TEMA 1
Diseño de experimentos: modelo unifactorial

José R. Berrendero
Departamento de Matemáticas
Universidad Autónoma de Madrid

Análisis de Datos - Grado en Biología

Esquema del tema

- Modelos estadísticos.
- El modelo unifactorial.
- El contraste de igualdad de medias y la tabla de análisis de la varianza (ANOVA).
- Intervalos y contrastes para los parámetros del modelo.
- Comparaciones múltiples.
- Diagnóstico del modelo.

Estructura de un modelo estadístico

Un modelo estadístico se utiliza para describir la relación entre una **variable respuesta** y un conjunto de **variables explicativas**.

La estructura habitual de un modelo es:

$$\textsc{Respuesta} = \textsc{Parte explicada} + \textsc{Error aleatorio}$$

La parte explicada es una función que describe cómo es el efecto de las variables explicativas sobre la respuesta media.

El término de error es una variable aleatoria que recoge el efecto de otras muchas variables que pueden influir en la respuesta pero no han sido tenidas en cuenta explícitamente en el modelo.

Un modelo muy sencillo (sin variables explicativas)

En un estudio para comparar la eficacia del uso de fertilizantes, se utilizan los fertilizantes en 30 parcelas y posteriormente se registra el peso en toneladas de la cosecha resultante en cada parcela. Los datos son:

6,27	5,36	6,39	4,85	5,99	7,14	5,08	4,07	4,35	4,95
3,07	3,29	4,04	4,19	3,41	3,75	4,87	3,94	6,28	3,15
4,04	3,79	4,56	4,55	4,55	4,53	3,53	3,71	7,00	4,61

En este caso, la variable respuesta es el peso de la cosecha y no hay variables explicativas.

Modelo:

$$\text{Cosecha} = \text{Cosecha media} + \text{Error aleatorio}$$

Un modelo muy sencillo (sin variables explicativas)

Modelo:

$$Y_i = \mu + u_i, \quad i = 1, \ldots, n,$$

donde μ es un parámetro que representa la cosecha media (la misma para todas las respuestas) y u_i es una v.a. que recoge el efecto de otras variables que hacen que las cosechas no sean iguales a la media.

Hipótesis habituales sobre las variables u_i:

- Tienen media 0 y varianza σ^2.
- Tienen distribución normal.
- Son independientes.

Los modelos que vamos a ver en el curso corresponden a diferentes formas de expresar μ de manera que se pueda cuantificar el efecto de las variables explicativas.

Estimación

El modelo es equivalente a:

$$Y_1, \ldots, Y_n \text{ son v.a. independientes con distribución } \mathrm{N}(\mu, \sigma)$$

Los estimadores usuales de los dos parámetros de este modelo son:

$$\begin{aligned} \hat{\mu} &= \bar{Y}, \\ \hat{\sigma}^2 &= S^2 = \frac{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}{n-1} = \frac{n}{n-1}\left(\frac{\sum_{i=1}^n Y_i^2}{n} - \bar{Y}^2\right) \end{aligned}$$

Varianza y grados de libertad

La dispersión de un conjunto de números $Y_1, \ldots, Y_n$ depende de las desviaciones de los datos a la media $(Y_1 - \bar{Y}), \ldots, (Y_n - \bar{Y})$.

Estas desviaciones corresponden a la parte de la respuesta no explicada por la media.

Se cumple que la suma de desviaciones es igual a cero:

$$(Y_1 - \bar{Y}) + \cdots + (Y_n - \bar{Y}) = 0$$

La dispersión es grande si lo es la **suma de cuadrados** de desviaciones:

$$(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2$$

Varianza y grados de libertad

La suma de cuadrados no se puede utilizar directamente para medir la variabilidad porque aumenta al incrementarse el número de datos, incluso cuando la variabilidad no aumenta.

Como las desviaciones suman 0, si nos dan $n-1$ de ellas podemos calcular la restante. Sólo tenemos $n-1$ desviaciones independientes para medir la variabilidad.

Llamamos **grados de libertad** (gl) al número de sumandos independientes que aparecen en una suma de cuadrados.

Para corregir por n dividimos la suma de cuadrados por sus gl. Esta operación nos da la **(cuasi)varianza**:

$$S^2 = \frac{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}{n-1}$$

El modelo unifactorial

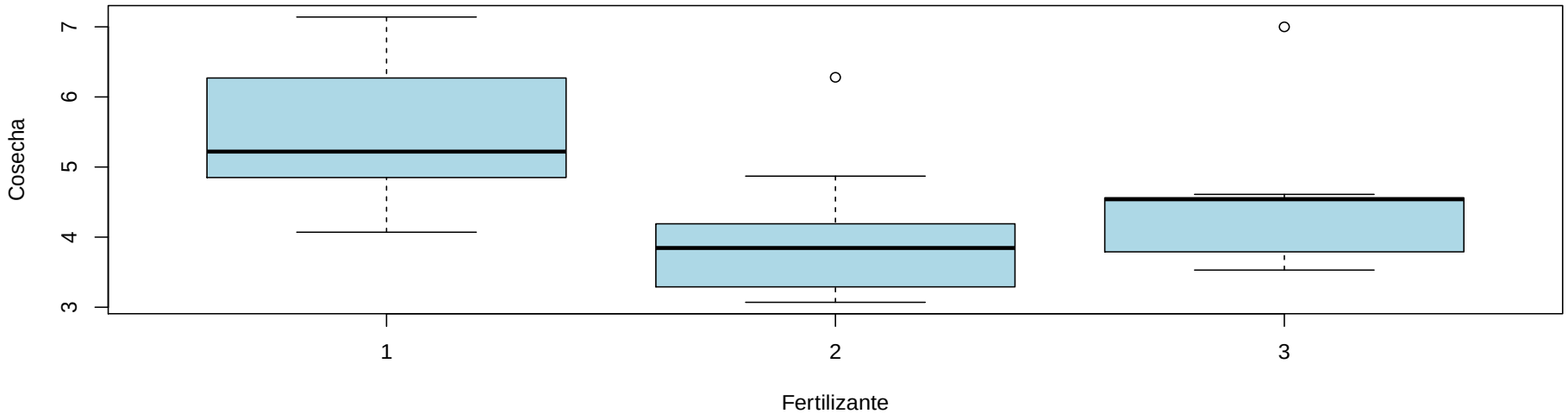
En un estudio para comparar la eficacia de tres fertilizantes se utiliza cada uno de ellos en 10 parcelas (asignando aleatoriamente cada parcela a uno de los tres fertilizantes) y posteriormente se registra el peso en toneladas de la cosecha resultante en cada parcela. Los datos son:

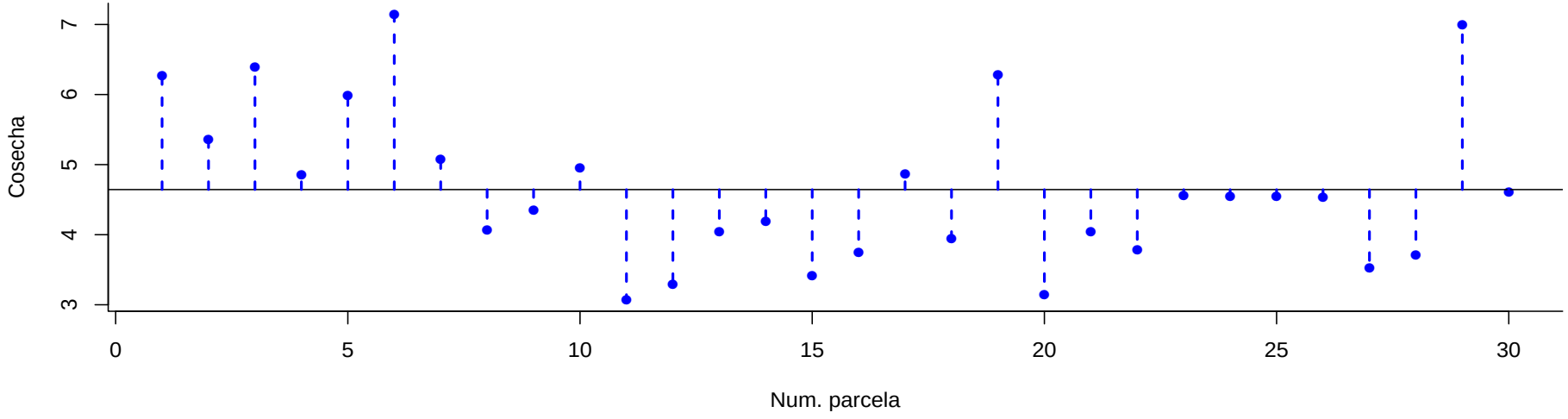
Fert. 1	6,27	5,36	6,39	4,85	5,99	7,14	5,08	4,07	4,35	4,95
Fert. 2	3,07	3,29	4,04	4,19	3,41	3,75	4,87	3,94	6,28	3,15
Fert. 3	4,04	3,79	4,56	4,55	4,55	4,53	3,53	3,71	7,00	4,61

Una variable explicativa cualitativa se llama **factor**. Los valores que toma se llaman **niveles**. En este modelo los niveles son los distintos **tratamientos** que aplicamos a las **unidades experimentales**.

En el ejemplo tenemos un factor (el tipo de fertilizante) que se presenta en tres niveles o tratamientos, que se aplican a las unidades experimentales (las parcelas).

Descripción de los datos





Principios básicos del diseño: réplicas

- Replicar un experimento da una idea de la variabilidad de las respuestas y, por lo tanto, permite evaluar la precisión de los estimadores.
- Cuanto mayor es el número de réplicas, mayor es la precisión de los estimadores y, por lo tanto, es más probable detectar diferencias significativas entre los tratamientos.
- Cuando el número de réplicas es el mismo para todos los tratamientos se dice que el diseño es **equilibrado**.
- En el ejemplo tenemos un diseño equilibrado con 10 réplicas para cada nivel del factor.

Principios básicos del diseño: aleatorización

- Para evitar sesgos, es importante asignar aleatoriamente las unidades experimentales a los tratamientos. El objetivo es que variables no controladas afecten por igual a todos los tratamientos.

- Cuando una unidad experimental puede recibir cada tratamiento con la misma probabilidad se dice que el diseño es **completamente aleatorizado**.

- La situación ideal es que las unidades experimentales sean idénticas, o lo más homogéneas que sea posible. Cuando existe un factor conocido de heterogeneidad que afecta a la respuesta, resulta conveniente repartir las unidades experimentales en **bloques** homogéneos.

- El diseño por **bloques aleatorizados completos**, consiste en asignar aleatoriamente las unidades a los tratamientos dentro de cada bloque de forma que cada bloque contenga todos los tratamientos.

Notación y principales medidas descriptivas

Disponemos de respuestas correspondientes a I niveles del factor, n_i es el tamaño muestral del grupo i y $n = n_1 + \cdots + n_I$ es el número total de respuestas.

Muestra	Respuestas				Medias	Desv. típicas
1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$\bar{Y}_{1.}$	S_1
1	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$\bar{Y}_{2.}$	S_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
I	Y_{I1}	Y_{I2}	$\cdots$	Y_{In_I}	$\bar{Y}_{I.}$	S_I

En el ejemplo: $I = 3$, $n_i = 10$, $n = 30$.

Muestra	n_i	$\bar{Y}_{i.}$	S_i
1	10	5,445	0,976
2	10	3,999	0,972
3	10	4,487	0,975

Formulación del modelo unifactorial

Si Y_{ij} representa la respuesta j para el nivel i,

$$Y_{ij} = \mu_i + u_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

- μ_i es el nivel medio de la respuesta para el nivel i del factor.
- u_{ij} es la variable de error que recoge el resto de variables que influyen en la respuesta. Estas variables son independientes y tienen distribución normal con media 0 y desviación típica σ.
- **Homocedasticidad**: La desviación típica es la misma para todos los niveles del factor.

Otra forma equivalente de escribir lo mismo:
Para $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, las variables Y_{ij} son independientes y, además,

$$Y_{ij} \equiv \mathrm{N}(\mu_i; \sigma)$$

Modelo unifactorial: una parametrización equivalente

Si definimos

$$\mu = \frac{\mu_1 + \cdots + \mu_I}{I}$$

y $\alpha_i = \mu_i - \mu$, entonces

$$Y_{ij} = \mu + \alpha_i + u_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

- μ es la media global de la variable respuesta.
- α_i representa el efecto adicional sobre la respuesta debido al nivel i del factor.
- Se verifica $\alpha_1 + \cdots + \alpha_I = 0$. Los efectos marginales se compensan unos con otros.
- Las variables de error u_{ij} se interpretan y distribuyen igual que en la formulación anterior del modelo.

Estimadores de los parámetros

Aplicando el método de máxima verosimilitud es posible encontrar los siguientes estimadores para los parámetros:

Para la primera forma de escribir el modelo:

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

¿Cuáles son las desviaciones típicas de estos estimadores?

Para la segunda forma de escribir el modelo:

- $\hat{\mu} = \bar{Y}_{..}$, donde $\bar{Y}_{..}$ es la **media global**

$$\bar{Y}_{..} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}}{n} = \frac{n_1 \bar{Y}_{1.} + \cdots + n_I \bar{Y}_{I.}}{n}.$$

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

Calcula estos estimadores para los datos de los fertilizantes.

El contraste de igualdad de medias

El objetivo principal es estudiar si hay diferencias significativas entre las medias de los I grupos.

Hipótesis nula (H_0): las medias de los I grupos son iguales ($\mu_1 = \cdots = \mu_I$).

Hipótesis alternativa (H_1): no todas las medias son iguales.

¿Por qué no comparar las tres medias de dos en dos utilizando el contraste que ya hemos estudiado?

La idea clave es rechazar H_0 cuando la variabilidad entre los grupos sea grande en relación a la variabilidad interna de los grupos.

Descomposición de la variabilidad

¿Qué factores pueden causar la variabilidad observada en las cosechas de las 30 parcelas del experimento?

Muchos factores posibles: diferencias entre los suelos de las parcelas, diferencias en humedad, etc.

El fertilizante es el factor en el que estamos interesados, así que vamos a dividir la variabilidad total en una parte debida al fertilizante y otra parte debida al resto de factores que no se controlan en el experimento.

Si la variabilidad debida al fertilizante es grande en relación a la debida al resto de factores, rechazamos H_0.

Variabilidad total

La variabilidad total se mide mediante la **suma de cuadrados total (SCT)**, que considera las desviaciones de las respuestas alrededor de la media global:

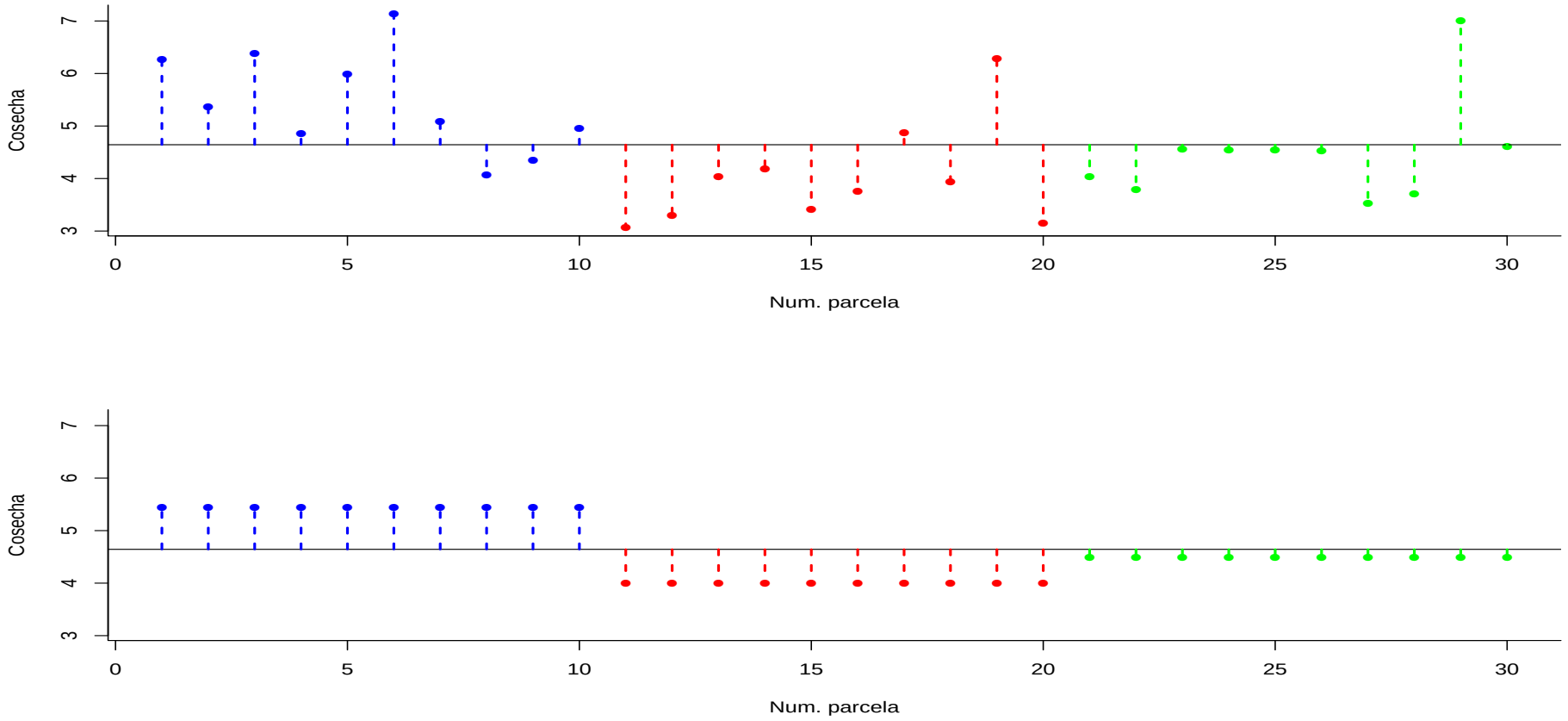
$$\text{SCT} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2.$$

¿Cuántos gl tiene esta suma de cuadrados?

La variabilidad total sólo depende de las respuestas observadas, no depende del modelo considerado.

Variabilidad explicada

Para medir la variabilidad debida al fertilizante se sustituye cada respuesta Y_{ij} por la respuesta media estimada a partir del modelo, $\bar{Y}_{i.}$, y se consideran las desviaciones $\bar{Y}_{i.} - \bar{Y}_{..}$



Variabilidad explicada

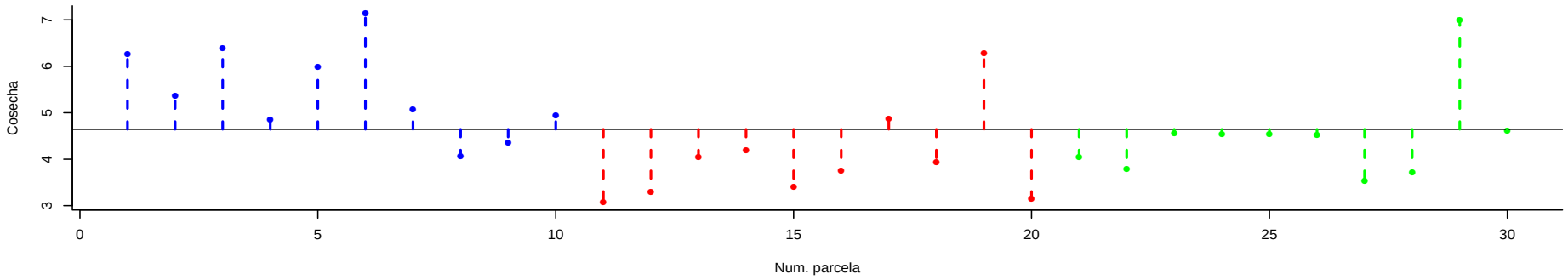
La variabilidad entre los grupos o variabilidad explicada se mide mediante la **suma de cuadrados explicada (SCE)**:

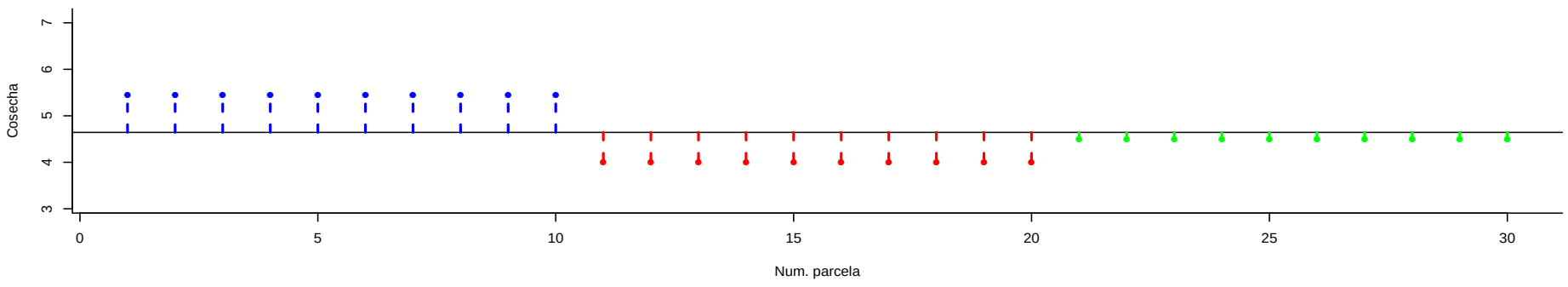
$$\text{SCE} = \sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

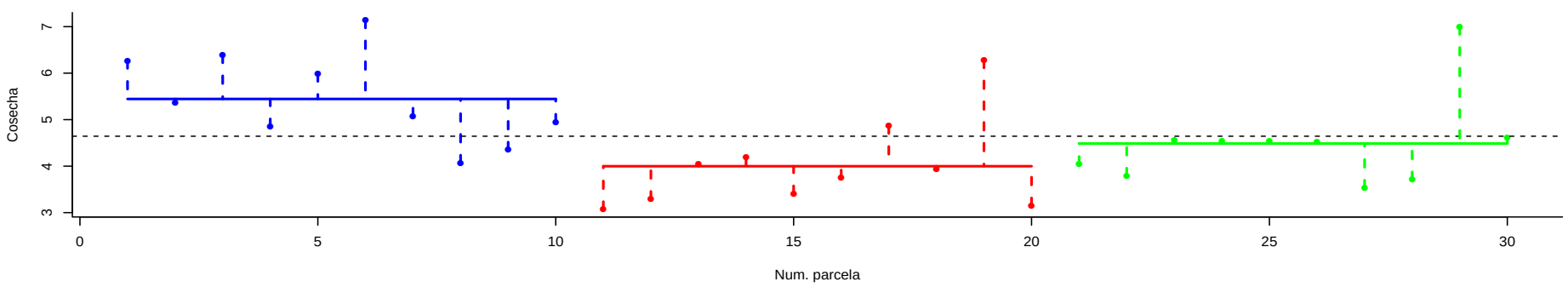
¿Cuántos gl tiene SCE?

¿Cuánto vale SCE en el ejemplo de los fertilizantes?

Variabilidad residual







Variabilidad residual

La variabilidad debida a factores que no son el fertilizante es la que explica las desviaciones $Y_{ij} - \bar{Y}_{i.}$

La variabilidad residual o no explicada se mide mediante la **suma de cuadrados residual (SCR)**:

$$\text{SCR} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = (n_1 - 1)S_1^2 + \ldots + (n_I - 1)S_I^2,$$

SCR es una combinación lineal de las varianzas de cada grupo. Los grupos de mayor tamaño reciben más peso.

¿Cuántos gl tiene?

¿Cuánto vale SCR en el ejemplo de los fertilizantes?

Descomposición de la variabilidad

Siempre se cumple:

$$Y_{ij} - \bar{Y}_{..} = (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.}).$$

Si elevamos al cuadrado y sumamos todos estos términos:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{I}\sum_{i=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2.$$

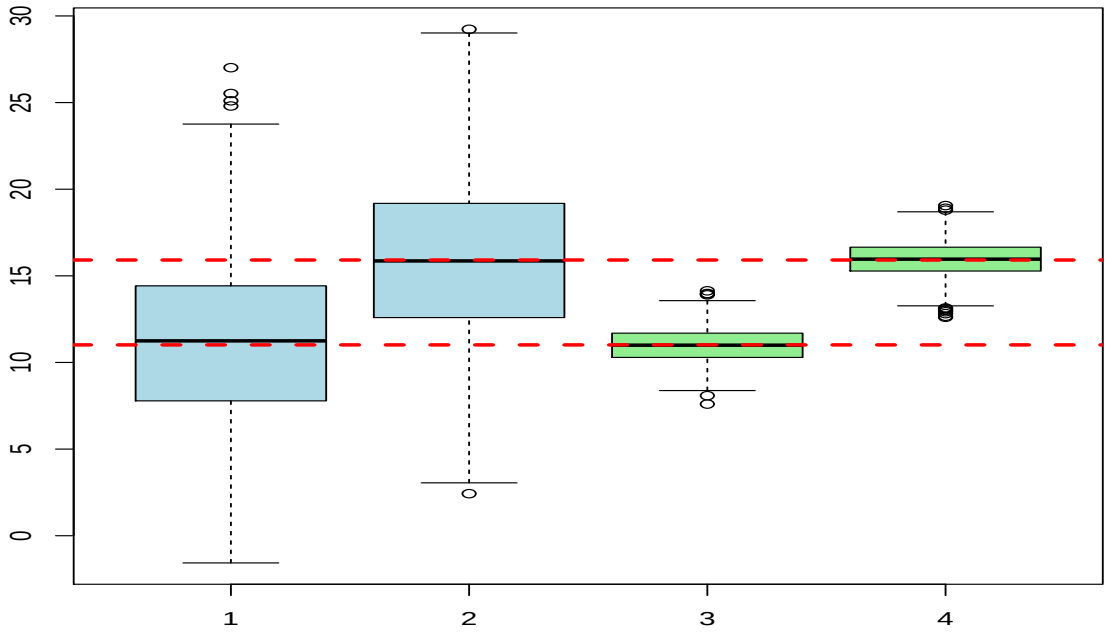
Por lo tanto:

$$\text{SCT} = \text{SCE} + \text{SCR}$$

$$\textsc{Var. Total} = \textsc{Var. Explicada} + \textsc{Var. Residual}$$

Diferencia entre grupos y variabilidad dentro de los grupos

Para decidir si la diferencia entre grupos es suficientemente grande hay que tener en cuenta la variabilidad existente dentro de los grupos.



Es más probable que la diferencia entre grupos sea significativa si la variabilidad dentro de los grupos es pequeña.

Diferencia entre grupos y variabilidad dentro de los grupos

Réplica	$T1$	$T2$	$T3$	$T1$	$T2$	$T3$
1	20	22	24	45	8	15
2	19	22	24	0	30	44
3	20	22	23	10	38	2
4	21	22	25	25	12	35
Medias	20	22	24	20	22	24

¿En qué situación parecen más significativas las diferencias entre las medias?

Dos casos extremos

Réplica	T1	T2	T3	T4
1	16	15	16	17
2	15	17	16	16
3	17	16	17	15
4	16	16	15	16
Media	16	16	16	16

Réplica	T1	T2	T3	T4
1	19.5	15	16.5	13
2	19.5	15	16.5	13
3	19.5	15	16.5	13
4	19.5	15	16.5	13
Media	19.5	15	16.5	13

Cuadrados medios y estadístico F

Los cuadrados medios se obtienen dividiendo las sumas de cuadrados por sus grados de libertad.

Suma de cuadrados	Cuadrados medios	Valores en el ejemplo
SCE	$\text{SCE}/(I-1)$	
SCR	$\text{SCR}/(n-I)$	
SCT	$\text{SCT}/(n-1)$	

Para contrastar $H_0 : \mu_1 = \cdots = \mu_I$ se compara $\text{SCE}/(I-1)$ con $\text{SCR}/(n-I)$ mediante el cociente:

$$F = \frac{\text{SCE}/(I-1)}{\text{SCR}/(n-I)}$$

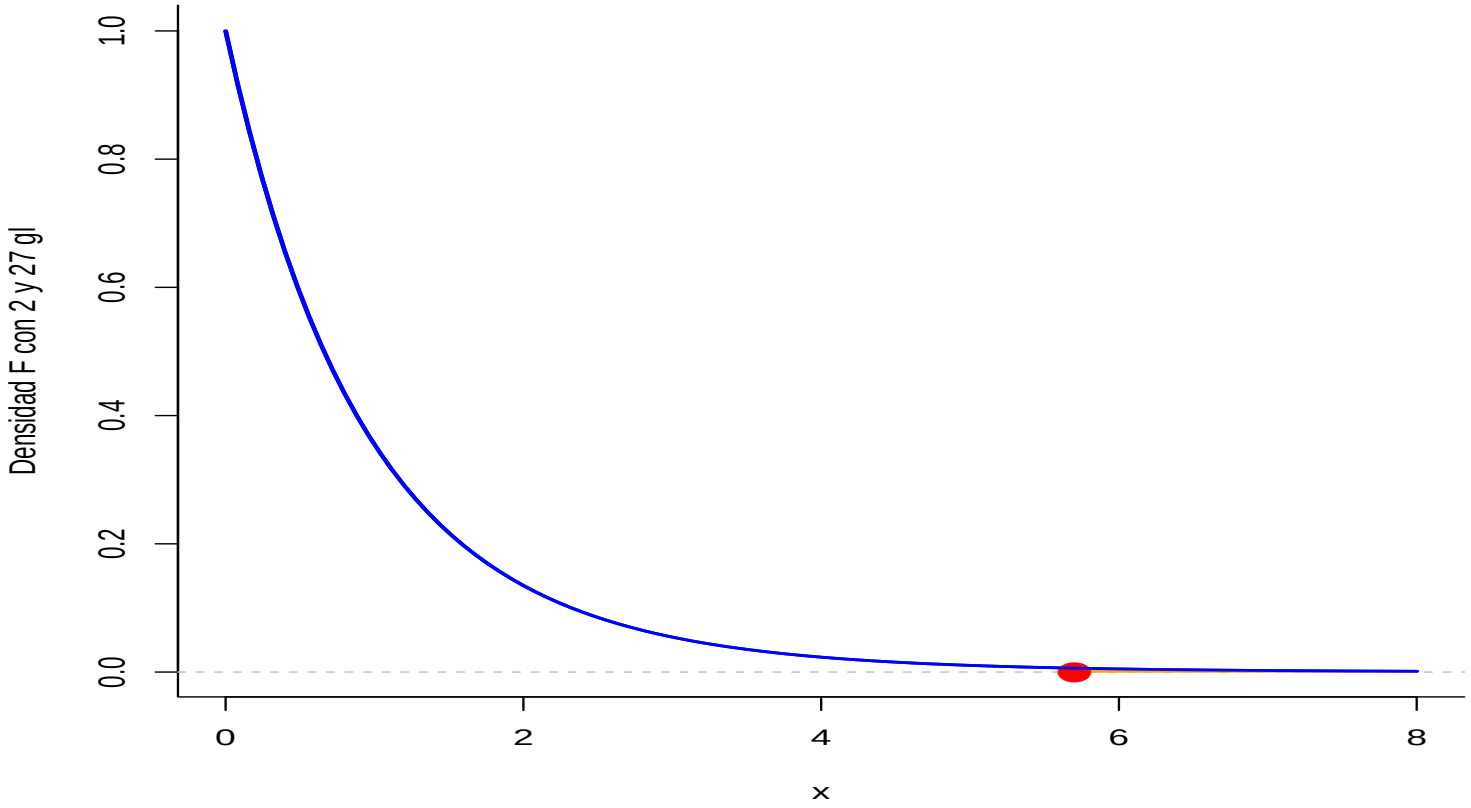
En el ejemplo de los fertilizantes, ¿cuánto vale F?

Se rechaza H_0 si F es "suficientemente grande".

La distribución F

Puede demostrarse que, cuando $H_0 : \mu_1 = \cdots = \mu_I$ es cierta, los valores de F se distribuyen de acuerdo con una distribución F con $I - 1$ y $n - I$ grados de libertad:

$$\text{Bajo } H_0, \quad F \equiv F_{I-1,n-I}$$



Tablas de la distribución F

n_1 y n_2: grados de libertad del numerador y del denominador respectivamente

Ejemplo: para $n_1 = 5$, $n_2 = 10$ y $\alpha = 0.01$, $F_{5,10;0.01} = 5.636$, significa que $P(F_{5,10} > 5.636) = 0.01$.

	n_1															
n_2	1	2	3	4	5	6	7	8	9	10	12	15	16	18	20	24
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6170	6192	6209	6235
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.44	99.45	99.46
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.83	26.75	26.69	26.60
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.15	14.08	14.02	13.93
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.051	9.888	9.722	9.680	9.610	9.553	9.466
6	13.75	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559	7.519	7.451	7.396	7.313
7	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314	6.275	6.209	6.155	6.074
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.477	5.412	5.359	5.279
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.924	4.860	4.808	4.729
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558	4.520	4.457	4.405	4.327
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.213	4.150	4.099	4.021
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010	3.972	3.909	3.858	3.780
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.815	3.778	3.716	3.665	3.587
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.656	3.619	3.556	3.505	3.427
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.485	3.423	3.372	3.294
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.409	3.372	3.310	3.259	3.181
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.312	3.275	3.212	3.162	3.084
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.190	3.128	3.077	2.999
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.116	3.054	3.003	2.925
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	3.051	2.989	2.938	2.859
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.030	2.993	2.931	2.880	2.801
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.941	2.879	2.827	2.749
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.931	2.894	2.832	2.781	2.702
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.852	2.789	2.738	2.659
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.850	2.813	2.751	2.699	2.620
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.815	2.778	2.715	2.664	2.585
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.783	2.746	2.683	2.632	2.552
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.753	2.716	2.653	2.602	2.522
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.726	2.689	2.626	2.574	2.495
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.700	2.663	2.600	2.549	2.469

La región crítica

El valor concreto a partir del cual se rechaza H_0 depende del grado de seguridad que queramos tener al rechazar.

Fijamos α, el nivel de significación del contraste, es decir, la probabilidad de error al rechazar (tipo 1).

Dado el valor de α se rechaza $H_0 : \mu_1 = \cdots = \mu_I$ en la región crítica:

$$R = \{F > F_{I-1,n-I;\alpha}\}$$

¿Qué decisión se adopta en el ejemplo de los fertilizantes si $\alpha = 0{,}05$? ¿Y si $\alpha = 0{,}01$?

¿Qué podemos decir sobre el p-valor del contraste con la tablas disponibles?

Algunos valores de referencia

Sea X es una variable cuyos valores se distribuyen de acuerdo con una distribución F_{n_1,n_2},

- Puede demostrarse que el valor esperado de X es $n_2/(n_2-2)$ (siempre que $n_2 > 2$).
- Cuando tenemos un número grande de datos n repartidos en un número pequeño de grupos I, ¿cuál es aproximadamente el valor esperado del estadístico F si no hay diferencias significativas entre los grupos?
- En el ejemplo, ¿cuál sería el valor esperado de F bajo H_0? ¿Difiere mucho este valor esperado del calculado a partir de los datos?
- Encuentra un valor c tal que una variable F con 2 y 10 gl es mayor que c con probabilidad 0.05. Este valor se denota $F_{2,10;0,05}$.
- En general, $F_{n_1,n_2;\alpha}$ es el valor tal que una variable F con n_1 y n_2 gl es mayor que $F_{n_1,n_2;\alpha}$ con probabilidad α.

Tabla ANOVA

Toda la información sobre las sumas de cuadrados y el contraste de igualdad de medias se suele ordenar en forma de tabla:

Fuente de variación	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada	$\sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$I-1$	$\frac{\sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2}{I-1}$	F
Residual	$\sum_{i=1}^{I}(n_i-1)S_i^2$	$n-I$	$\frac{\sum_{i=1}^{I}(n_i-1)S_i^2}{n-I}$	

Con los datos del ejemplo:

Fuente de variación	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada				
Residual				

Resultados con SPSS

Descriptivos

cosecha

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	10	5,4450	,97598	,30863	4,7468	6,1432	4,07	7,14
2	10	3,9990	,97175	,30729	3,3039	4,6941	3,07	6,28
3	10	4,4870	,97471	,30823	3,7897	5,1843	3,53	7,00
Total	30	4,6437	1,12104	,20467	4,2251	5,0623	3,07	7,14

ANOVA

cosecha

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	10,823	2	5,411	5,702	,009
Intra-grupos	25,622	27	,949		
Total	36,445	29			

Coeficiente de determinación

El **coeficiente de determinación** es la proporción de la variabilidad total explicada por los factores incluidos en un modelo.

En el modelo unifactorial, es la proporción de la variabilidad total que se puede atribuir al factor.

$$R^2 = \frac{\text{SCE}}{\text{SCT}}$$

La definición implica que siempre se cumple $0 \leq R^2 \leq 1$.

Calcula el coeficiente de determinación en el ejemplo de los fertilizantes.

Residuos y valores ajustados

Llamamos **valor ajustado o pronosticado** de Y_{ij} a un estimador de su valor esperado.
Los valores ajustados se obtienen sustituyendo en el modelo el término de error por 0 y todos los parámetros por sus estimadores:

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

Llamamos **residuo** a la diferencia entre la respuesta que se observa realmente y su valor ajustado:

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.}$$

Observaciones:

- Siempre se verifica que $Y_{ij} = \hat{Y}_{ij} + e_{ij}$.
- ¿Cuánto vale la media de los residuos?
- ¿Qué relación hay entre SCR y los residuos?

Estimación de la varianza

Un estimador insesgado de la varianza de los errores σ^2 viene dado por la SCR corregida por sus grados de libertad.

Este estimador se llama **varianza residual**:

$$S_R^2 = \frac{\text{SCR}}{n-I} = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij}^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$$

Distribución de S_R^2:

$$\frac{\text{SCR}}{\sigma^2} = \frac{(n-I)S_R^2}{\sigma^2} = \frac{\sum_{i=1}^{I}(n_i-1)S_i^2}{\sigma^2} \equiv \chi^2_{n-I}$$

Tabla de la distribución χ^2

π / ϕ	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	π / ϕ
1	3.93E-05	1.57E-04	9.82E-04	3.93E-03	1.58E-02	0.102	0.455	1.323	2.71	3.84	5.02	6.63	7.88	1
2	1.00E-02	2.01E-02	5.06E-02	0.103	0.211	0.575	1.386	2.77	4.61	5.99	7.38	9.21	10.60	2
3	7.17E-02	0.115	0.216	0.352	0.584	1.213	2.37	4.11	6.25	7.81	9.35	11.34	12.84	3
4	0.207	0.297	0.484	0.711	1.064	1.923	3.36	5.39	7.78	9.49	11.14	13.28	14.86	4
5	0.412	0.554	0.831	1.145	1.610	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	5
6	0.676	0.872	1.237	1.635	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	6
7	0.989	1.239	1.690	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.3	7
8	1.344	1.647	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.1	22.0	8
9	1.735	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.7	23.6	9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.5	23.2	25.2	10
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.9	24.7	26.8	11
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.0	23.3	26.2	28.3	12
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.4	24.7	27.7	29.8	13
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.1	23.7	26.1	29.1	31.3	14
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.3	25.0	27.5	30.6	32.8	15
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.5	26.3	28.8	32.0	34.3	16
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.5	24.8	27.6	30.2	33.4	35.7	17
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.6	26.0	28.9	31.5	34.8	37.2	18
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.7	27.2	30.1	32.9	36.2	38.6	19
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.8	28.4	31.4	34.2	37.6	40.0	20
21	8.03	8.90	10.28	11.59	13.24	16.34	20.3	24.9	29.6	32.7	35.5	38.9	41.4	21
22	8.64	9.54	10.98	12.34	14.04	17.24	21.3	26.0	30.8	33.9	36.8	40.3	42.8	22
23	9.26	10.20	11.69	13.09	14.85	18.14	22.3	27.1	32.0	35.2	38.1	41.6	44.2	23
24	9.89	10.86	12.40	13.85	15.66	19.04	23.3	28.2	33.2	36.4	39.4	43.0	45.6	24
25	10.52	11.52	13.12	14.61	16.47	19.94	24.3	29.3	34.4	37.7	40.6	44.3	46.9	25
26	11.16	12.20	13.84	15.38	17.29	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	26
27	11.81	12.88	14.57	16.15	18.11	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	27
28	12.46	13.56	15.31	16.93	18.94	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	28
29	13.12	14.26	16.05	17.71	19.77	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	29
30	13.79	14.95	16.79	18.49	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	30
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8	40
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5	50
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0	60
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100.4	104.2	70
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	101.9	106.6	112.3	116.3	80
90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	107.6	113.1	118.1	124.1	128.3	90
100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109.1	118.5	124.3	129.6	135.8	140.2	100

Intervalos de confianza para μ_i y σ^2

Un IC para σ^2 (nivel $1-\alpha$) es:

$$\text{IC}_{1-\alpha}(\sigma^2) = \left[\frac{\text{SCR}}{\chi^2_{n-I,\alpha/2}}, \frac{\text{SCR}}{\chi^2_{n-I,1-\alpha/2}}\right] = \left[\frac{(n-I)S_R^2}{\chi^2_{n-I,\alpha/2}}, \frac{(n-I)S_R^2}{\chi^2_{n-I,1-\alpha/2}}\right]$$

El error típico de $\hat{\mu}_i = \bar{Y}_{i.}$ es $S_R/\sqrt{n_i}$.

Un IC para μ_i (nivel $1-\alpha$) es:

$$\text{IC}_{1-\alpha}(\mu) = \left[\bar{Y}_{i.} \mp t_{n-I,\alpha/2}\frac{S_R}{\sqrt{n_i}}\right]$$

Comparando con los intervalos cuando sólo tenemos una muestra:

- Cambia el estimador de la desviación típica. Combinamos todas las muestras para estimar σ. Esto tiene sentido debido a la hipótesis de homocedasticidad.
- Por ello también cambian los gl.

Cuestiones

Responde utilizando los datos de los fertilizantes:

- ¿Cuánto valen los errores típicos de los estimadores de los parámetros μ_i?
- Calcula un intervalo de confianza de nivel 0,95 para σ^2.
- Calcula intervalos de confianza de nivel 0,95 para μ_1, μ_2 y μ_3.

Contrastes para μ_i

Bilateral: $H_0 : \mu_i = \mu_i^*$ frente a $H_1 : \mu_i \neq \mu_i^*$

$$R = \left\{ \frac{|\bar{Y}_{i.} - \mu_i^*|}{S_R/\sqrt{n_i}} > t_{n-I,\alpha/2} \right\}$$

Unilateral: $H_0 : \mu_i \leq \mu_i^*$ frente a $H_1 : \mu_i > \mu_i^*$

$$R = \left\{ \frac{\bar{Y}_{i.} - \mu_i^*}{S_R/\sqrt{n_i}} > t_{n-I,\alpha} \right\}$$

¿Cuál es la región crítica para contrastar $H_0 : \mu_i \geq \mu_i^*$ frente a $H_1 : \mu_i < \mu_i^*$?

Intervalo de confianza para $\mu_i - \mu_j$

¿Existen diferencias significativas entre los niveles i y j del factor?

IC para $\mu_i - \mu_j$:

$$\text{IC}_{1-\alpha}(\mu_i - \mu_j) = \left[(\bar{Y}_{i.} - \bar{Y}_{j.}) \mp t_{n-I,\alpha/2} S_R \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right]$$

Para contrastar $H_0 : \mu_i = \mu_j$ frente a $H_1 : \mu_i \neq \mu_j$ a nivel α, una posibilidad es rechazar H_0 cuando el intervalo anterior no contiene al 0.

Cuestiones

Responde utilizando los datos de los fertilizantes:

- ¿Podemos afirmar a nivel $\alpha = 0{,}05$ que la cosecha media en terrenos en los que se use el fertilizante 2 es superior a 3 toneladas?

- ¿Podemos afirmar a nivel $\alpha = 0{,}01$ que la cosecha esperada es diferente en los terrenos en los que se usa el fertilizante 2 que en aquellos en los que se usa el fertilizante 3?

- ¿Podemos afirmar a nivel $\alpha = 0{,}01$ que es más efectivo el fertilizante 1 que el fertilizante 2?

- Calcula un intervalo de confianza de nivel 95 % para $\mu_2 - \mu_3$.

Planteamiento del problema de comparaciones múltiples

Este problema aparece siempre que se llevan a cabo m inferencias (m contrastes o m intervalos de confianza).

Se desea que el nivel de significación global de los m contrastes sea inferior a un valor prefijado α_T. Es decir, α_T es la probabilidad de rechazar incorrectamente **alguna** de las m hipótesis nulas.

Para los IC se desea que el nivel de confianza global sea $1 - \alpha_T$. Es decir, $1 - \alpha_T$ es el nivel de confianza que corresponde a que **los m intervalos contengan a los m parámetros**.

Problema: ¿Cómo hay que llevar a cabo cada contraste o IC individual para conseguir el objetivo global?

Método de Bonferroni

- Calcular $\alpha = \alpha_T/m$.
- Para que el nivel de significación global sea α_T llevar a cabo cada contraste individual a nivel α.
- Para que el nivel de confianza global sea $1 - \alpha_T$ calcular cada IC individual a nivel $1 - \alpha$.

Por ejemplo si queremos comparar todos los pares de medias, tenemos que contrastar $m = I(I-1)/2$ hipótesis de la forma $H_0 : \mu_i = \mu_j$ frente a $H_1 : \mu_i \neq \mu_j$.

En el ejemplo de los fertilizantes, si queremos que el nivel global sea $\alpha_T = 0{,}05$, tenemos que hacer cada contraste individual a nivel $\alpha = 0{,}05/3 \approx 0{,}017$, puesto que $m = 3$ en este caso.

Método de Bonferroni: resultado SPSS

Comparaciones múltiples

cosecha

Bonferroni

(I) fertilizante	(J) fertilizante	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
1	2	1,44600*	,43565	,008	,3340	2,5580
	3	,95800	,43565	,110	-,1540	2,0700
2	1	-1,44600*	,43565	,008	-2,5580	-,3340
	3	-,48800	,43565	,818	-1,6000	,6240
3	1	-,95800	,43565	,110	-2,0700	,1540
	2	,48800	,43565	,818	-,6240	1,6000

*. La diferencia de medias es significativa al nivel 0.05.

Diagnóstico del modelo

Las técnicas de inferencia que hemos estudiado se basan en las siguientes hipótesis:

- Homogeneidad de las varianzas de las poblaciones (**homocedasticidad**)
- **Normalidad** de las observaciones.
- **Independencia** de las observaciones.

El **diagnóstico del modelo** consiste en valorar si es razonable asumir que estas hipótesis se verifican.

El diagnóstico del modelo se basa en el **análisis de los residuos**.

Propiedades de los residuos

Para $i = 1, \ldots, I$ y $j = 1, \ldots, n_i$, recordamos que el residuo e_{ij} es:

$$e_{ij} = Y_{ij} - \bar{Y}_{i.}$$

Bajo las hipótesis del modelo:

- Los residuos tienen distribución normal de media cero.
- Si el diseño es aproximadamente equilibrado, todos los residuos tienen aproximadamente la misma varianza.
- Si el número de datos n es grande respecto al número de grupos I, los residuos son aproximadamente independientes.

Si los residuos no se comportan de acuerdo con estas propiedades es posible que las hipótesis del modelo no se cumplan.

Homogeneidad de las varianzas

Cuando las varianzas son distintas en las poblaciones se dice que hay **heterocedasticidad**.

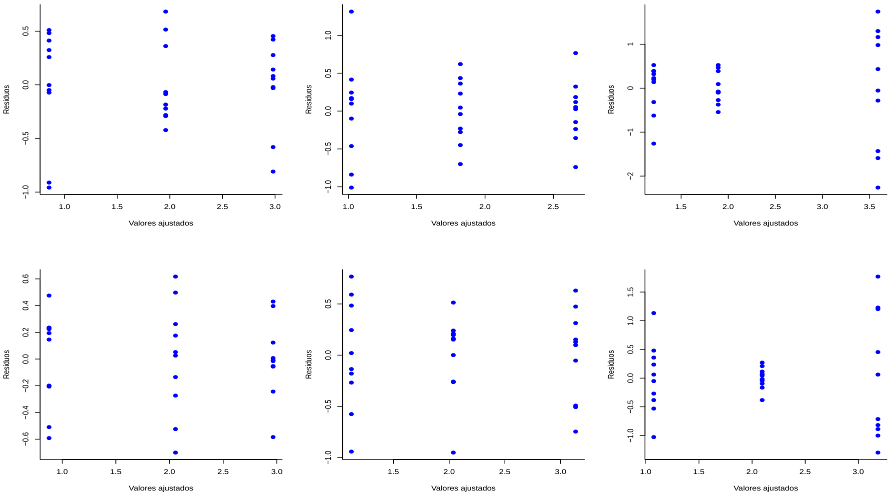
Para detectarla representamos cada valor ajustado $\bar{Y}_{i\cdot}$ frente a sus correspondientes residuos e_{ij}, $i = 1, \ldots, n_i$. Frecuentemente la varianza se incrementa al aumentar el valor ajustado.

Los resultados son relativamente robustos a esta hipótesis. Como regla aproximada, puede haber problemas si el cociente entre la máxima desviación típica y la mínima es mayor que 2.

Otra posibilidad es aplicar el **contraste de Levene** de igualdad de varianzas. La hipótesis nula de este contraste es $H_0 : \sigma_1^2 = \cdots = \sigma_I^2$.

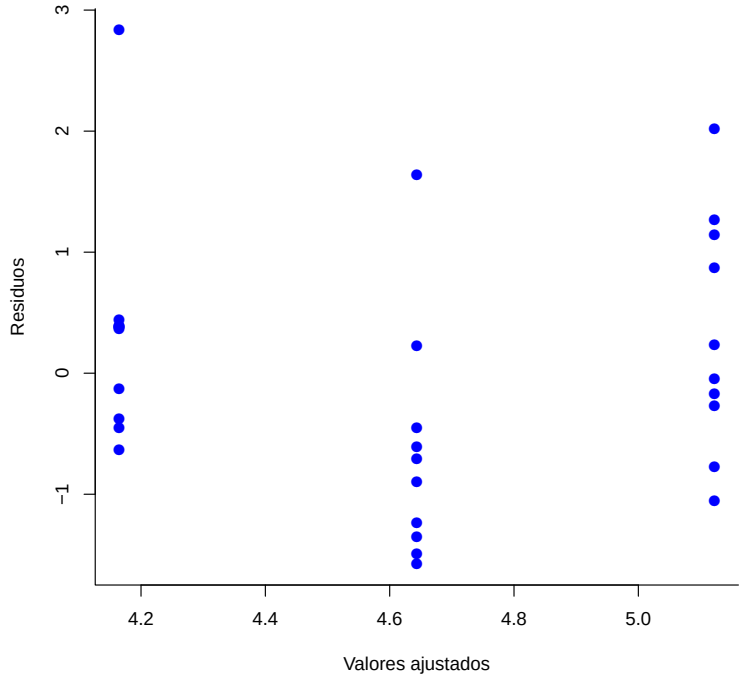
Residuos frente a valores ajustados

Determina los casos en que es posible que haya heterocedasticidad



Residuos frente a valores ajustados

Para los datos de los fertilizantes:



En este ejemplo:

$$\frac{\text{máx}\{S_1, S_2, S_3\}}{\text{mín}\{S_1, S_2, S_3\}} =$$

La hipótesis de normalidad

Para verificar la hipótesis de normalidad de los residuos se pueden utilizar:

- **Procedimientos gráficos:** histogramas o gráficos de probabilidades normales (*probability plots*)
- **Contrastes de bondad de ajuste:** contrastes basados en la distribución χ^2, contraste de Kolmogorov-Smirnov-Lilliefors, etc.

Gráficos de probabilidad

Supongamos que queremos valorar si unos datos $X_1, \ldots, X_n$ proceden de una distribución normal de media μ y desviación típica σ.

Un gráfico de probabilidad normal se basa en las siguientes ideas:

- Ordenamos los datos: $X_{(1)} < \cdots < X_{(n)}$.
- Se puede demostrar que, si los datos realmente proceden de una distribución normal,

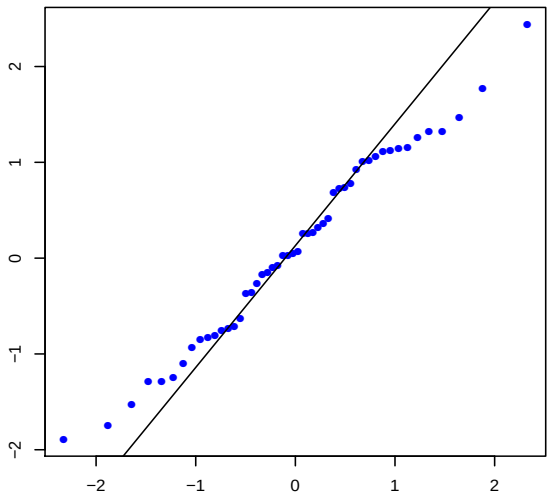
$$\mathrm{E}(X_{(i)}) \approx \mu + \sigma h\left(\frac{i}{n+1}\right),$$

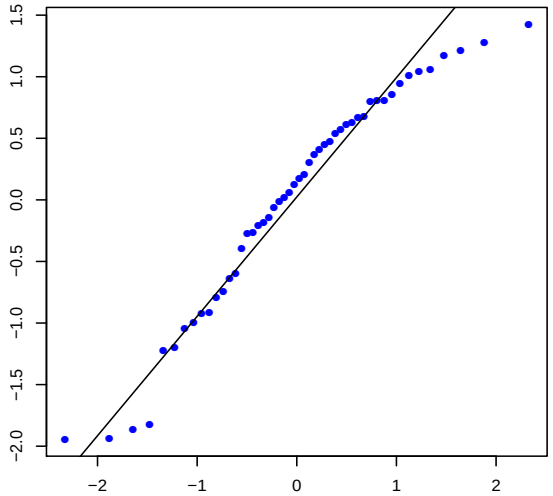
  donde h es una función conocida.
- Por lo tanto si representamos en un gráfico los puntos

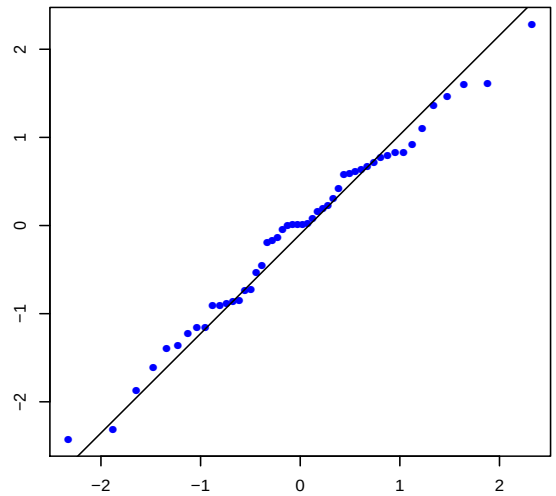
$$\left(X_{(i)}, h\left(\frac{i}{n+1}\right)\right), \quad i = 1, \ldots, n$$

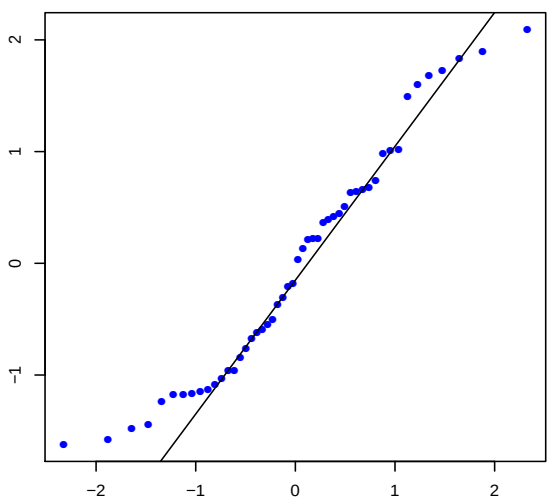
  y la hipótesis de normalidad se cumple, debemos observar que están aproximadamente alineados.

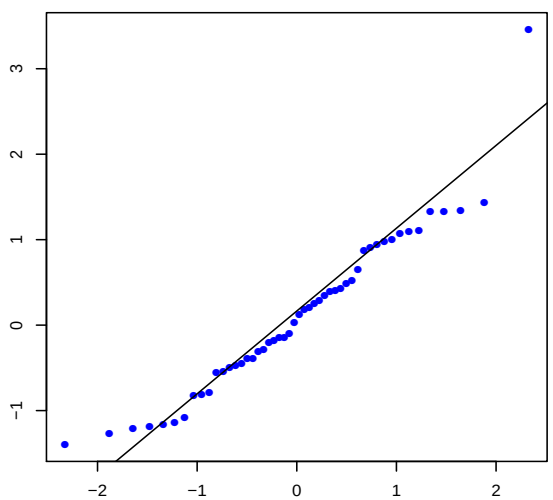
Gráficos de probabilidad para 6 muestras normales ($n = 50$)

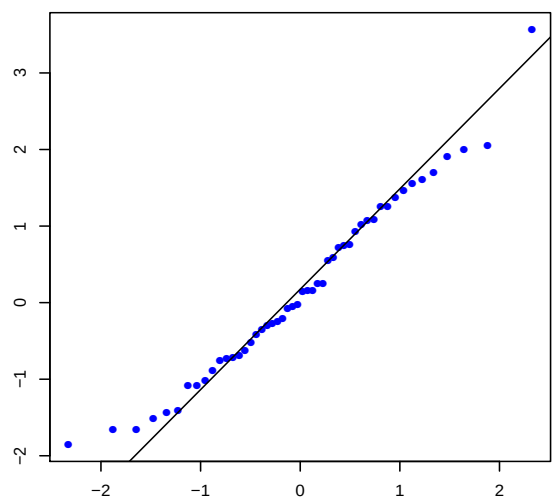




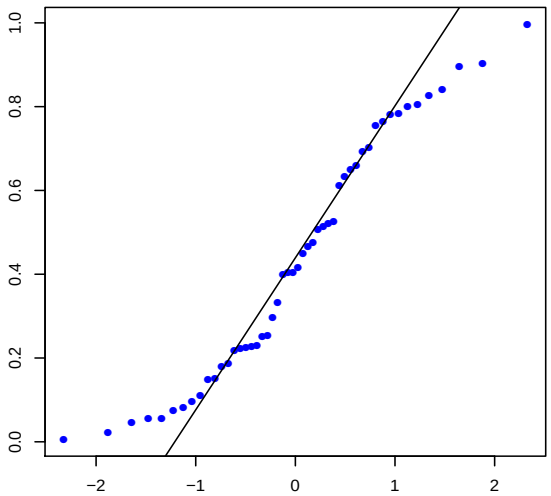


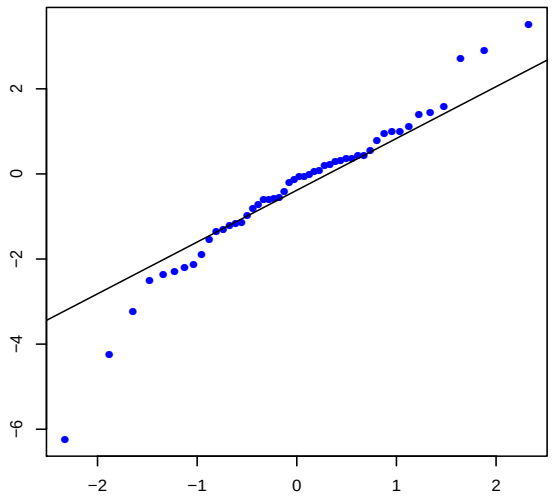


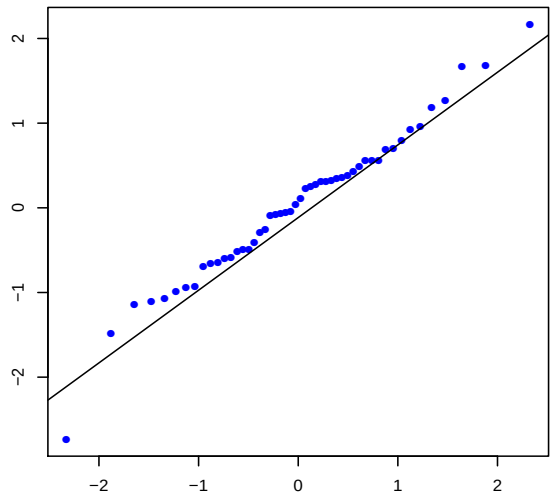


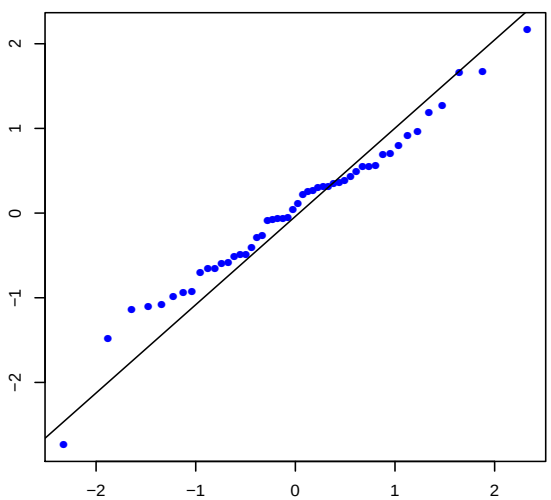


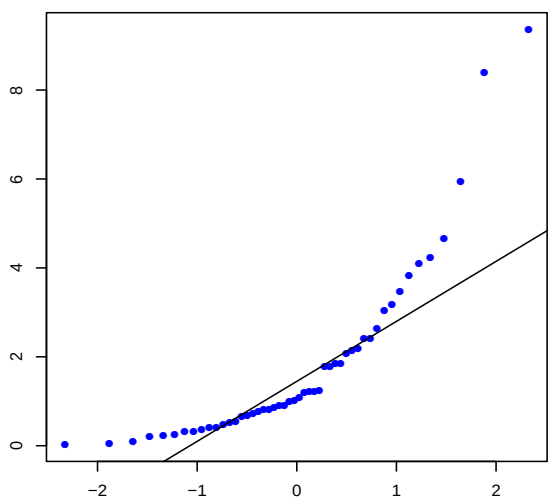
¿Cuáles de estas muestras ($n = 50$) proceden de una distribución normal?

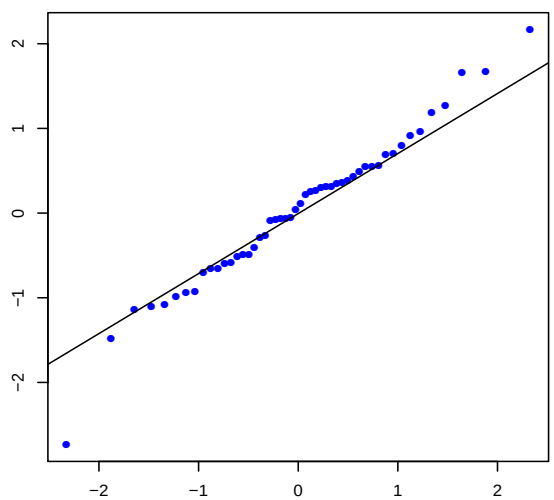




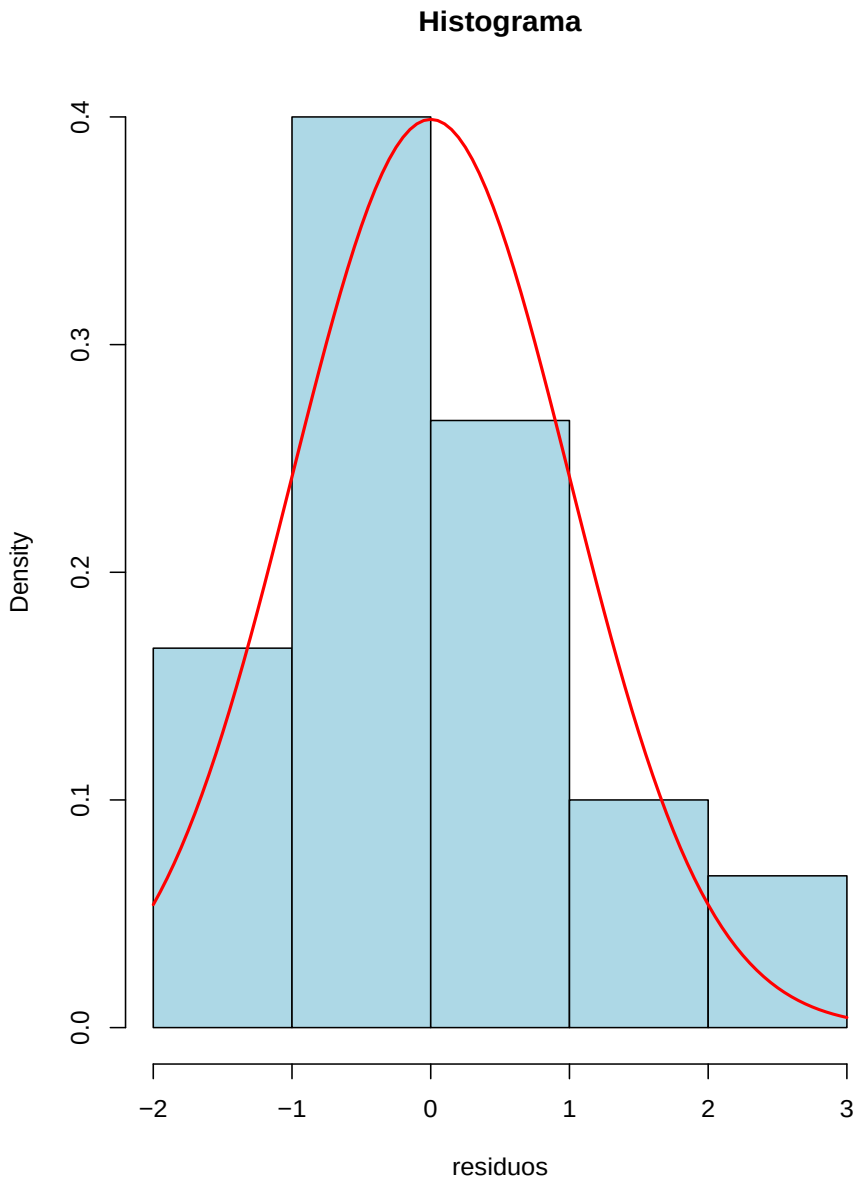


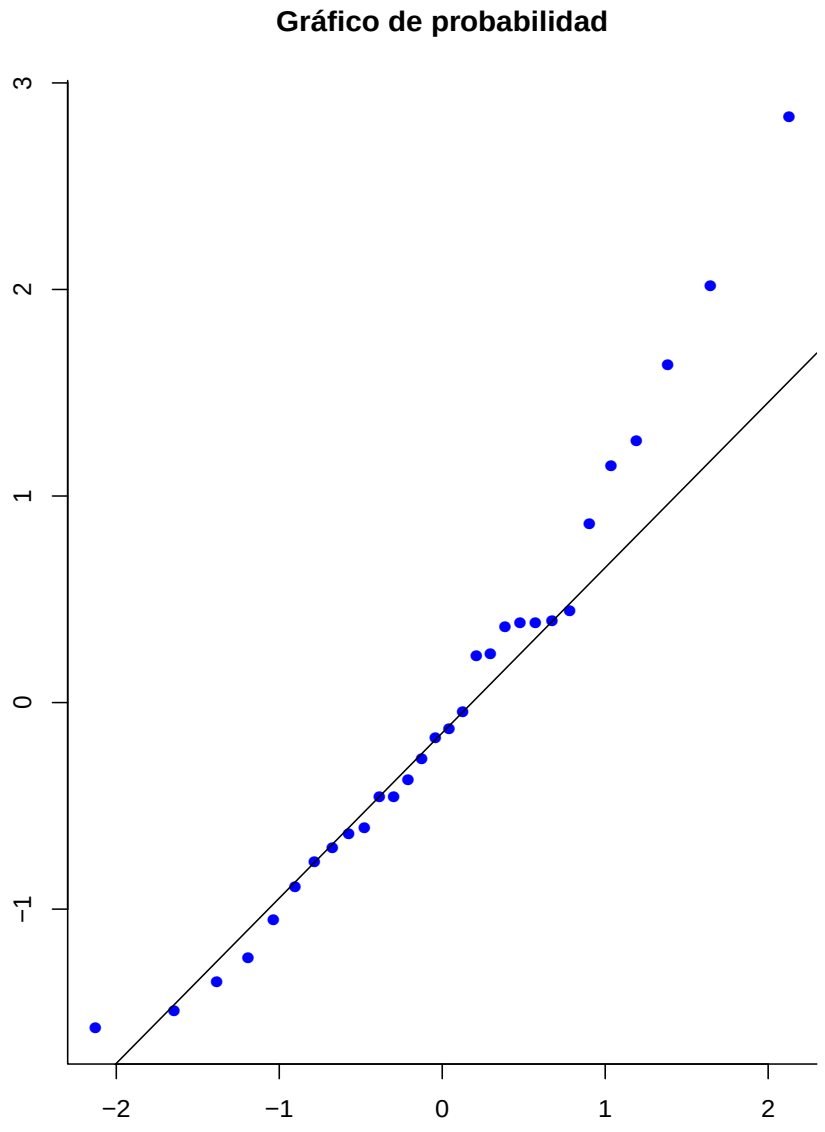






Datos de fertilizantes





Soluciones

Si alguna de las hipótesis del modelo no se cumple, ¿qué se puede hacer?

- Analizar los datos usando otros procedimientos (no-paramétricos) que requieran menos hipótesis.
- La heterocedasticidad o la falta de normalidad se pueden resolver transformando la variable respuesta.
- La transformación más frecuente es $\log Y_{ij}$. Otras transformaciones utilizadas son potencias Y_{ij}^{p}, donde p es un valor adecuado.
- La interpretación de los resultados es más difícil si transformamos los datos.

Ejemplo: desayuno y rendimiento escolar

Los jóvenes que desayunan mal rinden menos

Sólo un tercio de los chicos de 12 a 17 años se alimenta bien por la mañana

CARMEN GIRONA
Madrid

Son 7 de cada 10 adolescentes los que toman un desayuno insuficiente, y la calidad de esa alimentación está directamente relacionada con la nota media del curso, aunque dicha relación no es proporcional cuando se analizan las diferentes asignaturas. Éstas son algunas de las ideas que se desprenden de un estudio coordinado por María Victorina Aguilar Vilas, directora del departamento de Nutrición, Bromatología y Toxicología de la Universidad de Alcalá de Henares de Madrid. En la calidad de la alimentación de primera hora influyen, además, otros factores, como el estado nutricional del joven y la situación económica, cultural y social de la familia.

Para evaluar la calidad de la primera comida del día el grupo madrileño clasificó los desayunos en "completo" (cubre el 25% de las necesidades diarias de energía e incluye alimentos de cuatro grupos: lácteos, cereales, frutas y aceites), de "buena calidad" (contiene los cuatro grupos pero no llega al 25% del aporte energético), de "mejorable calidad" (alimentos de tres grupos), de "insuficiente calidad" (sólo de dos), y de "mala calidad" (no se desayuna). Para evaluar el rendimiento se usó la nota media final del curso y la de seis asignaturas obligatorias relacionadas con la comprensión, la memoria y la actividad física.

En el estudio, publicado en el número de julio y agosto de la revista *Nutrición Hospitalaria,* han participado 467 escolares de 12 a 17 años del curso 2003-2004. Los datos revelan que el 3,65% no desayuna; otro 3,65% toma un desayuno de insuficiente calidad; el 68,29% toma uno de calidad mejorable; el 29,7%, un desayuno de buena calidad, y sólo el 4,88%, un desayuno completo. El trabajo también revela que las chicas desayunan peor que los chicos: el 3,33% de las de 12 a 14 años y el 8,33% de las de 15 a 17 van al colegio sin haber tomado nada. Asimismo, sólo el 4,17% de las chicas mayores toman un desayuno completo, frente al 18,18% de los chicos.

Respecto a la nota media, los datos reflejan que cuanto más completo es el desayuno, mejores notas (6,18 para los que desayunan mal y 7,17 para los que toman un desayuno completo). Por asignaturas, a mayor calidad del desayuno, mejores notas en las asignaturas que pre-

- Fuente: *El País, 11 de noviembre de 2008*.
- Los datos se refieren a $N = 467$ escolares.
- Los escolares se dividen en 5 grupos de acuerdo con el desayuno: completo, buena calidad, mejorable calidad, insuficiente calidad, mala calidad.
- Para cada estudiante se registra su nota media final y la nota final para 6 asignaturas.

Los datos

Tabla II
Grupos de desayuno considerados según su calidad

Desayuno completo	25% de las necesidades diarias de energía e incluir alimentos de, al menos, cuatro grupos distintos: lácteos, cereales, frutas, aceites y grasa, etc.
Buena calidad	Contiene un alimento, al menos, del grupo de lácteos, cereales y fruta.
Mejorable calidad	Falta uno de los grupos.
Insuficiente calidad	Faltan dos de los grupos.
Mala calidad	No desayuna.

Tabla V
Relación entre la calidad de su desayuno y calificación en diversas asignaturas cursadas

Calidad del desayuno	*% población*	*Media*	*Lengua*	*Matemáticas*	*Física-Química*	*Biología*	*Ciencias Sociales*	*Educación física*
Completo	4,88	7,17 ± 1,74	5,83 ± 1,11	6,00 ± 1,33	7,0 ± 1,14	6,16 ± 0,54	7,66 ± 0,56	7,4 ± 0,24
Buena calidad	29,27	6,84 ± 0,30	6,58 ± 0,42	6,08 ± 0,29	6,0 ± 0,01	6,08 ± 0,47	7,13 ± 0,47	7,29 ± 9,28
Mejorable calidad	68,29	6,61 ± 0,16	6,61 ± 0,38	5,92 ± 0,52	7,4 ± 0,45	6,10 ± 0,36	7,45 ± 0,11	7,24 ± 0,24
Insuficiente calidad	3,65	6,48 ± 0,01	7,00 ± 0,14	5,33 ± 0,06	6,0 ± 0,01	5,0 ± 0,01	6,00 ± 0,16	8,33 ± 0,04
Mala calidad	3,65	6,18 ± 1,89	6,00 ± 0,38	5,66 ± 3,59	2,0 ± 0,01	6,33 ± 1,00	6,33 ± 1,94	8,33 ± 0,28

Para cada grupo, los datos de la tabla se presentan en la forma

$$\texttt{media} \pm \texttt{desviación típica}$$

El objetivo es analizar si hay diferencias significativas entre los 5 grupos.

En el ejemplo (considerando como respuesta la nota media):

Grupo	1	2	3	4	5
n_i	21	125	291	15	15
Media	7,17	6,84	6,61	6,48	6,18
S_i	1,74	0,30	0,16	0,01	1,89

Sumas de cuadrados y estadístico F

$$\bar{Y}_{\cdot\cdot} = \frac{(21 \times 7{,}17) + \cdots + (15 \times 6{,}18)}{467} = \frac{3118{,}98}{467} \approx 6{,}68$$

$$\text{SCE} = 21(7{,}17 - 6{,}68)^2 + 125(6{,}84 - 6{,}68)^2 + \cdots + 15(6{,}18 - 6{,}68)^2 \approx 14{,}02$$

$$\text{SCR} = 20 \times 1{,}74^2 + 124 \times 0{,}30^2 + \cdots + 14 \times 1{,}89^2 \approx 129{,}15$$

$$F = \frac{\text{SCE}/(I-1)}{\text{SCR}/(n-I)} = \frac{14{,}02/4}{129{,}15/462} \approx 12{,}54$$

Tabla ANOVA

Fuente	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada	14.02	4	3.50	12.54
Residual	129.15	462	0.28	

De acuerdo con las tablas de la distribución F,

$$F_{4,462,0,05} \approx 2{,}39$$

Como $12{,}54 > 2{,}39$, estamos en la región crítica.

Los datos aportan evidencia a nivel $\alpha = 0{,}05$ de que las medias no son iguales, es decir, el tipo de desayuno influye en la nota media.

¿Qué crítica se podría hacer respecto a esta conclusión teniendo en cuenta cómo se han obtenido los datos?