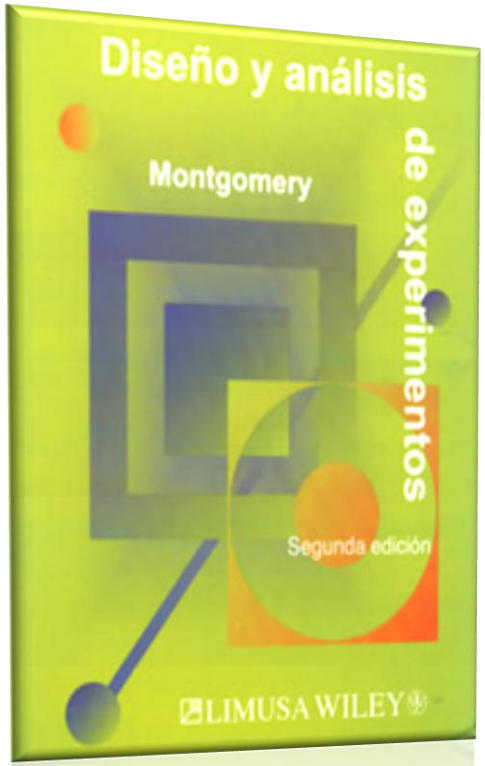
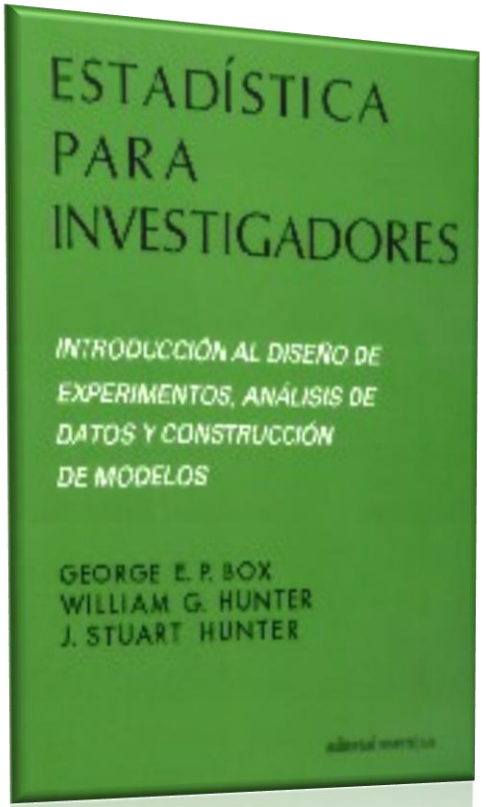
EXPERIMENTOS CON UN SOLO FACTOR: EL ANÁLISIS DE VARIANZA

PROF. ZORITZA BRAVO

Bibliografía recomendada





Notación


Factores
• Niveles = tratamientos
Nº de niveles
• a
Nº de réplicas
• n

Introducción

Este modelo es el más sencillo del diseño de experimentos, en el cual la variable respuesta puede depender de la influencia de un único factor, de forma que el resto de las causas de variación se engloban en el error experimental

Se supone que el experimento ha sido aleatorizado por completo, es decir, todas las unidades experimentales han sido asignadas al azar a los tratamientos

Existen dos tipos de modelos: el de efectos fijos y el de efectos aleatorios

Efectos fijos y aleatorios

I. Los niveles del factor se seleccionan de modo específico por el experimentador. Esto constituye el llamado modelo de ***efectos fijos***.

II. Los niveles de un factor son una muestra aleatoria de una población mayor de tratamientos. Esto es el modelo de ***efectos aleatorios***.

Ejemplos

Una firma comercial desea conocer la influencia que tiene el nivel cultural de las familias en el éxito de una campaña publicitaria sobre cierto producto. Para ello, aprovecha los resultados de una encuesta anterior clasificando las respuestas en tantos grupos como niveles culturales ha establecido.

Un solo *factor*, ya que la firma sólo está interesada en averiguar si los distintos niveles culturales influyen o no de la misma manera sobre las ventas, no importándole la influencia del resto de los factores que pueden inducir a una mayor o menor tendencia a la compra

Diseño de efectos fijos

Modelo de efectos fijos

Y: variable respuesta

Consideramos ***a*** poblaciones diferentes y comparamos la respuesta a un tratamiento, o único nivel de un factor.

En la población *i-ésima* ($i = 1, \ldots, a$) se toman $\boldsymbol{n_i}$ observaciones.

La respuesta se cuantifica mediante $\boldsymbol{y_{ij}}$, donde $i = 1, \ldots, a$ se refiere a la población en estudio y $j = 1, \ldots, n_i$ se refiere a la observación *j -ésima.*

Modelo de efectos fijos

Y: variable respuesta

Consideramos ahora un factor con ***a*** niveles, es decir, en total ***a*** tratamientos, y una única población.

Se observa la respuesta y_{ij} del tratamiento *i-ésimo* a n_i observaciones de la población.

Modelo de efectos fijos

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

$$\begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n_i \\ \sum_{i=1}^{a} n_i = N \end{cases}$$

El valor medio de *Y*, la variable respuesta, en la población o nivel *i-ésimo*

Error aleatorio

Modelo de efectos fijos

Alternativamente, se puede expresar de esta manera:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

$$i = 1, \ldots, a;\ j = 1, \ldots, n$$

suponiendo grupos de igual tamaño

Modelo de efectos fijos

y_{ij}: es la observación *(i, j)-ésima*

μ: la media global

τ_i: es el efecto del *i-ésimo* tratamiento

ε_{ij} es el error aleatorio, tal que $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes entre sí, $E[\varepsilon_{ij}] = 0$ y $Var[\varepsilon_{ij}] = \sigma^2$

Modelo de efectos fijos

Se supone, además, que las unidades experimentales están en un ambiente uniforme, lo cual lleva a un diseño completamente aleatorizado.

En el modelo de efectos fijos, los efectos de los tratamientos τ_i se definen como desviaciones respecto a la media general, por lo que:

$$\sum_{j=1}^{n}\sum_{i=1}^{a}\tau_i = 0$$

$$\sum_{i=1}^{a} n\tau_i = 0 \Longrightarrow$$

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

Modelo de efectos fijos

$$E[y_{ij}] = \mu + \tau_i$$

Esperanza del tratamiento *i*

$$i = 1, \ldots, a$$

Prueba de Hipótesis

$$H_0 \equiv \mu_1 = \mu_2 = \cdots = \mu_a$$

$$H_1 \equiv \mu_i \neq \mu_j \text{ (para al menos un par)}$$

$$H_0 \equiv \tau_1 = \tau_2 = \cdots = \tau_a$$

$$H_1 \equiv \tau_i \neq 0, \quad \exists i$$

Modelo de efectos fijos

Nivel	Observaciones	Totales	Promedios
1	$y_{11} \quad y_{12} \quad \cdots \quad y_{1n}$	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	$y_{21} \quad y_{22} \quad \cdots \quad y_{2n}$	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\cdots\cdots\cdots$	$\cdots$	$\cdots$
a	$y_{a1} \quad y_{a2} \quad \cdots \quad y_{an}$	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
		$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

$$y_{i\bullet} = \sum_{i=1}^{n} y_{ij}$$

$$\overline{y}_{i\bullet} = y_{i\bullet} / n,\ i = 1,\ldots,a$$

$$y_{\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$$

$$\overline{y}_{\bullet\bullet} = y_{\bullet\bullet} / N,\ N = an$$

Descomposición de la suma de cuadrados total

La idea es descubrir cómo se reparte la variabilidad total de la muestra. Una posible medida de variabilidad total es la suma de cuadrados, denominada total, o suma total de cuadrados corregida:

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}((\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}))^2 =$$

$$= n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 =$$

$$= SCTra + SCE.$$

Grados de libertad

Se tiene un total de ***an*** observaciones y ***a*** tratamientos

- *SCT* tiene (an − 1) grados de libertad.
- *SCTra* tiene (a − 1) grados de libertad.
- *SCE* tiene a(n−1) grados de libertad, porque hay *n* réplicas dentro de cada tratamiento, es decir, se tienen *(n−1)* grados de libertad para estimar el error experimental.

Al tener ***a*** tratamientos, se tiene un total de *a(n − 1)* grados de libertad.

Estimadores de la varianza

$$SCE = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2\right]$$

Si el término entre paréntesis se divide entre *n−1*, se obtiene la varianza del tratamiento *i*

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2$$

Estimadores de la varianza

Se puede estimar la varianza poblacional combinando dichas varianzas por grupos:

$$\frac{(n-1)s_1^2+(n-1)s_2^2+\cdots+(n-1)s_a^2}{(n-1)+(n-1)+\cdots+(n-1)} = \frac{\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2\right]}{\sum_{i=1}^{a}(n-1)} =$$

$$= \frac{SCE}{N-a} \qquad N = a \cdot n$$

Si no hay diferencias entre los ***a*** tratamientos, se puede estimar la varianza poblacional σ^2 como →

$$\frac{SCTra}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}{a-1}$$

Estimadores de la varianza

Se dispone, así de dos posibles estimadores de σ^2

$$MCTra = \frac{SCTra}{a-1}$$

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

Cuando no existen diferencias entre las medias de los tratamientos, las estimaciones deben ser similares.

Estimadores de la varianza

Si consideramos las medias de cuadrados anteriores, entonces, se puede demostrar, sustituyendo, que

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

$$E(MCTra) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}.$$

De este modo, si para algún $\tau_i \neq 0$, entonces $E(MCTra) > \sigma^2$

Análisis estadístico

¿Cómo llevamos a cabo una prueba de hipótesis?

$$H_0 \equiv \mu_1 = \mu_2 = \cdots = \mu_a$$

$$H_1 \equiv \mu_i \neq \mu_j \text{ (para al menos un par)}$$

No hay diferencia en las medias de los tratamientos

$$H_0 \equiv \tau_1 = \tau_2 = \cdots = \tau_a$$

$$H_1 \equiv \tau_i \neq 0, \quad \exists i$$

Análisis estadístico

Como los errores ε_{ij} se distribuyen independientemente entre sí, según una *N(0, σ)*, entonces

$$\frac{SCE}{\sigma^2} \sim \chi^2_{N-a}$$
$$\frac{SCTra}{\sigma^2} \sim \chi^2_{a-1}$$

→ **Fisher**

Aplicando el teorema de *Cochran*, se tiene que SSE/σ² y SSTra/σ² son independientes, por lo que si $\tau_i = 0, \forall i$ →

$$F_0 = \frac{\frac{SCTra}{a-1}}{\frac{SCE}{N-a}} = \frac{MCTra}{MCE}$$

↓

Se distribuye como una *F* de Snedecor, $F_{a-1,N-a}$

Análisis estadístico

Si algún $\tau_i \neq 0$, entonces $E(MSTra) > \sigma^2$ entonces el valor del estadístico F_0 es mayor, obteniéndose una región crítica superior, de modo que se rechaza, a nivel α, la hipótesis nula de igualdad de tratamientos, si

$$F_0 > F_{\alpha,a-1,N-a}$$

Tabla ANOVA

$$H_0 \equiv \tau_1 = \tau_2 = \cdots \tau_a$$

$$H_1 \equiv \tau_i \neq 0, \quad \exists i$$

F. Variación	S. Cuadrados	gl	M. Cuadrados	F_0
Factor	$SCTra = n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$a-1$	$MCTra = \frac{SCTra}{a-1}$	$F_o = \frac{MCTra}{MCE}$
Error	$SCE = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2$	$N-a$	$MCE = \frac{SCE}{n-a}$	
Total	$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$N-1$		

Se rechaza H_0 a nivel α cuando

$F_0 > F_{\alpha,a-1,N-a}$

Estimación de los parámetros

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

donde $i = 1, \ldots, a$; $j = 1, \ldots, n$, se pueden estimar los parámetros μ y τ_i por el método de los Mínimos Cuadrados.

$$L = \sum_{i=1}^{a}\sum_{j=1}^{n} \varepsilon_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \mu - \tau_i)^2$$

Suma de los cuadrados de los errores

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

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$$

$$i = 1, 2, \ldots, a$$

Intervalos de confianza

Si se asume que los errores están distribuidos según una normal, entonces cada

$$\bar{y}_{i\cdot} \sim N\left(\mu_i, \sigma^2/n\right)$$

De este modo, cuando σ^2 es desconocida un intervalo de confianza al $100(1-\alpha)\%$ es

$$\left[\bar{y}_{i\cdot} \pm t_{\frac{\alpha}{2},N-a}\sqrt{\frac{MCE}{n}}\right]$$

Intervalo de confianza para la media μ_i del tratamiento *i-ésimo*

Intervalos de confianza

$$\left[(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \pm t_{\frac{\alpha}{2},N-a}\sqrt{\frac{2MCE}{n}}\right]$$

Intervalo de confianza para la diferencia en las medias de dos tratamientos cualesquiera μ_i - μ_j

Ejemplo

- 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).

Ejemplo

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. Las 25 observaciones deben asignarse al azar. Para ilustrar la forma en que puede aleatorizarse el orden de ejecución, supóngase que las observaciones se numeran como sigue:

% algodón					
15	1	2	3	4	5
20	6	7	8	9	10
25	11	12	13	14	15
30	16	17	18	19	20
35	21	22	23	24	25

Ejemplo

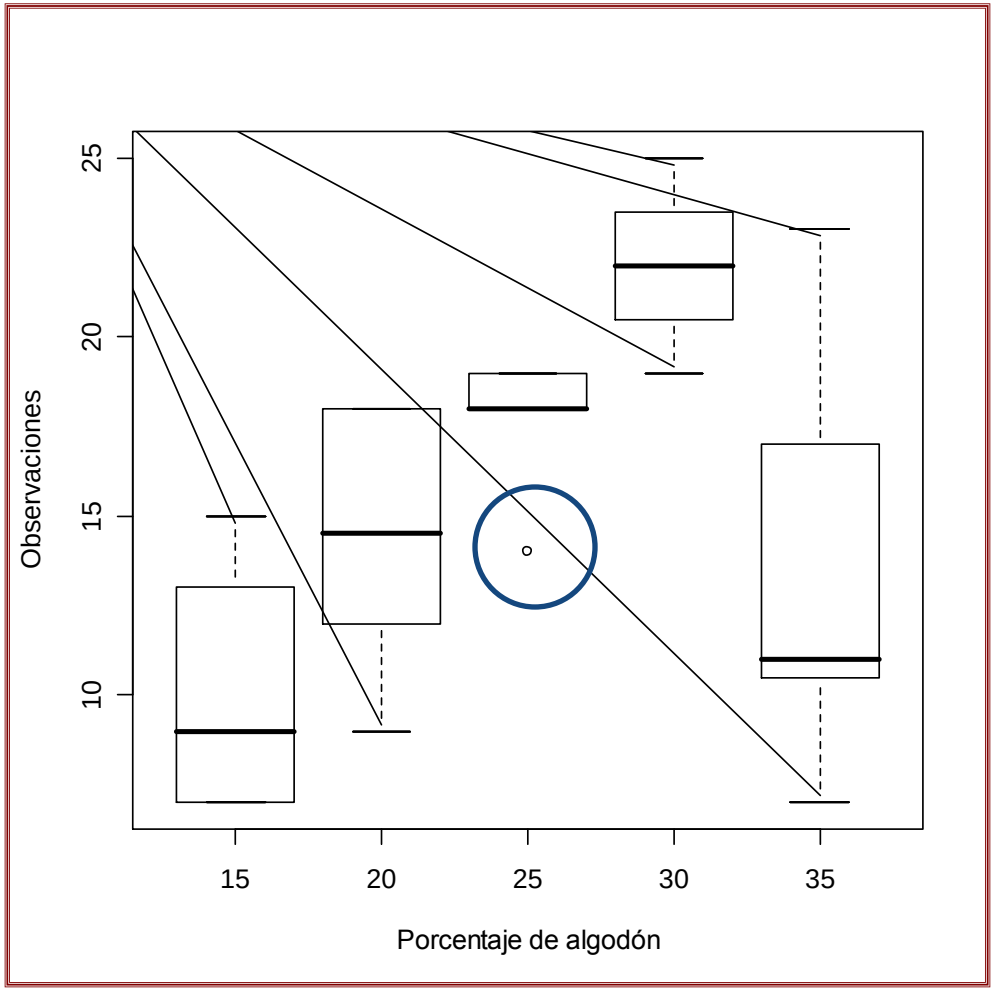
Ahora se elige al azar un número entre 1 y 25. Supongamos que es el 8, entonces la observación 8ª se ejecuta primero (es decir, a un 20% de algodón). A continuación se elige un número al azar entre 1 y 25, quitando el 8. Supongamos que es el 4, entonces la observación 4ª se ejecuta en segundo lugar (a un 15% de algodón). Se repite el proceso hasta completar las 25 observaciones.

Esta secuencia de prueba aleatorizada es necesaria para evitar que los resultados se contaminen por los efectos de variables desconocidas que pueden salir de control durante el experimento.

Ejemplo

% de algodón	Observaciones					Suma	Media
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						376	**15.04**

Ejemplo



- ✓ La gráfica indica que la resistencia a la tensión aumenta con el contenido de algodón hasta el 30%
- ✓ Más allá del 30% ocurre un notable decrecimiento en la resistencia
- ✓ La falta de traslape de las cajas sugiere una diferencia significativa entre los contenidos medios de las resistencias entre los grupos
- ✓Usando un 30% de algodón parece que se fabrican las mejores fibras, es decir, las de mayor fortaleza

Ejemplo

% de algodón	Observaciones	Suma	Media
15	7 7 15 11 9	49	9.8
20	12 17 12 18 18	77	15.4
25	14 18 18 19 19	88	17.6
30	19 25 22 19 23	108	21.6
35	7 10 11 15 11	54	10.8
		376	**15.04**

$\bar{y}_{1\cdot}$

$y_{\cdot\cdot}$ $y_{1\cdot}$ $\bar{y}_{\cdot\cdot}$

Hipótesis del modelo

Normalidad: ε_{ij} sigue una distribución normal

$E(\varepsilon_{ij}) = 0$

Homocedasticidad: $Var(\varepsilon_{ij}) = \sigma^2$

Independencia: ε_{ij} son independientes entre sí

Metodología

I. Estimar los parámetros del modelo.
II. Contrastar si el factor influye en la respuesta, es decir, si los valores medios de *Y* son diferentes al cambiar el nivel del factor.
III. Si el factor influye en la variable respuesta, es decir, las medias no son iguales, buscar las diferencias entre poblaciones (o niveles del factor).
IV. Diagnosis del modelo: comprobar si las hipótesis del modelo son ciertas mediante el análisis de los residuos.

Estimación de los parámetros

En este ejemplo, $a = 5$, $n_i = 5$ y $N = 25$. Las estimaciones puntuales de los parámetros son las siguientes:

$$\hat{\mu}_1 = \bar{y}_{1\cdot} = 9{,}8$$

$$\hat{\mu}_2 = \bar{y}_{2\cdot} = 15{,}4$$

$$\hat{\mu}_3 = \bar{y}_{3\cdot} = 17{,}6$$

$$\hat{\mu}_4 = \bar{y}_{4\cdot} = 21{,}6$$

$$\hat{\mu}_5 = \bar{y}_{5\cdot} = 10{,}8$$

mean(resistencia[porcentaje==15])

Análisis de varianza

$$\begin{cases} H_0 : \mu_1 = \cdots = \mu_a \text{ (el factor no influye)} \\ H_1 : \text{algún factor es diferente (el factor influye)} \\ \text{nivel de significación } \alpha \end{cases}$$

$$SCT = \sum\sum y_{ij}^2 - n\bar{y}_{..}^2$$

$$SCTra = \sum n_i \bar{y}_{i.}^2 - n\bar{y}_{..}^2$$

$$SCE = SCT - SCTra$$

mode1=aov(resistencia~porcentaje)
summary(mode1)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
porcentaje	4	475.76	118.94	14.757	9.128e06***
Residuals	20	161.20	8.06		

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

Análisis de varianza

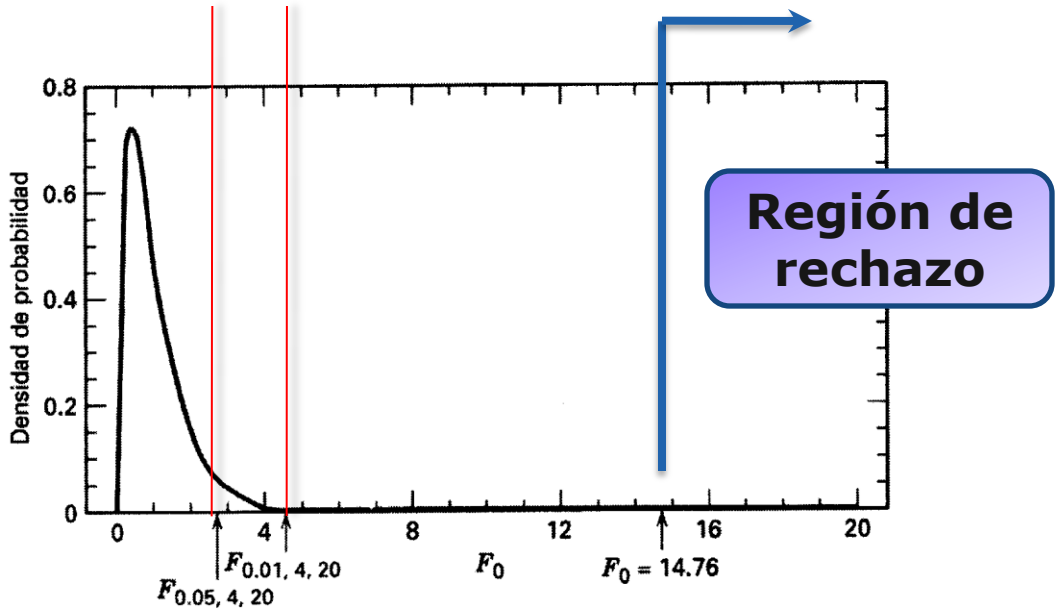
$$F_{4,20;0,1} = 2,2489$$

$$F_{4,20;0,05} = 2,8661$$

$$F_{4,20;0,01} = 4,4307$$

qf(0.95, 4, 20)=2,8661

Por lo tanto, rechazamos H_0 a los niveles anteriores y concluimos que hay diferencias entre los tratamientos.



Diagnosis del modelo

$$e_{ij} = y_{ij} - y_{ij}$$

$$y_{ij} = \mu + \hat{\tau}_i = \bar{y}_{\bullet\bullet} + \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right) = \bar{y}_{i\bullet}$$

$$d_{ij} = \frac{e_{ij}}{\sqrt{MCE}}$$

```
par(mfrow=c(1,3),oma=c(1,1,1,1))
hist(rstandard(mode1),main='Histograma de los residuos
estandarizados',col="gray60")
boxplot(rstandard(mode1),main="Diagrama de cajas de los
residuos",col='gray')
qqnorm(rstandard(mode1), main='Gráfica de probabilidad
normal de los residuos')
qqline(rstandard(mode1))
title("Chequeando normalidad de los residuos",outer=TRUE)
```

Diagnosis del modelo: Normalidad

Chequeando normalidad de los residuos

Histograma de los residuos estandarizados

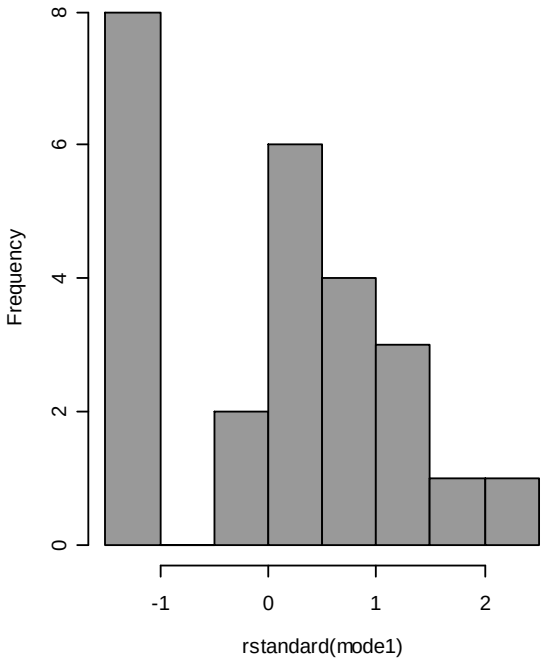
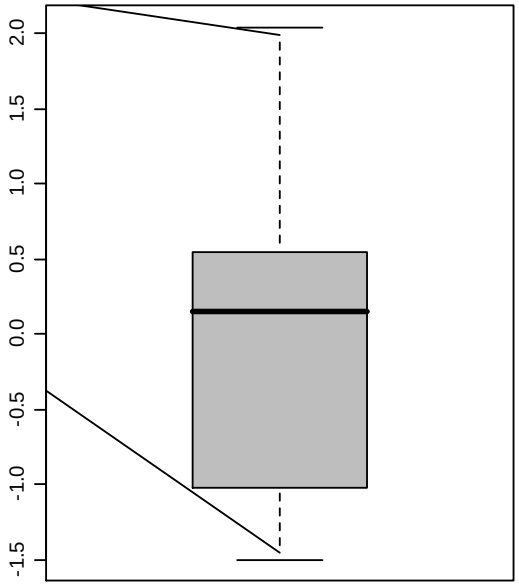
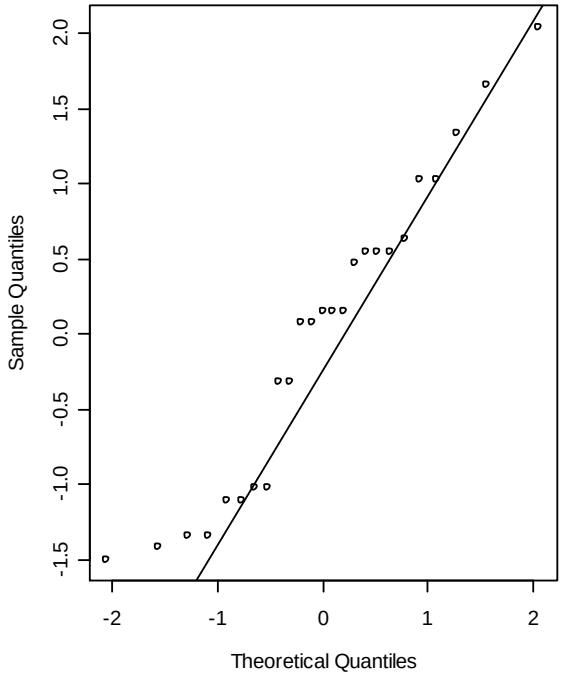


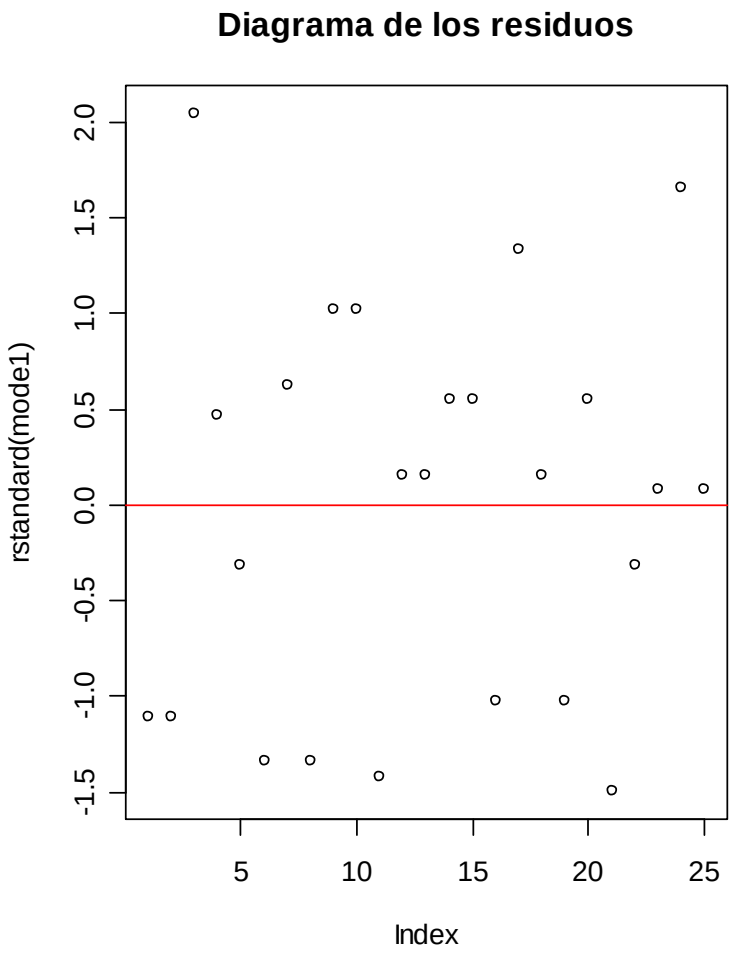
Diagrama de cajas de los residuos

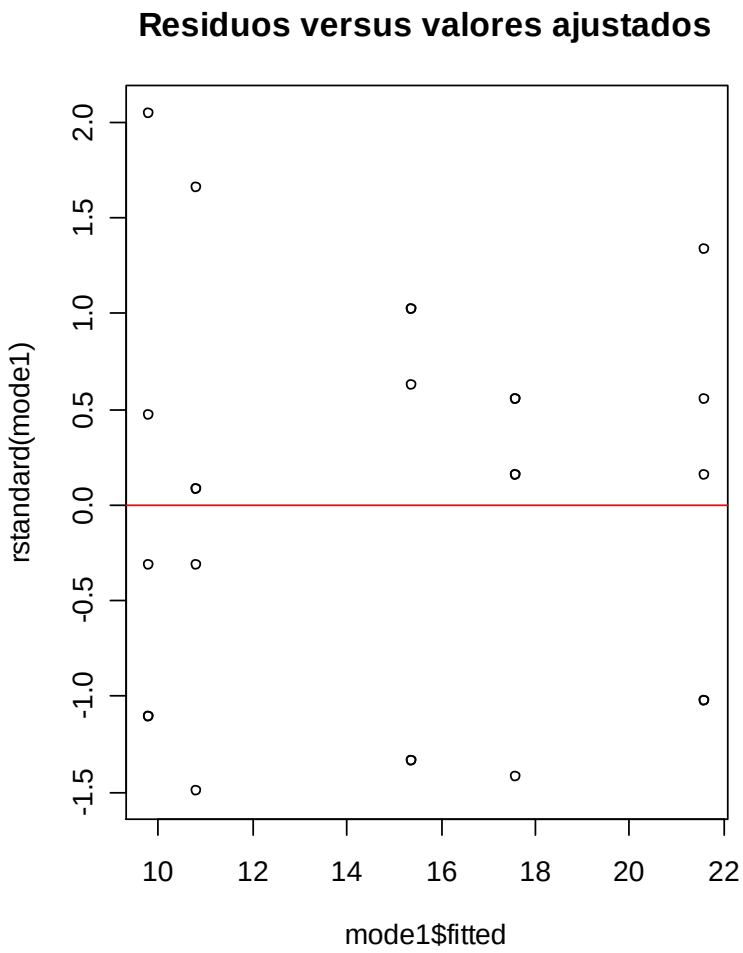


Gráfica de probabilidad normal de los residuos

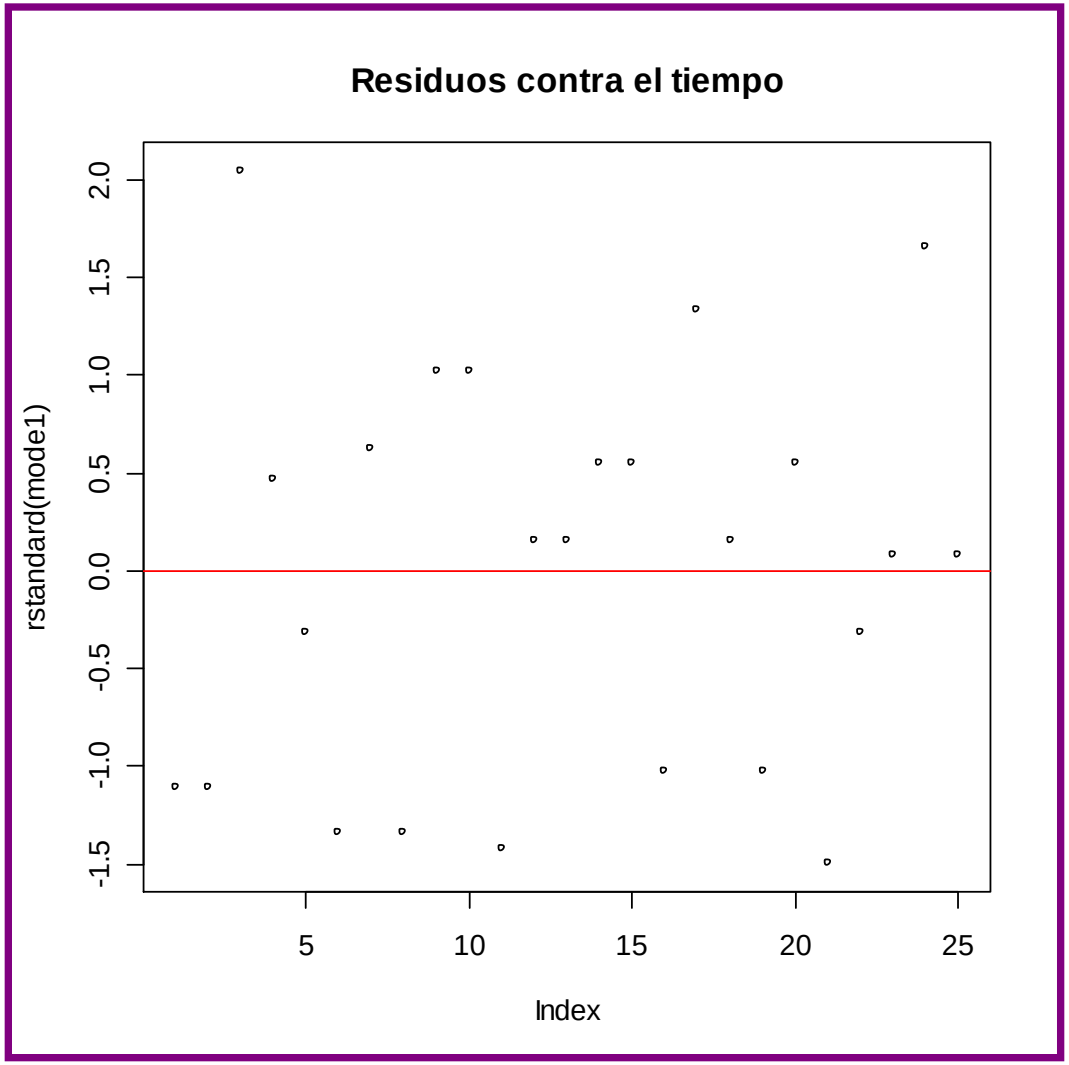


Diagnosis del modelo: homocedasticidad





Diagnosis del modelo: independencia



Comparaciones entre medias

Una vez obtenidas diferencias significativas entre los tratamientos, conviene estudiar por qué se rechaza la igualdad entre medias, comparando todos los pares de medias, porque puede ser que se rechace la igualdad de medias porque haya un par de medias diferentes entre sí. Se considera, entonces, los siguientes contrastes:

$$H_0 \equiv \mu_i = \mu_j, \qquad i \neq j$$

$$H_0 \equiv \mu_i \neq \mu_j, \qquad i \neq j$$

Diferencia significativa mínima

LSD de Fisher (Least significant difference)

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-a}$$

Bajo la hipótesis nula

$$LSD_\alpha = t_{N-a,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \rightarrow$$

Diferencia significativa mínima

Si $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > LSD_\alpha \Longrightarrow$ Se rechaza que $\mu_i = \mu_j$ a nivel α.

Si $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| < LSD_\alpha \Longrightarrow$ Se acepta que $\mu_i = \mu_j$ a nivel α.

Método de Bonferroni

En este criterio se rechaza $\mu_i = \mu_j$ (i ≠ j) si

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > t_{N-a,\frac{\alpha}{2p}}\,\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

donde p es el número de comparaciones que se pueden obtener

$$1 \leq p \leq \binom{a}{2}$$

Se puede aproximar $t_{N-a,\frac{\alpha}{2p}}$ por una normal:

$$t_{v,\alpha} = z_\alpha + \frac{1}{4v}\left(z_\alpha^3 - z_\alpha\right) \text{ siendo } z_\alpha \sim N(0,1)$$

Distribución de recorrido estudentizada

$$Z_1, \ldots, Z_a \sim N(0,1)$$
$$U \sim \chi^2_m$$

→ Independientes

$$Q = \max_{i \neq j} \frac{|Z_i - Z_j|}{\sqrt{\frac{U}{m}}} = \frac{Z_{(a)} - Z_{(1)}}{\sqrt{\frac{U}{m}}} \sim q_{a,m}$$

se distribuye con una distribución de recorrido estudentizado de parámetros *a* y *m*.

Método de Tuckey

Se requiere que $n_i = n$, $i = 1, \ldots, a$.

Si esto no se cumple, entonces se toma $n = min_i\{ni\}$

$$\text{Si } |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > q_{a,N-a;\alpha}\hat{\sigma}\sqrt{\frac{1}{n}} \Longrightarrow \text{Se rechaza que } \mu_i = \mu_j \text{ a nivel } \alpha.$$

$$\text{Si } |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| < q_{a,N-a;\alpha}\hat{\sigma}\sqrt{\frac{1}{n}} \Longrightarrow \text{Se acepta que } \mu_i = \mu_j \text{ a nivel } \alpha.$$

```
pairwise.t.test(resistencia,porcentaje,p.adjust.method='none')

pairwise.t.test(resistencia,porcentaje,p.adjust.method='bonferroni')

TukeyHSD(aov(resistencia~porcentaje))
```