114
radius_mean -0.3174039  0.1506897
```

Los intervalos de confianza de las pendientes, se solapan. Podrían asumirse pendientes iguales.

```
summary(aditivo)$adj.r.squared
[1] 0.1703737

anova(modeldiag,aditivo)
Analysis of Variance Table

Model 1: texture_mean ~ diagnosis
Model 2: texture_mean ~ diagnosis + radius_meancent
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    567 8696.1
2    566 8686.5  1    9.6255 0.6272 0.4287

summary(aditivo)$coefficients
                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)     19.7232198 0.17601952 112.051322 0.000000e+00
diagnosis1      -1.7013936 0.24851513  -6.846238 1.978756e-11
radius_meancent  0.0540519 0.06825157   0.791951 4.287210e-01
```

Incorporar la variable radius_mean no implica mejoras considerables dado que ya se encuentra diagnosis en el modelo.