Análisis de la Varianza de dos factores con replicaciones: Caso Balanceado (Scheffé, 1959)

En este ejemplo nos interesa el tiempo de coagulación (en minutos) del plasma sanguíneo para 3 tratamientos y 2 concentraciones de adrenalina mezclada con el plasma.
Para cada combinación de tratamiento y concentración de adrenalina, se tomaron 3 observaciones independientes. Se obtuvieron los siguientes datos:

	Concentración	
Tratamiento	**1**	**2**
1	9.8 10.1 9.8	11.3 10.7 10.7
2	9.2 8.6 9.2	10.3 10.7 10.2
3	8.4 7.9 8.0	9.8 10.1 10.1

En este caso tenemos dos factores:

- **Factor A**: Tratamiento (con tres niveles)
- **Factor B**: Concentración (dos niveles)

y dentro de cada casillero tenemos la misma **cantidad de replicaciones K**, en este caso K=3.

Podemos pensar que nuestros datos se disponen en **una tabla de doble** entrada como la anterior (una entrada para el factor A y otra para B) y en la que en **cada casilla** tendremos las **replicaciones** de cada una de las **combinaciones de los factores A y B**.

	Factor B					
Factor A	**1**	**2**		**.**	**.**	**J**
1	Y_{111} Y_{112} . . Y_{11K}	Y_{121} Y_{122} . . Y_{12K}	.	.	.	Y_{1J1} Y_{1J2} . . Y_{1JK}
2	Y_{211} Y_{212} . . Y_{21K}	Y_{221} Y_{222} . . Y_{22K}	.	.	.	Y_{2J1} Y_{2J2} . . Y_{2JK}
.						
.	.	.		Y_{ijl}	.	.
.	.	.		.	.	.
I	Y_{I11} Y_{I12} . . Y_{I1K}	Y_{I21} Y_{I22} . . Y_{I2K}	.	.	.	Y_{IJ1} Y_{IJ2} . . Y_{IJK}

Cada observación Y_{ijk}puede escribirse como:

$$Y_{ijk} = \eta_{ij} + \varepsilon_{ijk}$$

donde ε_{ijk} representa el error, la media η_{ij} (que depende del cada nivel i del Factor A (Filas) y de cada j nivel del Factor B (Columnas)) y el subíndice k identifica la replicación dentro de cada casillero.

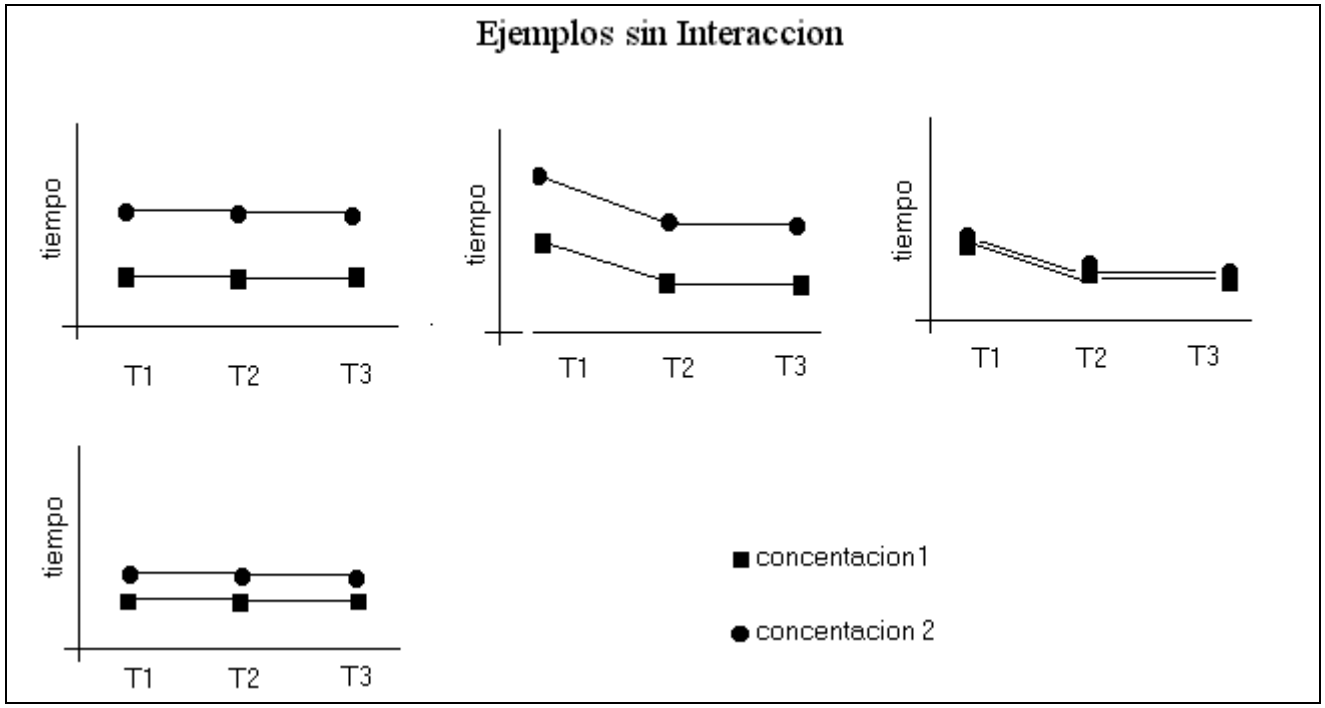
Asumiremos que $\varepsilon_{ijk} \sim N(0,\sigma^2)$ independientes.

Cuando el número de observaciones dentro de cada casillero es constante decimos que el **diseño es balanceado**. Vamos a considerar el caso balanceado.
Para cada observación, podríamos considerar un modelo que involucre una **media general**, el **efecto del tratamiento** y el **efecto de la concentración de adrenalina:**

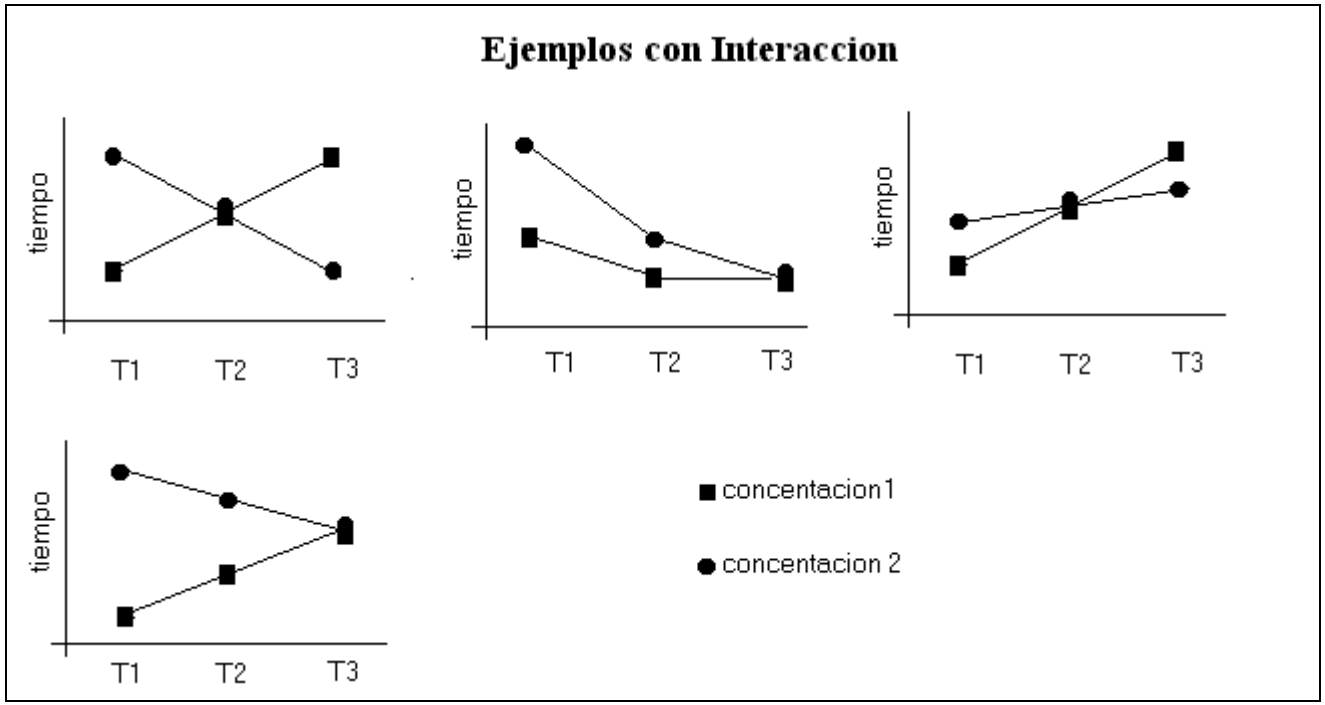
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

Esto es lo que conocemos como Modelo Aditivo.

Veamos que podría ocurrir con η_{ij} .



Sin embargo, podría ocurrir que el efecto de cierto tratamiento no sea el mismo para los distintos niveles de concentración de adrenalina. En este caso diríamos que hay interacción.



¿Cómo representar esto en el modelo? **Deberíamos pensar en un Modelo No Aditivo.**

Escribimos cada observación Y_{ijk} puede escribirse como:

$$Y_{ijk} = \eta_{ij} + \varepsilon_{ijk}$$

Podemos pensar que cada η_{ij} es una suma de 4 términos:

- Una media general, μ
- Efecto del nivel i del Factor A: α_i
- Efecto del nivel j del Factor B: β_j
- Interacciones ij: γ_{ij}

Luego

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

Notemos que

$$\eta_{ij} = \overline{\eta}_{..} + (\overline{\eta}_{i.} - \overline{\eta}_{..}) + (\overline{\eta}_{.j} - \overline{\eta}_{..}) + (\eta_{ij} - \overline{\eta}_{i.} - \overline{\eta}_{.j} + \overline{\eta}_{..})$$

que es de la forma $\mu + \alpha_i + \beta_j + \gamma_{ij}$ donde

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

Estimación

Tenemos que minimizar

$$S = \sum_{i,j} \sum_{k} (Y_{ijk} - \eta_{ij})^2$$

Obtenemos el estimador de mínimos cuadrados de η_{ij} resolviendo

$$\frac{\partial S}{\partial \eta_{ij}} = (-2) \sum_{k} (Y_{ijk} - \eta_{ij}) = 0$$

con lo cual

$$\hat{\eta}_{ij} = \bar{Y}_{ij.}$$

y queda

$$S_{\Omega} = \sum_{i,j} \sum_{k} (Y_{ijk} - \bar{Y}_{ij.})^2$$

Notemos que en este caso la matriz de diseño **X** es:

$$\begin{bmatrix} 1 & 0 & .. & 0 \\ : & : & .. & 0 \\ 1 & 0 & .. & : \\ 0 & 1 & .. & 0 \\ : & : & .. & : \\ 0 & 1 & .. & : \\ : & 0 & .. & 1 \\ 0 & : & .. & 1 \end{bmatrix} \qquad \eta = \begin{bmatrix} \eta_{11} \\ : \\ : \\ \eta_{IJ} \end{bmatrix} \qquad \text{rg(X)} = \text{p} = \text{IJ}$$

Por lo tanto todas funciones paramétricas son estimables, en particular:

$$\mu \text{ , } \alpha_i \text{ , } \beta_j \text{ y } \gamma_{ij}$$

Luego, por el Teorema de Gauss-Markov, los estimadores de mínimos cuadrados de μ , α_i ,β_j y γ_{ij} los obtenemos reemplazando a η_{ij} por su estimador $\hat{\eta}_{ij}$

Así obtenemos:

$$\hat{\mu} = \overline{\hat{\eta}}_{..}$$

$$\hat{\alpha}_i = (\overline{\hat{\eta}}_{i.} - \overline{\hat{\eta}}_{..})$$

$$\hat{\beta}_j = (\overline{\hat{\eta}}_{.j} - \overline{\hat{\eta}}_{..})$$

$$\hat{\gamma}_{ij} = (\hat{\eta}_{ij} - \overline{\hat{\eta}}_{i.} - \overline{\hat{\eta}}_{.j} + \overline{\hat{\eta}}_{..})$$

Resultando

$$\hat{\mu} = \bar{y}_{...}$$

$$\hat{\alpha}_i = (\bar{y}_{i..} - \bar{y}_{...})$$

$$\hat{\beta}_j = (\bar{y}_{.j.} - \bar{y}_{...})$$

$$\hat{\gamma}_{ij} = (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})$$

La hipótesis de igualdad de los efectos de los I niveles del Factor A (filas) puede plantearse mediante la hipótesis nula:

$$H_A: \alpha_1 = \alpha_2 = ...= \alpha_I = 0,$$

la hipótesis de igualdad de los J niveles del Factor B (columnas) se plantea como:

$$H_B: \beta_1 = \beta_2 = ...= \beta_J = 0,$$

mientras que la ausencia de interacciones, la testearíamos a través de la hipótesis

$$H_{AB}: \gamma_{11} = \gamma_{12} = ...= \gamma_{IJ} = 0.$$

La ausencia de interacciones implica que la diferencia de medias de dos niveles de un factor es la misma para todos los niveles del otro factor.

La suma de cuadrados puede ser reescrita como:

$$S = \sum_{i,j}\sum_{k} (Y_{ijk} - \eta_{ij})^2 = \sum_{i,j}\sum_{k} (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

$$= \sum_{i,j}\sum_{k} ((Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_{ij}) + (\hat{\mu} - \mu) + (\hat{\alpha}_i - \alpha_i) + (\hat{\beta}_j - \beta_j) + (\hat{\gamma}_{ij} - \gamma_{ij}))^2$$

y usando las restricciones

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

queda

$$S = S_\Omega + IJK(\hat{\mu} - \mu)^2 + JK\sum_i (\hat{\alpha}_i - \alpha_i)^2 + IK\sum_j (\hat{\beta}_j - \beta_j)^2 + K\sum_{i,j} (\hat{\gamma}_{ij} - \gamma_{ij})^2$$

Esta expresión es muy útil pues bajo H_A, H_B, o H_{AB} permite ver que los estimadores son los mismos que bajo Ω.

Por ejemplo, bajo H_A: $\alpha_1 = \alpha_2 = ...= \alpha_I = 0,$ tendríamos

$$S = S_\Omega + IJK(\hat{\mu} - \mu)^2 + JK\sum_i \hat{\alpha}_i^{\,2} + IK\sum_j (\hat{\beta}_j - \beta_j)^2 + K\sum_{i,j} (\hat{\gamma}_{ij} - \gamma_{ij})^2$$

Por lo tanto, S se minimiza cuando

$$\mu = \hat{\mu}, \beta_j = \hat{\beta}_j \text{ y ademas } \gamma_{ij} = \hat{\gamma}_{ij}$$

En este caso además tendríamos

$$S_{A\omega} = S_\Omega + JK\sum_i \hat{\alpha}_i^{\,2}$$

Análogamente

$$\mathrm{S}_{B\omega} = S_\Omega + IK\sum_j \hat{\beta}_j^2$$

$$\mathrm{S}_{AB\omega} = S_\Omega + K\sum_{i,j} \hat{\gamma}_{ij}^2$$

Para testear, por ejemplo H_A

$$\frac{\mathrm{n-r}}{\mathrm{q}}\frac{\mathrm{S}_{A\omega} - S_\Omega}{S_\Omega} = \frac{\mathrm{n-IJ}}{\mathrm{I-1}}\frac{JK\sum_i \hat{\alpha}_i^{\,2}}{S_\Omega} = \frac{\mathrm{n-IJ}}{\mathrm{I-1}}\frac{JK\sum_i \hat{\alpha}_i^{\,2}}{\sum_{i,j}\sum_k (Y_{ijk} - \bar{Y}_{ij.})^2}$$

En cuanto a los grados de libertad de cada una, es decir q, es el número de condiciones l.i. estimables impuestas por cada hipótesis.

Los grados de libertad de cada una de estas sumas son:

- **S_A: I-1**
- **S_B: J-1**
- **S_{AB}: (I-1)(J-1)**

- **SE: IJ(K-1)**
- **ST: n-1=I*J*K-1**

Por lo tanto la Tabla de Análisis de la Varianza será:

(Extraída de Scheffé, 1959)

TABLE 4.3.1

ANALYSIS OF VARIANCE OF THE TWO-WAY LAYOUT WITH K OBSERVATIONS PER CELL

Source	SS	d.f.
A main effects	$\mathrm{SS}_A = JK\sum_i (y_{i..} - y_{...})^2$	$I-1$
B main effects	$\mathrm{SS}_B = IK\sum_j (y_{.j.} - y_{...})^2$	$J-1$
AB interactions	$\mathrm{SS}_{AB} = K\sum_i\sum_j (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$	$(I-1)(J-1)$
Error	$\mathrm{SS}_e = \sum_i\sum_j\sum_k (y_{ijk} - y_{ij.})^2$	$IJ(K-1)$
"Total"	$\mathrm{SS}_{\text{"tot"}} = \sum_i\sum_j\sum_k (y_{ijk} - y_{...})^2$	$IJK-1$

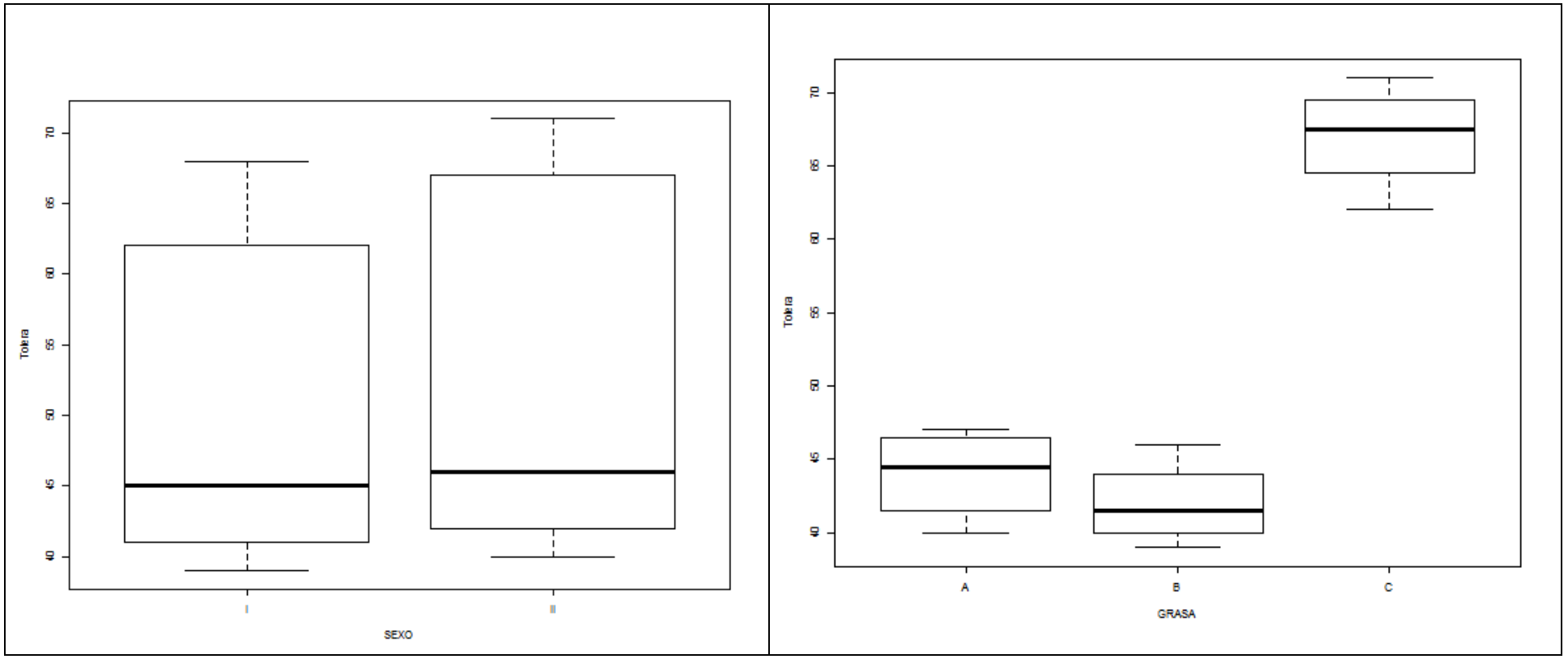
EJEMPLO: 2 factores con replicaciones.

Supongamos que nos interesa estudiar el efecto del porcentaje de grasa corporal (factor A, 3 niveles) y del sexo (factor B) en la **tolerancia** al ejercicio físico en personas de 25 a 35 años de edad. Esta tolerancia se mide *en minutos antes de que ocurra la fatiga* en sujetos realizando bicicleta fija.

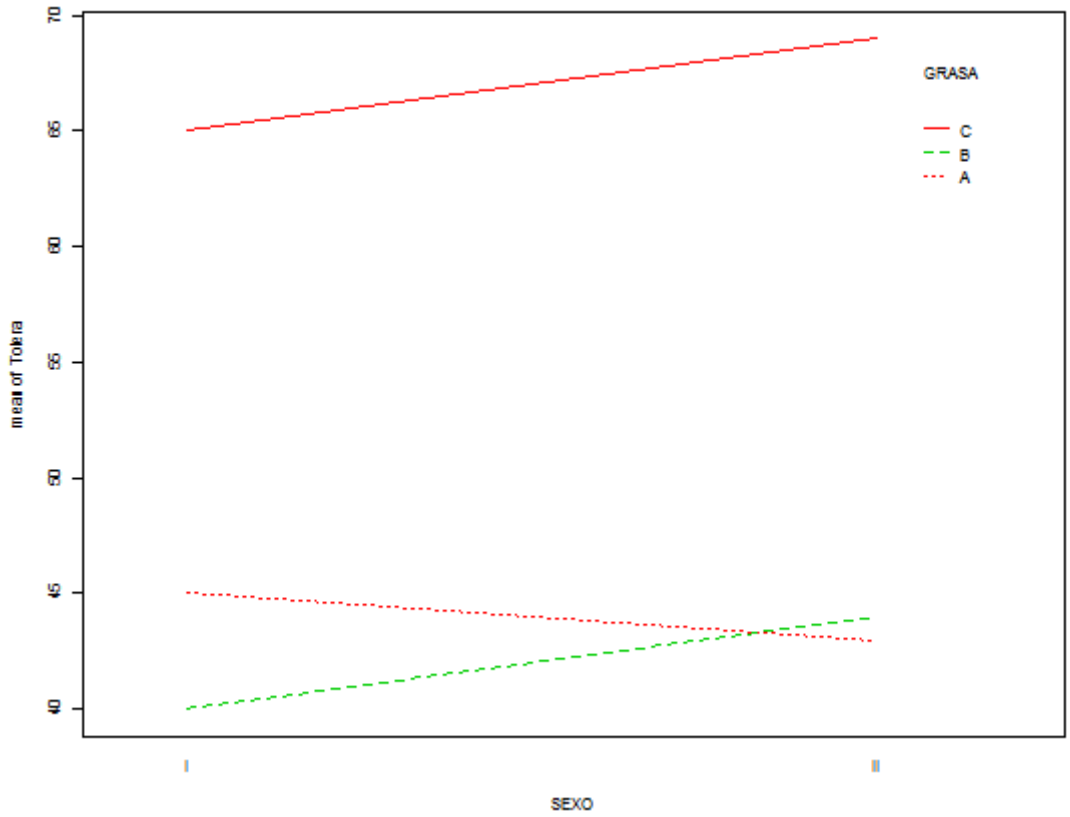
Dos sujetos fueron sometidos al test de tolerancia para cada grupo de sexo-grasa. A partir de los datos obtenidos se calculó la siguiente tabla de análisis de la varianza para el modelo:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad i=1, 2, 3 \quad j=1, 2, \quad k=1, 2$$

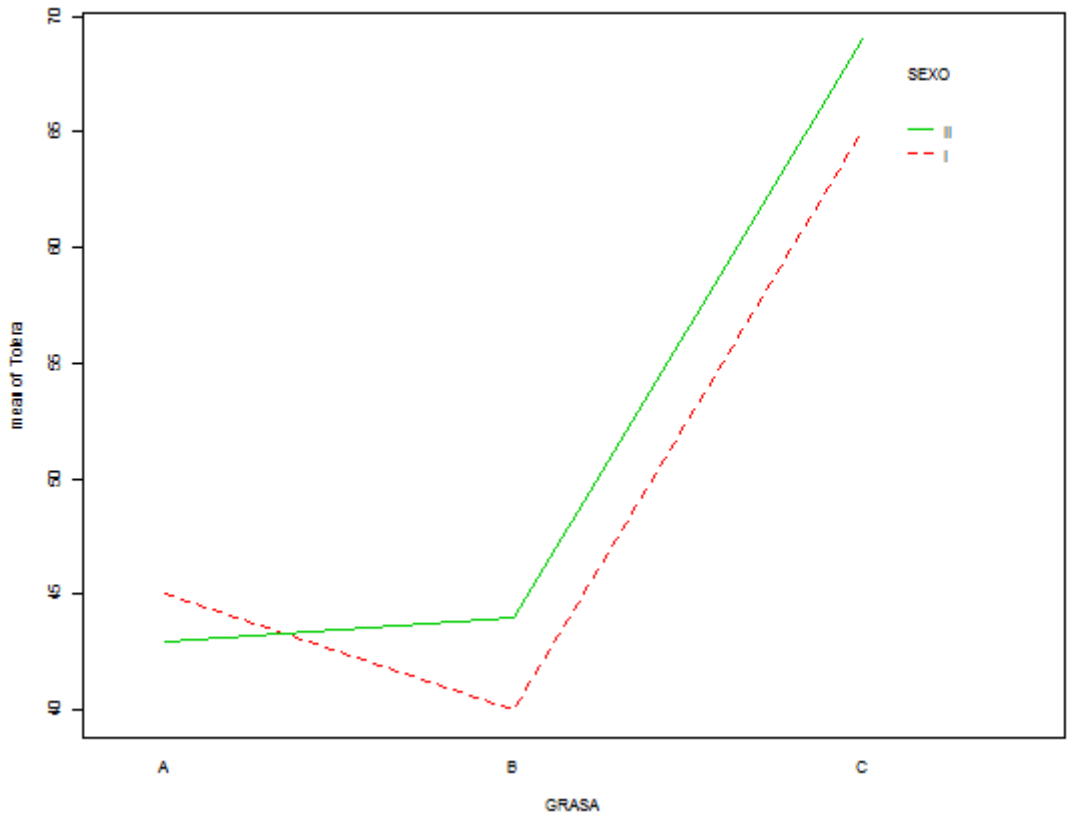
```
grasa<-
read.table("C:\\Users\\Ana\\ModeloLineal\\doctex\\grasa.txt",header=T)
grasa
attach(grasa)
names(grasa)
plot(Tolera~ SEXO + GRASA, data=grasa)
```



```
interaction.plot(SEXO,GRASA,Tolera,col=2:3)
```



```
interaction.plot(SEXO,GRASA,Tolera,col=2:3)
```



ANALYSIS OF VARIANCE TABLE FOR TOLERA

```
g <- lm(Tolera~GRASA*SEXO, grasa)
anova(g)

Analysis of Variance Table

Response: Tolera
           Df Sum Sq Mean Sq F value    Pr(>F)
GRASA       2   1544  772.00 74.7097 5.754e-05 ***
SEXO        1     12   12.00  1.1613    0.3226
GRASA:SEXO  2     24   12.00  1.1613    0.3747
Residuals   6     62   10.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comenzamos por testear la hipótesis de ausencia de interacciones

$$H_{AB}: \gamma_{11} = \gamma_{12} = ...= 0$$

Como el **p-valor** obtenido para el test de F correspondiente es **0.3747**, no podemos rechazar H_{AB},

¿Por qué testeamos primero H_{AB}?

No tiene tiene sentido testear los efectos principales cuando hay interacción, a menos que hubiera un interés específico. Un p-valor bajo en el test para H_{AB} sugiere que cada factor tiene un efecto en la variable de respuesta, pero el tamaño de este efecto depende del nivel del otro factor. Por esta razón testeamos en primer término H_{AB}.

Si el p-valor para testear H_{AB} **no** es pequeño, testeamos H_A y $H_{B.}$

Si en cambio, el p-valor es pequeño, no podemos descartar la presencia de interacciones y comparamos las medias entre los distintos niveles de un factor, fijado el nivel del otro factor.

Como en este ejemplo **p-valor** es **0.3747** y no podemos rechazar H_{AB}, estamos en condiciones de testear H_A y H_B.
Si deseáramos verificar si el sexo tiene algún efecto sobre la tolerancia al ejercicio físico deberíamos testear

$$H_B: \beta_1 = \beta_2 = 0,$$

y como el **p-valor** del test correspondiente es `0.3226`, no podemos rechazar la hipótesis de que el efecto del sexo sea nulo.

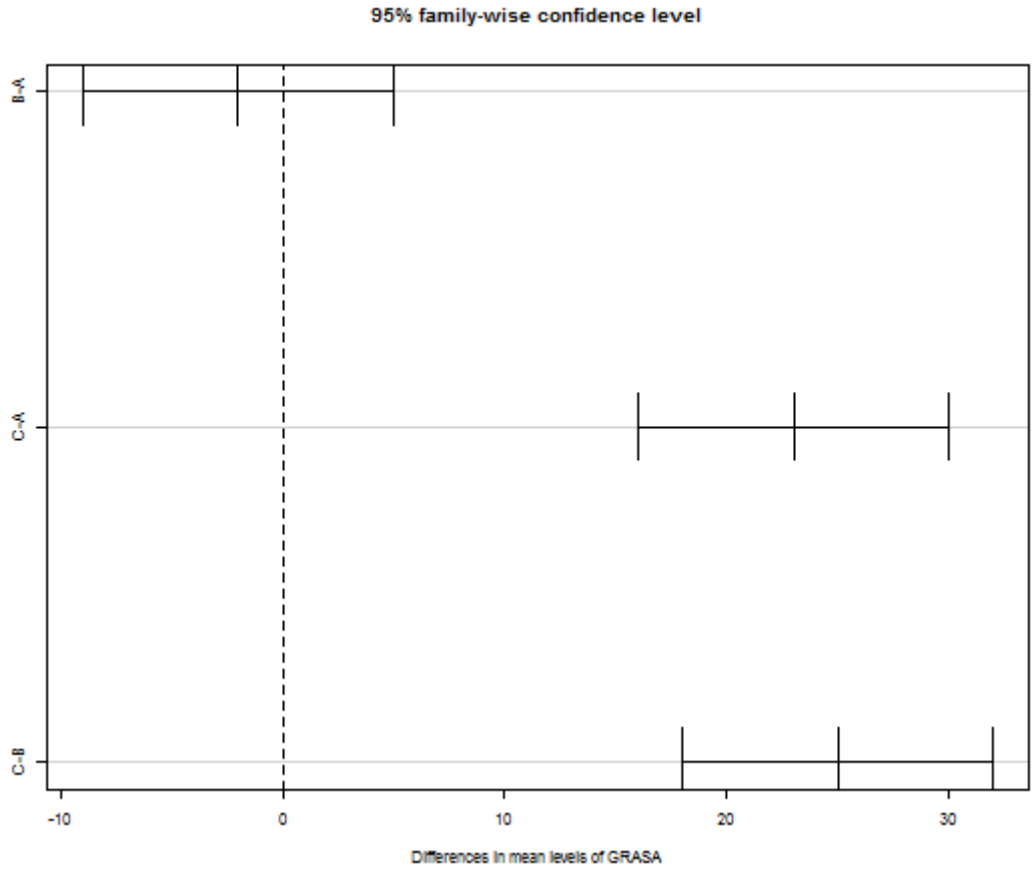
Por otra parte, podría interesarnos testear

$$H_A: \alpha_1 = \alpha_2 = \alpha_3 = 0.$$

El **p-valor** obtenido para el test de F correspondiente es **0.0001**, en consecuencia rechazamos la hipótesis de que el efecto del porcentaje de grasa es el mismo para los tres niveles.

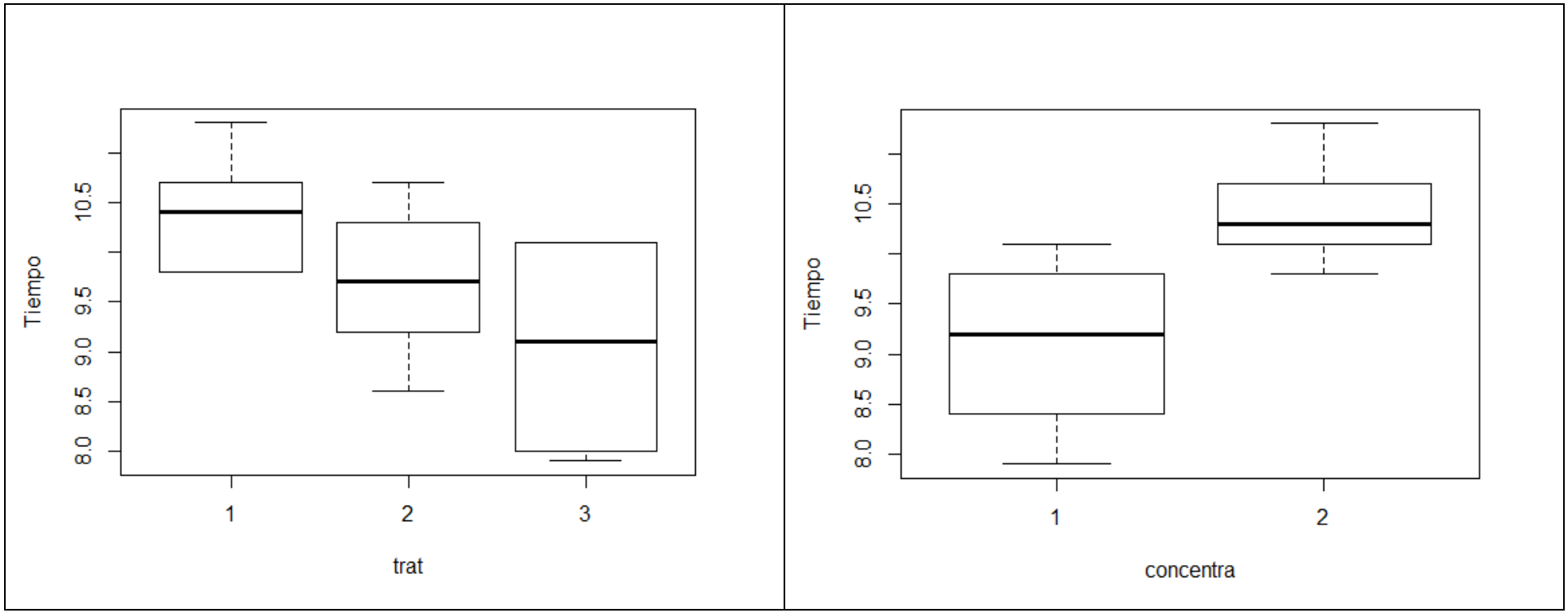
Si nos interesase realizar intervalos de confianza simultáneos para las diferencias entre las medias de los niveles de porcentaje de grasa podemos calcular los intervalos mediante el método de Tukey con un nivel global de 95%:

```
salida<-aov(Tolera~SEXO*GRASA)
tolera.tuk<-TukeyHSD(salida,"GRASA",ordered=FALSE,conf.level=0.95)
plot(tolera.tuk)
```



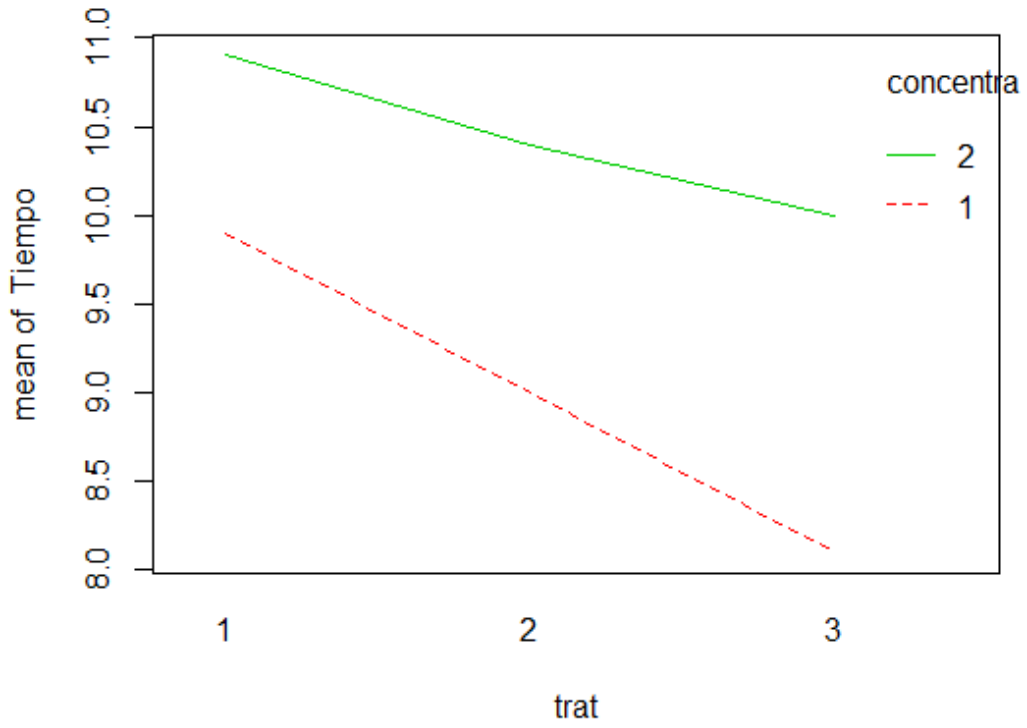
EJEMPLO: 2 factores con replicaciones

```
plasma<-
read.table("C:\\Users\\Ana\\ModeloLineal\\doctex\\plasma.txt",header=T)
attach(plasma)
names(plasma)
trat<- factor(TRATA)
concentra<- factor(CONCENTRA)
plot(Tiempo~trat + concentra, data=plasma)
```

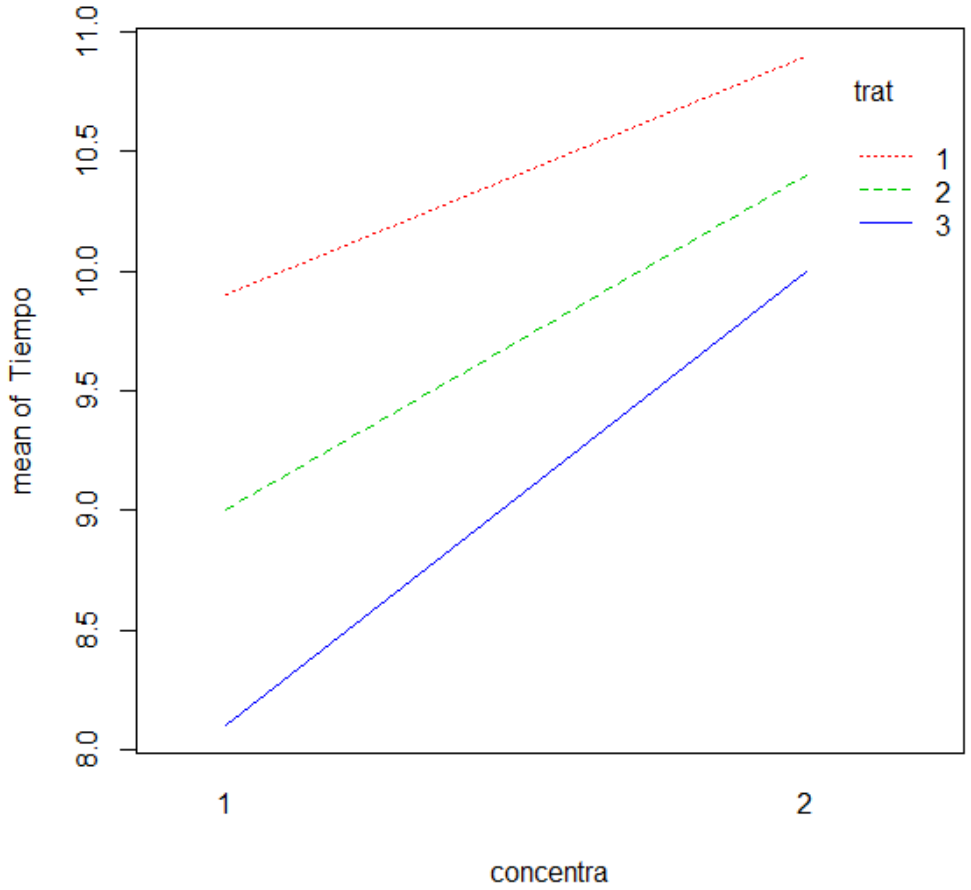


$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad i=1, 2, 3 \quad j=1, 2, \quad k=1, 2,3$$

```
interaction.plot(trat,concentra,Tiempo,col=2:3)
```



```
interaction.plot(concentra,trat,Tiempo,col=2:4)
```



```
g <- lm(Tiempo~trat*concentra, plasma)
anova(g)
```

```
Analysis of Variance Table

Response: Tiempo
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
trat	2	5.470	2.7350	37.2955	7.084e-06	***
concentra	1	9.245	9.2450	126.0682	1.011e-07	***
trat:concentra	2	0.610	0.3050	4.1591	0.04244	*
Residuals	12	0.880	0.0733			

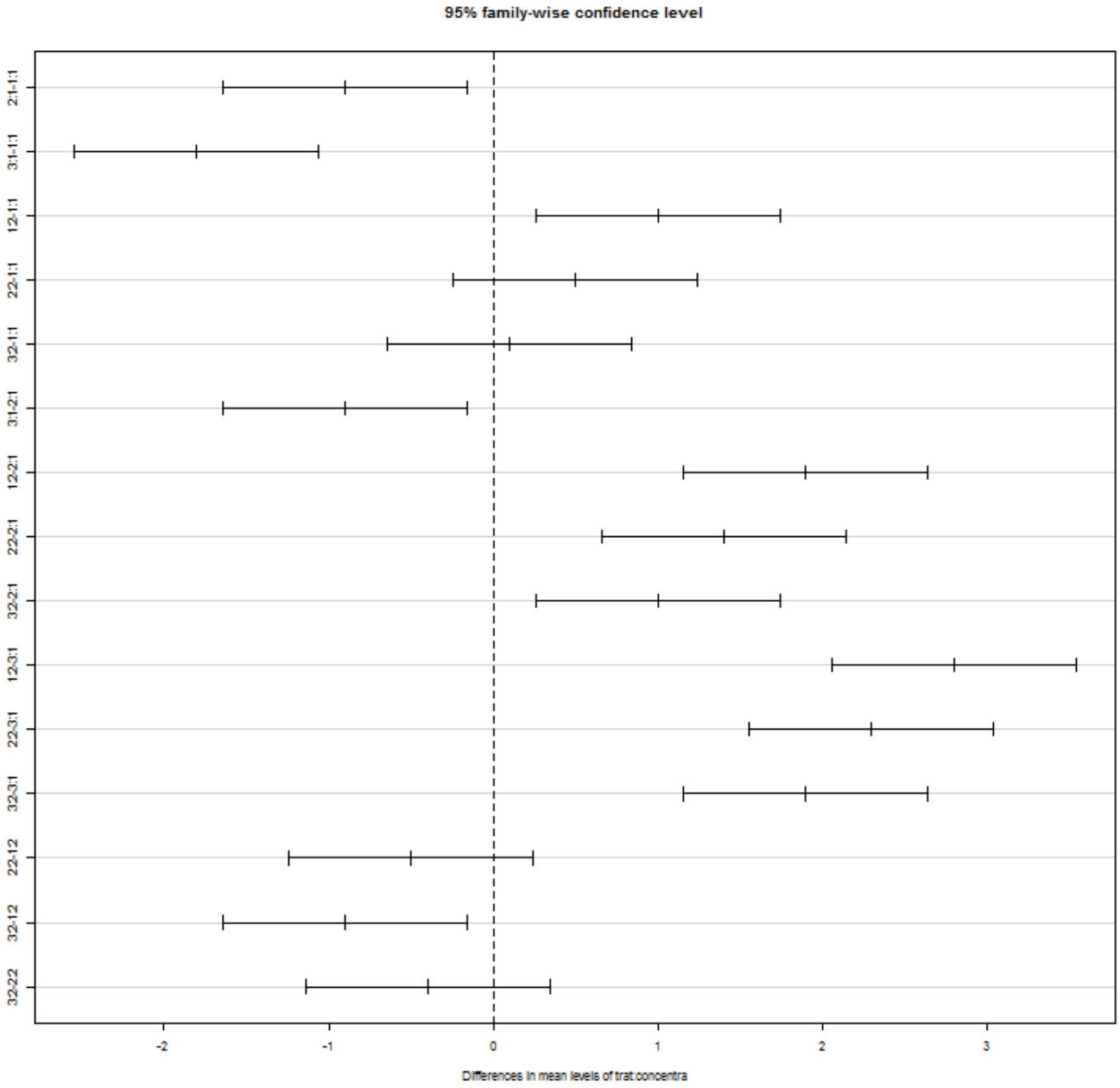
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como antes comenzamos por testear la hipótesis nula H_{AB}. En este caso la hipótesis nula es rechazada al 5%. Compararemos las medias de todas las combinaciones.

```
tiempo.tuk<-TukeyHSD(salida,ordered=FALSE,conf.level=0.95)

par(cex=0.5)
plot(tiempo.tuk,cex=2)


Tambien podria escribirse:
tiempo.tuk<-
TukeyHSD(salida,"trat:concentra",ordered=FALSE,conf.level=0.95)
```


95% family-wise confidence level
2:1-1:1
3:1-1:1
1:2-1:1
2:2-1:1
3:2-1:1
3:1-2:1
1:2-2:1
2:2-2:1
3:2-2:1
1:2-3:1
2:2-3:1
3:2-3:1
2:2-1:2
3:2-1:2
3:2-2:2
-2
-1
0
1
2
3
Differences in mean levels of trat:concentra

Una forma de resumir esta información es considerando:

TRAT	CONCENTRA	MEAN	GROUPS
1	2	10.900	I
2	2	10.400	I I
3	2	10.000	.. I
1	1	9.9000	.. I
2	1	9.0000	 I
3	1	8.1000	 I

Donde se ve que hay cuatro grupos de medias que no difieren significativamente unas de otras.