Tema 2
Análisis de la varianza multifactorial

Estudia la influencia de dos o más factores (variables explicativas) **sobre la media de una variable aleatoria** (variable respuesta)

- **Definición de la variable a explicar**

- **Definición de los distintos factores que pueden influir en la respuesta y, en cada uno de ellos, sus distintos niveles o grupos.**

Estudiaremos tres casos:

1. **Dos factores sin interacción (diseño por bloques)**
2. **Dos factores con posible interacción entre ellos**
3. **Tres factores (Cuadrados latinos)**

Análisis de la varianza con dos factores
Diseño por bloques

Modelo:

$$Y_{ij} = \mu + \alpha_i + \beta_j + U_{ij} \quad i = 1, 2, \ldots, I \quad j = 1, 2, \ldots, J$$

Y_{ij} es la respuesta de la variable en el i-ésimo nivel del factor 1 (α) y en el j-ésimo nivel del factor 2 (β).

$\mu_{ij} = E(Y_{ij}) = \mu + \alpha_i + \beta_j$ es el valor medio de Y_{ij}

α_i representa el efecto que sobre la media global μ tiene del nivel i del factor 1

β_j representa el efecto que sobre la media global μ tiene del nivel j del factor 2

U_{ij} es la variación aleatoria de las Y_{ij} (igual en distribución para todas ellas)

Supondremos que U_{ij} sigue una distribución $N(0,\sigma)$ lo que implica que Y_{ij} sigue una distribución $N(\mu_{ij},\sigma)$ y que no hay interacción entre los factores.

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$$

Muestra aleatoria (una observación por casilla)

Factor 2 (β)

Factor 1 (α) Niveles	1	2	...	...	J	Medias por filas
1	Y_{11}	Y_{12}	...	...	Y_{1J}	$\overline{Y}_{1.}$
2	Y_{21}	Y_{22}	...	...	Y_{2J}	$\overline{Y}_{2.}$
...	...	...	...	...	...	...
I	Y_{I1}	Y_{I2}	...	...	Y_{IJ}	$\overline{Y}_{I.}$
Medias por columnas	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	...	...	$\overline{Y}_{.J}$	$\overline{Y}_{..}$

$$Y_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2), \text{ independientes}$$

Datos (Ejemplo 1)

Se desea estudiar la eficiencia (en cuanto a menor emisión de CO_2) de 5 máquinas desaladoras. Se piensa que la cantidad de sal en el agua puede influir en dicha eficiencia.
Factor 1: distintas máquinas (I=5)
Factor 2: nivel de sal (J=3)

	Poca sal	Bastante sal	Mucha sal	$\bar{Y}_{i.}$
Mq I	24	26	29	26.3
Mq II	27	30	32	29.6
Mq III	26	27	30	27.6
Mq IV	25	28	28	27
Mq V	28	29	31	29.3
$\bar{Y}_{.j}$	26	28	30	$\bar{Y}_{..} = 28$

Análisis estadístico:

Estimación de los parámetros desconocidos

Parámetros desconocidos del modelo (I + J) :

$$\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \sigma$$

Estimaciones de los parámetros:

$$\hat{\mu} = \bar{y}_{..} = \frac{1}{IJ}\sum_i \sum_j y_{ij}$$

$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} = \frac{1}{J}\sum_j y_{ij} - \bar{y}_{..}$$

$$\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..} = \frac{1}{I}\sum_i y_{ij} - \bar{y}_{..}$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{(I-1)(J-1)}\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

Análisis estadístico: ANOVA

$$SCE(\alpha) = J \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SCE(\beta) = I \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2$$

$$SCR = \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

$$SCT = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$$

Se cumple que:
SCE(α) + SCE(β) + SCR = SCT

SCE(α) Suma de cuadrados explicada (variabilidad debida a que hay distintos niveles del factor 1)

SCE(β) Suma de cuadrados explicada (variabilidad debida a que hay distintos niveles del factor 2)

SCR Suma de cuadrados residual (variabilidad no debida a los factores)

SCT Suma de cuadrados total (variabilidad total de todos los datos)

Análisis estadístico: ANOVA

(Contrastes del efecto de cada factor)

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

$$H_1 : \text{Algún } \alpha_i \neq 0$$

El factor 1 no influye

Estadístico de contraste $F(\alpha) = \dfrac{SCE(\alpha)/(I-1)}{SCR/(I-1)(J-1)}$:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$

$$H_1 : \text{Algún } \beta_j \neq 0$$

El factor 2 no influye

Estadístico de contraste $F(\beta) = \dfrac{SCE(\beta)/(J-1)}{SCR/(I-1)(J-1)}$

Análisis estadístico: ANOVA
(Tabla)

Tabla ANOVA

Suma de cuadrados	g.l.	Varianza	Estadístico
$SCE(\alpha) = J\sum_i \hat{\alpha}_i^2$	$I-1$	$\frac{SCE(\alpha)}{I-1}$	$F(\alpha)$
$SCE(\beta) = I\sum_j \hat{\beta}_i^2$	$J-1$	$\frac{SCE(\beta)}{J-1}$	$F(\beta)$
$SCR = \sum_i \sum_j \hat{e}_{ij}^2$	$(I-1)(J-1)$	$\frac{SCR}{(I-1)(J-1)}$	
$SCT = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	$IJ-1$		

$$F(\alpha) = \frac{SCE(\alpha)/(I-1)}{SCR/(I-1)(J-1)}\ ; \qquad F(\beta) = \frac{SCE(\beta)/(J-1)}{SCR/(I-1)(J-1)}$$

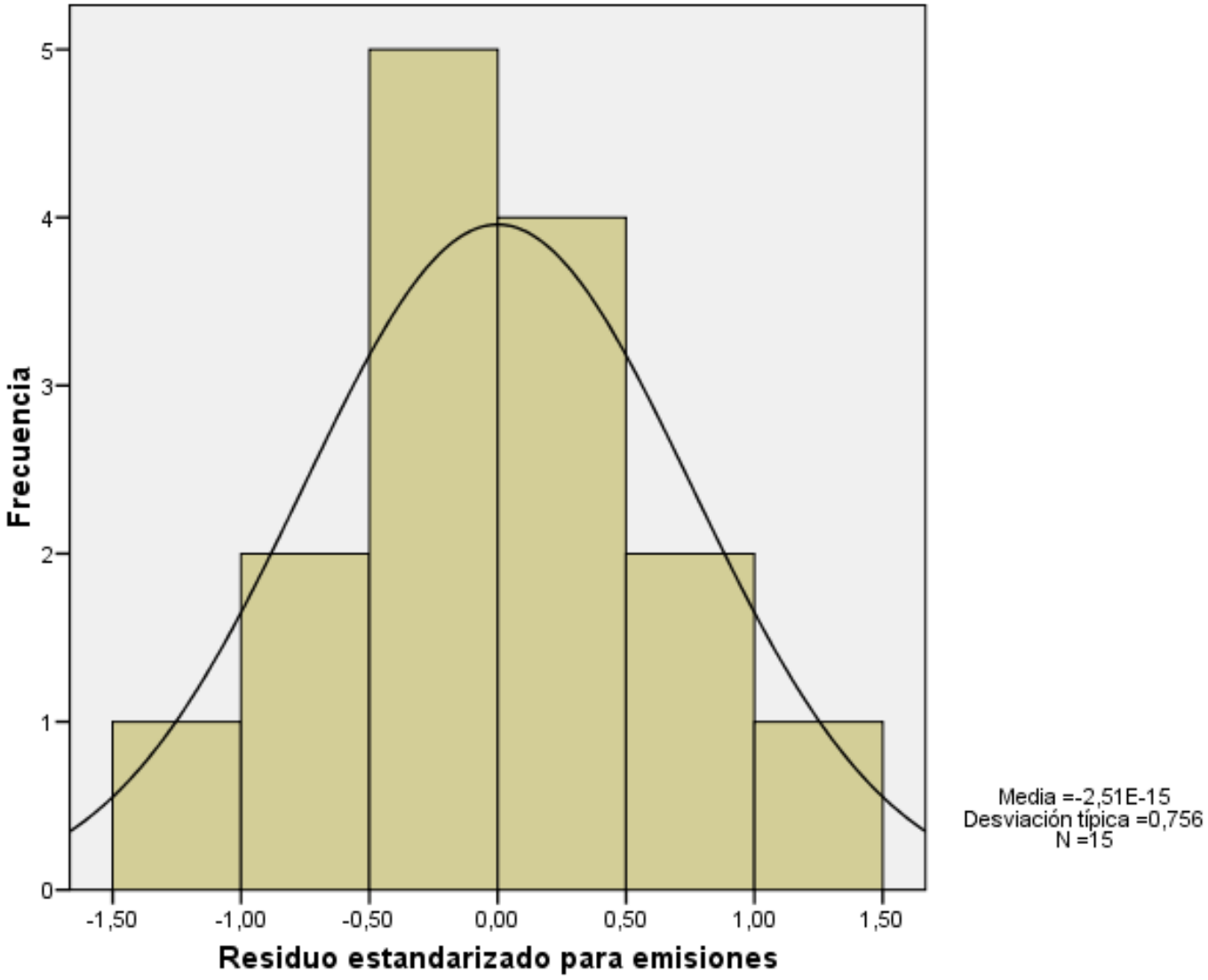
Con los datos del ejemplo 1:

Fuente de variación	Suma de cuadrados	g.l.	Varianzas	Test F	p-valor
Máquina	25.33	4	6.33	10.857	0.0026
Tipo agua	40	2	20	34.286	0.0001
Residual	4.66	8	0.583		
Total	70	14			

En cuanto a las emisiones de CO_2 las 5 máquinas no son iguales (p-valor 0.0026) y también influye la cantidad de sal (p-valor 0.0001).

¿Y si no hubiéramos tenido en cuenta el factor "cantidad de sal" ?

Fuente de variación	Suma de cuadrados	g.l.	Varianzas	Test F	p-valor
Máquina	25.33	4	6.33	1.418	0.2972
Residual	44.66	10	4.46		
Total	70	14			


Frecuencia
5
4
3
2
1
0
-1,50
-1,00
-0,50
0,00
0,50
1,00
1,50
Residuo estandarizado para emisiones
Media =-2,51E-15
Desviación típica =0,756
N =15

Coeficiente de determinación:

$$R^2 = \frac{\mathsf{SCE}}{\mathsf{SCT}} = \frac{\mathsf{SCE}(\alpha)+\mathsf{SCE}(\beta)}{\mathsf{SCT}}$$

El porcentaje de variabilidad explicada por el modelo es

$$R^2 \times 100$$

y, en particular, por cada uno de los factores:

$$\frac{\mathsf{SCE}(\alpha)}{\mathsf{SCT}} \times 100 \text{ y } \frac{\mathsf{SCE}(\beta)}{\mathsf{SCT}} \times 100$$

En el ejemplo R^2 x 100 = 93.3 = 36.2 (máquinas) + 57.1 (sal)

Análisis posteriores al rechazo de H_0

H_0 : No hay diferencia entre los niveles i, j del factor 1

Con nivel de significación α rechazamos H_0 si el cero no está en el siguiente intervalo de confianza:

$$IC_{1-\alpha}(\alpha_i - \alpha_j) = \left(\bar{y}_{i.} - \bar{y}_{j.} \pm t_{(I-1)(J-1);\alpha/2} S_R \sqrt{\frac{1}{J} + \frac{1}{J}} \right)$$

H_0 : No hay diferencia entre los niveles i, j del factor 2

Con nivel de significación α rechazamos H_0 si el cero no está en el siguiente intervalo de confianza:

$$IC_{1-\alpha}(\beta_i - \beta_j) = \left(\bar{y}_{.i} - \bar{y}_{.j} \pm t_{(I-1)(J-1);\alpha/2} S_R \sqrt{\frac{1}{I} + \frac{1}{I}} \right)$$

Error típico

Comparaciones múltiples:

Pruebas Post hoc: Test de Bonferroni

Al igual que en el análisis de la varianza con un factor podemos hacer pruebas simultáneas entre todas las posibles parejas de niveles en cada factor. Por ejemplo utilizando el Test de Bonferroni.

En el ejemplo 1:

Comparaciones múltiples

Variable dependiente: EMISIONES

Bonferroni

(I) Calidad del agua (bloque)	(J) Calidad del agua (bloque)	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 85%.	
					Límite inferior	Límite superior
Poca sal	Poca sal					
	Bastante sal	-2,0000*	,48305	,010	-3,1139	-,8861
	Mucha sal	-4,0000*	,48305	,000	-5,1139	-2,8861
Bastante sal	Poca sal	2,0000*	,48305	,010	,8861	3,1139
	Bastante sal					
	Mucha sal	-2,0000*	,48305	,010	-3,1139	-,8861
Mucha sal	Poca sal	4,0000*	,48305	,000	2,8861	5,1139
	Bastante sal	2,0000*	,48305	,010	,8861	3,1139
	Mucha sal					

Basado en las medias observadas.

*. La diferencia de medias es significativa al nivel ,15.

Comparaciones múltiples

Variable dependiente: EMISIONES

Bonferroni

(I) Máquina (factor)	(J) Máquina (factor)	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
Máquina I	Máquina I					
	Máquina II	-3,3333*	,62361	,007	-5,7233	-,9433
	Máquina III	-1,3333	,62361	,650	-3,7233	1,0567
	Máquina IV	-,6667	,62361	1,000	-3,0567	1,7233
	Máquina V	-3,0000*	,62361	,013	-5,3900	-,6100
Máquina II	Máquina I	3,3333*	,62361	,007	,9433	5,7233
	Máquina II					
	Máquina III	2,0000	,62361	,125	-,3900	4,3900
	Máquina IV	2,6667*	,62361	,027	,2767	5,0567
	Máquina V	,3333	,62361	1,000	-2,0567	2,7233
Máquina III	Máquina I	1,3333	,62361	,650	-1,0567	3,7233
	Máquina II	-2,0000	,62361	,125	-4,3900	,3900
	Máquina III					
	Máquina IV	,6667	,62361	1,000	-1,7233	3,0567
	Máquina V	-1,6667	,62361	,282	-4,0567	,7233
Máquina IV	Máquina I	,6667	,62361	1,000	-1,7233	3,0567
	Máquina II	-2,6667*	,62361	,027	-5,0567	-,2767
	Máquina III	-,6667	,62361	1,000	-3,0567	1,7233
	Máquina IV					
	Máquina V	-2,3333	,62361	,057	-4,7233	,0567
Máquina V	Máquina I	3,0000*	,62361	,013	,6100	5,3900
	Máquina II	-,3333	,62361	1,000	-2,7233	2,0567
	Máquina III	1,6667	,62361	,282	-,7233	4,0567
	Máquina IV	2,3333	,62361	,057	-,0567	4,7233
	Máquina V					

Basado en las medias observadas.

*. La diferencia de medias es significativa al nivel ,05.

Análisis de la varianza con dos factores e interacción

Modelo:

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + U_{ij} \quad i = 1, 2, \ldots, I \quad j = 1, 2, \ldots, J$$

Y_{ij} representa la respuesta de la variable en el i-ésimo nivel del factor 1 (α) y en el j-ésimo nivel del factor 2 (β).

$\mu_{ij} = E(Y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ es el valor medio de Y_{ij}

α_i representa el efecto que sobre la media global μ tiene del nivel i del factor 1

β_j representa el efecto que sobre la media global μ tiene del nivel j del factor 2

$(\alpha\beta)_{ij}$ representa el efecto de la interacción entre el nivel i del factor 1 y el nivel j del factor 2

U_{ij} es la variación aleatoria de las Y_{ij} (igual para todas ellas)

Supondremos que U_{ij} sigue una distribución $N(0,\sigma)$ lo que implica que Y_{ij} sigue una distribución $N(\mu_{ij},\sigma)$

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$$

Muestra aleatoria (n_{ij} observaciones en la casilla i,j)

Factor 2 (β)

Factor 1 (α)

Niveles	1	2	...	...	J	Medias por filas
1	Y_{111} .... Y_{11n11}	Y_{121} .... Y_{12n12}	...	...	Y_{1J1} .... Y_{1Jn1J}	$\overline{Y}_{1..}$
			Y_{ijk}			
...	...	...	...	...	...	...
I	Y_{I11} .... Y_{I1nI1}	Y_{I21} .... Y_{I2nI2}	...	...	Y_{IJ1} .... Y_{IJnIJ}	$\overline{Y}_{I..}$
Medias por columnas	$\overline{Y}_{.1.}$	$\overline{Y}_{.2.}$	...	...	$\overline{Y}_{.J.}$	$\overline{Y}_{...}$

$$Y_{ijk} \sim N(\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij};\sigma^2) \text{ independientes;}$$

En un diseño equilibrado todas las casillas tendrán el mismo número de datos (K)

$n_{ij} = K$ para todo i,j

Ejemplo 2 Eysenck (1974)

En un estudio sobre memoria verbal se seleccionaron al azar 50 personas mayores y 50 jóvenes (factor 1: edad). Dentro de cada uno de estos grupos se asignaron, al azar, 10 personas a 5 distintos grupos a los que se les presentó una misma lista de 27 palabras. A cada uno de los 5 grupos se les dieron las siguientes instrucciones (factor 2: método)

Grupo 1 (contar): se les pidió que contasen el nº de letras de cada palabra
Grupo 2 (rimar): se les pidió que rimasen cada palabra con otra
Grupo 3 (adjetivar): se les pidió que a cada palabra le asignasen un adjetivo
Grupo 4 (imaginar): se les pidió que a cada palabra le asignasen una imagen
Grupo 5 (recordar): se les pidió que memorizasen las palabras.

A los 4 primeros grupos no se les dijo que deberían recordar las palabras. Finalmente, tras revisar la lista 3 veces, se recogió el nº de palabras recordadas por cada grupo (variable respuesta).

Datos

I=2, J=5, K=10

		Factor 2 Método				
		Contar	Rimar	Adjetivar	Imaginar	Recordar
Factor 1 Edad	Mayores	9	7	11	12	10
		8	9	13	11	19
		6	6	8	16	14
		8	6	6	11	5
		10	6	14	9	10
		4	11	11	23	11
		6	6	13	12	14
		5	3	13	10	15
		7	8	10	19	11
		7	7	11	11	11
	Jóvenes	8	10	14	20	21
		6	7	11	16	19
		4	8	18	16	17
		6	10	14	15	15
		7	4	13	18	22
		6	7	22	16	16
		5	10	17	20	22
		7	6	16	22	22
		9	7	12	14	18
		7	7	11	19	21

Análisis estadístico:

Estimación de los parámetros desconocidos

Parámetros desconocidos del modelo (IJ + 1) :

$$\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{IJ}, \sigma$$

Estimaciones de los parámetros:

$$\hat{\mu} = \bar{y}_{...} = \frac{1}{IJK} \sum_i \sum_j \sum_k y_{ijk}$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} = \frac{1}{JK} \sum_j \sum_k y_{ijk} - \bar{y}_{...}$$

$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} = \frac{1}{IK} \sum_i \sum_k y_{ijk} - \bar{y}_{...}$$

$$\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} = \frac{1}{K} \sum_k y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{IJ(K-1)} \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$$

Tenemos $IJ + 1$ parámetros desconocidos

$\boldsymbol{\mu}$	1 parámetro	$IJ + 1$
$\boldsymbol{\alpha_i}$	$(I - 1)$ parámetros	
$\boldsymbol{\beta_j}$	$(J - 1)$ parámetros	
$\boldsymbol{(\alpha\beta)_{ij}}$	$(I - 1)(J - 1)$ parámetros	
$\boldsymbol{\sigma^2}$	1 parámetro	

Tenemos:

- $\boldsymbol{IJ + 1}$ parámetros desconocidos
- $\boldsymbol{n = IJK}$ datos

Si K es mayor que 2, siempre $n > IJ + 1$

Análisis estadístico: ANOVA

$$SCE(\alpha) = JK \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SCE(\beta) = IK \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SCE(\alpha\beta) = K \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SCR = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$$

$$SCT = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$$

Se cumple que:
SCE(α) +
SCE(β) +
SCE(αβ) +
SCR = SCT

SCE(α) Suma de cuadrados explicada (variabilidad debida a que hay distintos niveles del factor 1)

SCE(β) Suma de cuadrados explicada (variabilidad debida a que hay distintos niveles del factor 2)

SCE (αβ) Suma de cuadrados explicada (variabilidad debida a las interacciones)

SCR Suma de cuadrados residual (variabilidad no debida a los factores)

SCT Suma de cuadrados total (variabilidad total de todos los datos)

Análisis estadístico: ANOVA

(Contrastes del efecto de cada factor)

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$
$$H_1 : \text{Algún } \alpha_i \neq 0$$

← El factor 1 no influye

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$
$$H_1 : \text{Algún } \beta_j \neq 0$$

← El factor 2 no influye

$$H_0 : (\alpha\beta)_{ij} = 0 \text{ para todos } i, j, \; H_1 : \text{algún } (\alpha\beta)_{ij} \neq 0$$

No hay interacciones

Estadísticos de contraste

$$F(\alpha) = \frac{SCE(\alpha)/(I-1)}{SCR/IJ(K-1)}; \qquad F(\beta) = \frac{SCE(\beta)/(J-1)}{SCR/IJ(K-1)}$$
$$F(\alpha\beta) = \frac{SCE(\alpha\beta)/(I-1)(J-1)}{SCR/IJ(K-1)}$$

Análisis estadístico: Tabla ANOVA

Tabla ANOVA

Suma de cuadrados	G.l.	Varianza	F
$SCE(\alpha) = JK \sum_i \hat{\alpha}_i^2$	$I-1$	$\frac{SCE(\alpha)}{I-1}$	$F(\alpha)$
$SCE(\beta) = IK \sum_j \hat{\beta}_j^2$	$J-1$	$\frac{SCE(\beta)}{J-1}$	$F(\beta)$
$SCE(\alpha\beta) = K \sum_i \sum_j (\hat{\alpha\beta})_{ij}^2$	$(I-1)(J-1)$	$\frac{SCE(\alpha\beta)}{(I-1)(J-1)}$	$F(\alpha\beta)$
$SCR = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$	$IJ(K-1)$	$\frac{SCR}{IJ(K-1)}$	
$SCT = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$	$IJK-1$		

$$F(\alpha) = \frac{SCE(\alpha)/(I-1)}{SCR/IJ(K-1)} \quad F(\beta) = \frac{SCE(\beta)/(J-1)}{SCR/IJ(K-1)} \quad F(\alpha\beta) = \frac{SCE(\alpha\beta)/(I-1)(J-1)}{SCR/IJ(K-1)}$$

Recordemos que siempre deben cumplirse los siguientes requisitos previos

1. **Normalidad:** los datos obtenidos en cada nivel de los factores se ajustan razonablemente a una distribución Normal (gráficos y contrastes)
   Y_{ij} sigue una distribución $N(\mu_{ij}, \sigma)$ para cada i,j
2. **Homocedasticidad:** la variabilidad de los datos en cada nivel de los factores es similar (contraste de igualdad de varianzas)
   $\sigma^2 = Var(Y_{ij})$ igual para todo i,j
3. **Linealidad:** los residuos tipificados se distribuyen alrededor del cero
   $E(U_{ij}) = 0$
4. **Independencia:** las observaciones se realizan de forma independiente unas de otras (diseño de la obtención de datos)

SI HAY DESVIACIONES SIGNIFICATIVAS SOBRE ESTOS REQUISITOS LOS RESULTADOS POSTERIORES PUEDEN SER INCORRECTOS

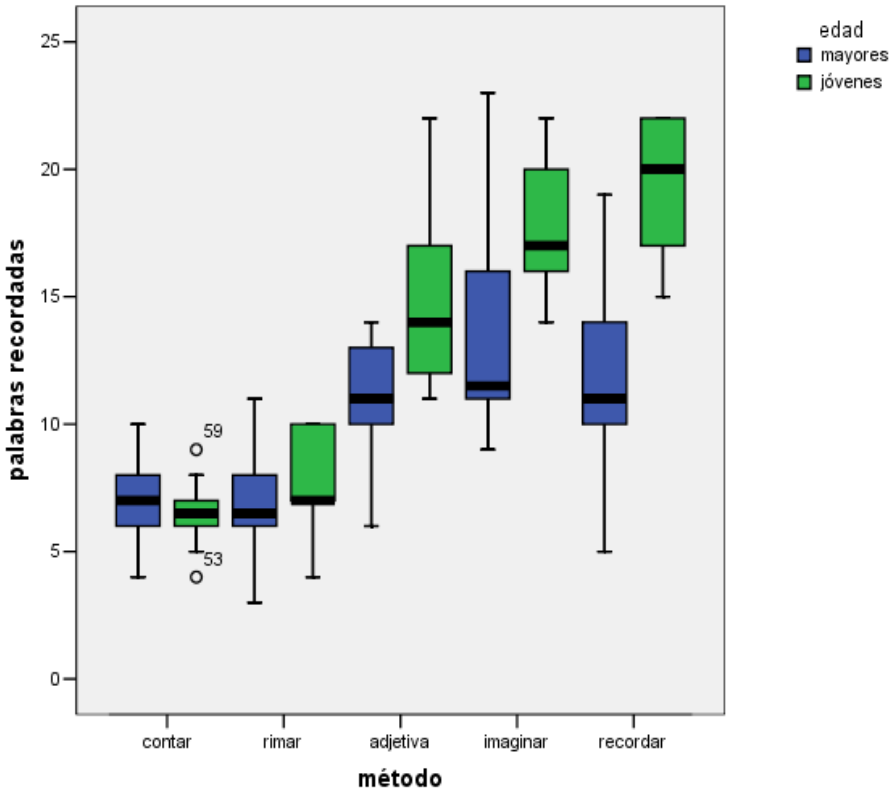
Ejemplos
con Excel y SPSS

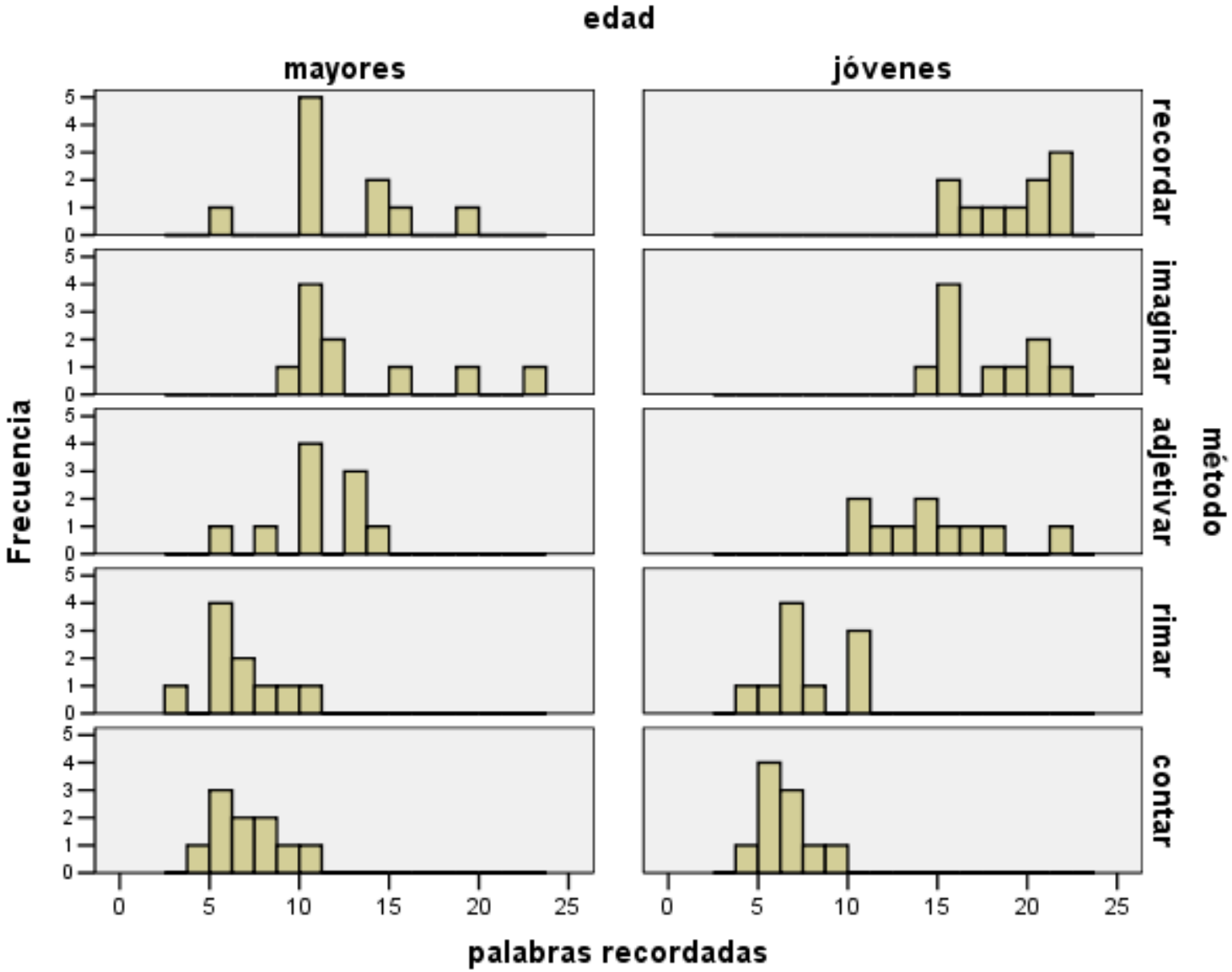
Ejemplo 2 Eysenck (1974)

Estadísticos descriptivos

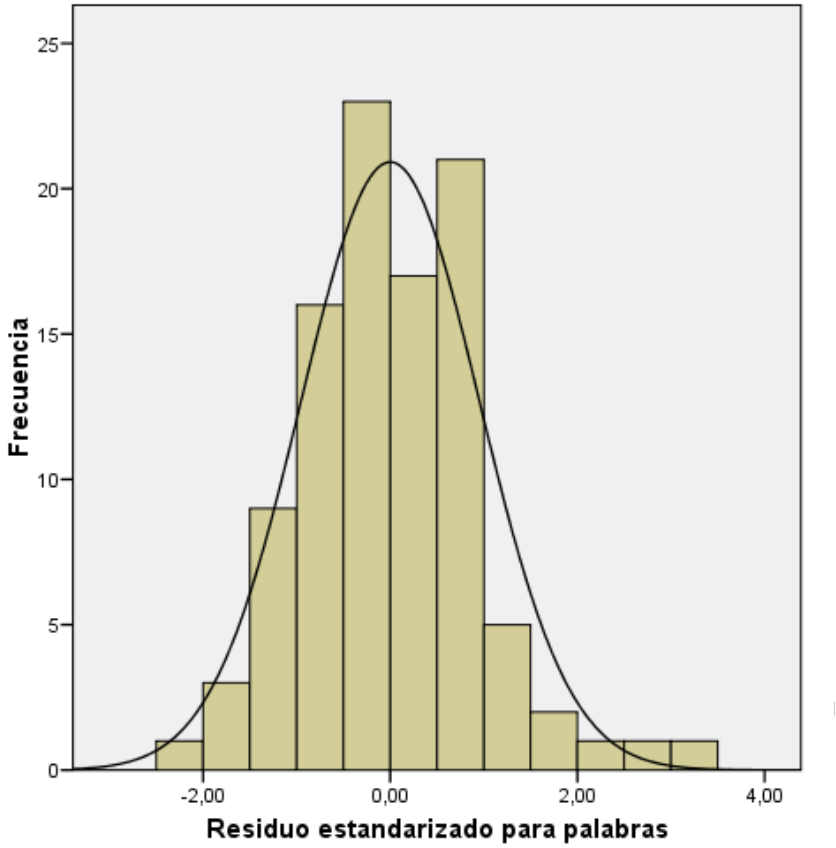
Variable dependiente: palabras recordadas

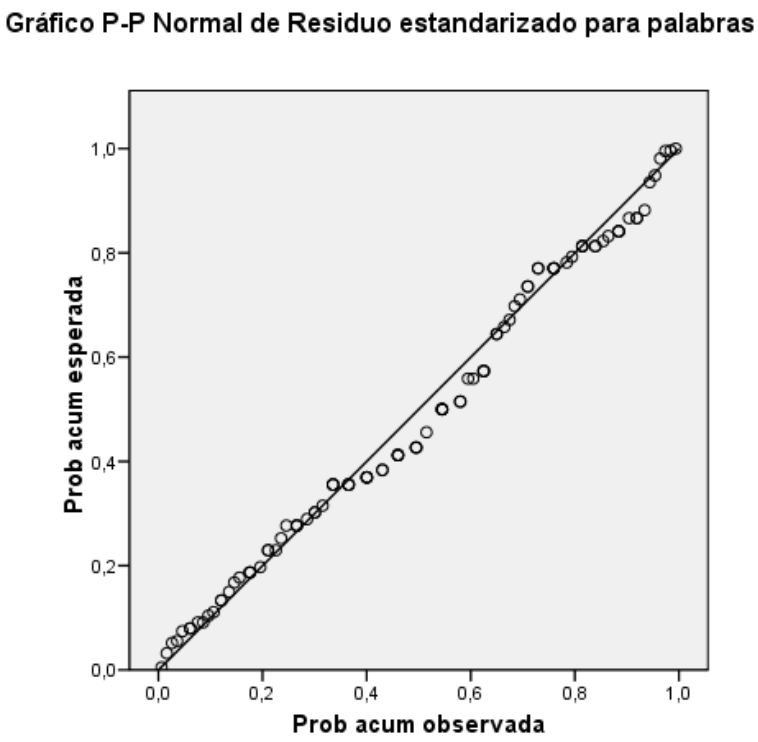
edad	método	Media	Desv. típ.	N
mayores	contar	7,00	1,826	10
	rimar	6,90	2,132	10
	adjetivar	11,00	2,494	10
	imaginar	13,40	4,502	10
	recordar	12,00	3,742	10
	Total	10,06	4,007	50
jóvenes	contar	6,50	1,434	10
	rimar	7,60	1,955	10
	adjetivar	14,80	3,490	10
	imaginar	17,60	2,591	10
	recordar	19,30	2,669	10
	Total	13,16	5,787	50
Total	contar	6,75	1,618	20
	rimar	7,25	2,023	20
	adjetivar	12,90	3,538	20
	imaginar	15,50	4,174	20
	recordar	15,65	4,902	20
	Total	11,61	5,191	100




edad
mayores
jóvenes
recordar
imaginar
adjetivar
rimar
contar
método
Frecuencia
palabras recordadas
0
5
10
15
20
25

Residuos tipificados (SPSS)





Excel

ANÁLISIS DE VARIANZA					
Origen de las variaciones	*Suma de cuadrados*	*g.l.*	*Promedio de los cuadrados*	*F*	*p-valor*
Edad	240,25	1	240,25	29,94	3,9814E-07
Método	1514,94	4	378,74	47,19	2,5301E-21
Interacción	190,3	4	47,58	5,93	0,00027927
Error	722,3	90	8,03		
Total	2667,79	99			

SPSS

Pruebas de los efectos inter-sujetos

Variable dependiente: palabras recordadas

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	1945,490[a]	9	216,166	26,935	,000
Intersección	13479,210	1	13479,210	1679,536	,000
edad	240,250	1	240,250	29,936	,000
método	1514,940	4	378,735	47,191	,000
edad * método	190,300	4	47,575	5,928	,000
Error	722,300	90	8,026		
Total	16147,000	100			
Total corregida	2667,790	99			

a. R cuadrado = ,729 (R cuadrado corregida = ,702)

Comparaciones múltiples

Variable dependiente: palabras recordadas

Bonferroni

(I) método	(J) método	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
contar	rimar	-,50	,896	1,000	-3,08	2,08
	adjetivar	-6,15*	,896	,000	-8,73	-3,57
	imaginar	-8,75*	,896	,000	-11,33	-6,17
	recordar	-8,90*	,896	,000	-11,48	-6,32
rimar	contar	,50	,896	1,000	-2,08	3,08
	adjetivar	-5,65*	,896	,000	-8,23	-3,07
	imaginar	-8,25*	,896	,000	-10,83	-5,67
	recordar	-8,40*	,896	,000	-10,98	-5,82
adjetivar	contar	6,15*	,896	,000	3,57	8,73
	rimar	5,65*	,896	,000	3,07	8,23
	imaginar	-2,60*	,896	,047	-5,18	-,02
	recordar	-2,75*	,896	,028	-5,33	-,17
imaginar	contar	8,75*	,896	,000	6,17	11,33
	rimar	8,25*	,896	,000	5,67	10,83
	adjetivar	2,60*	,896	,047	,02	5,18
	recordar	-,15	,896	1,000	-2,73	2,43
recordar	contar	8,90*	,896	,000	6,32	11,48
	rimar	8,40*	,896	,000	5,82	10,98
	adjetivar	2,75*	,896	,028	,17	5,33
	imaginar	,15	,896	1,000	-2,43	2,73

Basado en las medias observadas.

*. La diferencia de medias es significativa al nivel ,05.

Sólo los jóvenes : ANOVA 1

Descriptivos

palabras recordadas

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
contar	10	6,50	1,434	,453	5,47	7,53	4	9
rimar	10	7,60	1,955	,618	6,20	9,00	4	10
adjetivar	10	14,80	3,490	1,104	12,30	17,30	11	22
imaginar	10	17,60	2,591	,819	15,75	19,45	14	22
recordar	10	19,30	2,669	,844	17,39	21,21	15	22
Total	50	13,16	5,787	,818	11,52	14,80	4	22

ANOVA

palabras recordadas

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1353,720	4	338,430	53,064	,000
Intra-grupos	287,000	45	6,378		
Total	1640,720	49			

Prueba de homogeneidad de varianzas

palabras recordadas

Estadístico de Levene	gl1	gl2	Sig.
2,642	4	45	,046

Comparaciones múltiples

Variable dependiente: palabras recordadas

Bonferroni

(I) metjov	(J) metjov	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
contar	rimar	-1,100	1,129	1,000	-4,43	2,23
	adjetivar	-8,300*	1,129	,000	-11,63	-4,97
	imaginar	-11,100*	1,129	,000	-14,43	-7,77
	recordar	-12,800*	1,129	,000	-16,13	-9,47
rimar	contar	1,100	1,129	1,000	-2,23	4,43
	adjetivar	-7,200*	1,129	,000	-10,53	-3,87
	imaginar	-10,000*	1,129	,000	-13,33	-6,67
	recordar	-11,700*	1,129	,000	-15,03	-8,37
adjetivar	contar	8,300*	1,129	,000	4,97	11,63
	rimar	7,200*	1,129	,000	3,87	10,53
	imaginar	-2,800	1,129	,170	-6,13	,53
	recordar	-4,500*	1,129	,002	-7,83	-1,17
imaginar	contar	11,100*	1,129	,000	7,77	14,43
	rimar	10,000*	1,129	,000	6,67	13,33
	adjetivar	2,800	1,129	,170	-,53	6,13
	recordar	-1,700	1,129	1,000	-5,03	1,63
recordar	contar	12,800*	1,129	,000	9,47	16,13
	rimar	11,700*	1,129	,000	8,37	15,03
	adjetivar	4,500*	1,129	,002	1,17	7,83
	imaginar	1,700	1,129	1,000	-1,63	5,03

*. La diferencia de medias es significativa al nivel .05.

Sólo los mayores : ANOVA 1

Descriptivos

palabras recordadas

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
contar	10	7,00	1,826	,577	5,69	8,31	4	10
rimar	10	6,90	2,132	,674	5,38	8,42	3	11
adjetivar	10	11,00	2,494	,789	9,22	12,78	6	14
imaginar	10	13,40	4,502	1,424	10,18	16,62	9	23
recordar	10	12,00	3,742	1,183	9,32	14,68	5	19
Total	50	10,06	4,007	,567	8,92	11,20	3	23

ANOVA

palabras recordadas

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	351,520	4	87,880	9,085	,000
Intra-grupos	435,300	45	9,673		
Total	786,820	49			

Prueba de homogeneidad de varianzas

palabras recordadas

Estadístico de Levene	gl1	gl2	Sig.
2,529	4	45	,054

Comparaciones múltiples

Variable dependiente: palabras recordadas

Bonferroni

(I) metmay	(J) metmay	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
contar	rimar	,100	1,391	1,000	-4,01	4,21
	adjetivar	-4,000	1,391	,061	-8,11	,11
	imaginar	-6,400*	1,391	,000	-10,51	-2,29
	recordar	-5,000*	1,391	,008	-9,11	-,89
rimar	contar	-,100	1,391	1,000	-4,21	4,01
	adjetivar	-4,100	1,391	,051	-8,21	,01
	imaginar	-6,500*	1,391	,000	-10,61	-2,39
	recordar	-5,100*	1,391	,006	-9,21	-,99
adjetivar	contar	4,000	1,391	,061	-,11	8,11
	rimar	4,100	1,391	,051	-,01	8,21
	imaginar	-2,400	1,391	,913	-6,51	1,71
	recordar	-1,000	1,391	1,000	-5,11	3,11
imaginar	contar	6,400*	1,391	,000	2,29	10,51
	rimar	6,500*	1,391	,000	2,39	10,61
	adjetivar	2,400	1,391	,913	-1,71	6,51
	recordar	1,400	1,391	1,000	-2,71	5,51
recordar	contar	5,000*	1,391	,008	,89	9,11
	rimar	5,100*	1,391	,006	,99	9,21
	adjetivar	1,000	1,391	1,000	-3,11	5,11
	imaginar	-1,400	1,391	1,000	-5,51	2,71

*. La diferencia de medias es significativa al nivel .05.

Ejemplo 3

www.zoology.ubc.ca/.../ANOVA/ANOVA.html

La mariposa tropical *Heliconius erato* tiene un sabor desagradable que le proporciona una cierta protección de los pájaros. Éstos aprenden a reconocerlas para evitarlas. A su vez, para protegerse, el resto de las mariposas de una zona particular evolucionan para parecerse a las de mal sabor.

En América del sur existen diferentes formas de *Heliconius erato*. Localmente casi el 100% son de la misma forma.

En un estudio se tomaron mariposas de la forma "rayada" (más común al norte) y de la forma "cartero" (más común al sur) y se intercambiaron de zona midiendo posteriormente su supervivencia.

Mediante un ANOVA de dos factores se contrastaron las siguientes hipótesis:

H_0: La supervivencia media es igual en las dos zonas

H_0: La supervivencia media es igual para las dos formas (morph)

H_0: No hay interacción entre zona y forma

Tabla ANOVA

Source of Variation	SS	df	MS	F	P
Zone	9.05	1	9.05	0.965	0.327
Morph	34.553	1	34.55	3.685	0.056
Zone*morph	80.548	1	80.55	8.590	0.004
Error	1837.947	196	9.38		

Los autores concluyen:

So there is no mean difference in the life span in the two habitats, nor between the two morphs on average (although there is a tendency for there to be a difference between the morphs). There is however a significant interaction between zone and morphology. In other words the life span of a particular morph varies as a function of where it is, just as we imagined because of the putative function of the coloration.

Análisis de la varianza con tres factores

Modelo general

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + u_{ijk}, \quad \begin{matrix} i = 1, \ldots, I \\ j = 1, \ldots, J \\ k = 1, \ldots, k \end{matrix}$$

$u_{ijk} \sim \mathcal{N}(0, \sigma^2)$ independientes

Tenemos IJK + 1 parámetros desconocidos.
El número de datos debe superar el número de parámetros

Estudiaremos un modelo más sencillo aunque con importantes restricciones.

Análisis de la varianza con tres factores
Cuadrados latinos

Modelo:

$$Y_{ij(k)} = \mu + \alpha_i + \beta_j + \gamma_k + u_{ij(k)}, \qquad \begin{matrix} i = 1, \ldots, I \\ j = 1, \ldots, J \\ k = 1, \ldots, k \end{matrix}$$

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{k=1}^{K} \gamma_k = 0$$

$$u_{ijk} \sim \mathcal{N}(0, \sigma^2) \text{ independientes}$$

α_i es el *efecto fila*

β_j es el *efecto columna*

γ_k es el *efecto letra*

I = J = K

Análisis de la varianza con tres factores
Cuadrados latinos: diseño

Cada nivel de un factor se cruza sólo una vez con cada uno de los niveles de los otros factores

Se puede aplicar el diseño cuando:

1. Existen 3 factores
2. El número de niveles (I) es el mismo en cada factor
3. No hay interacción entre los factores

En primer lugar, se elige un cuadrado con I filas, I columnas e I letras de forma que no haya letras repetidas en ninguna fila ni en ninguna columna

Análisis de la varianza con tres factores

Cuadrados latinos: diseño

Cuadrado latino con I = 3 (12 diseños posibles)
Tres factores (fila, columna, letra)
Se rellena con n = 9 datos

Diseño

A	C	B
C	B	A
B	A	C

Datos

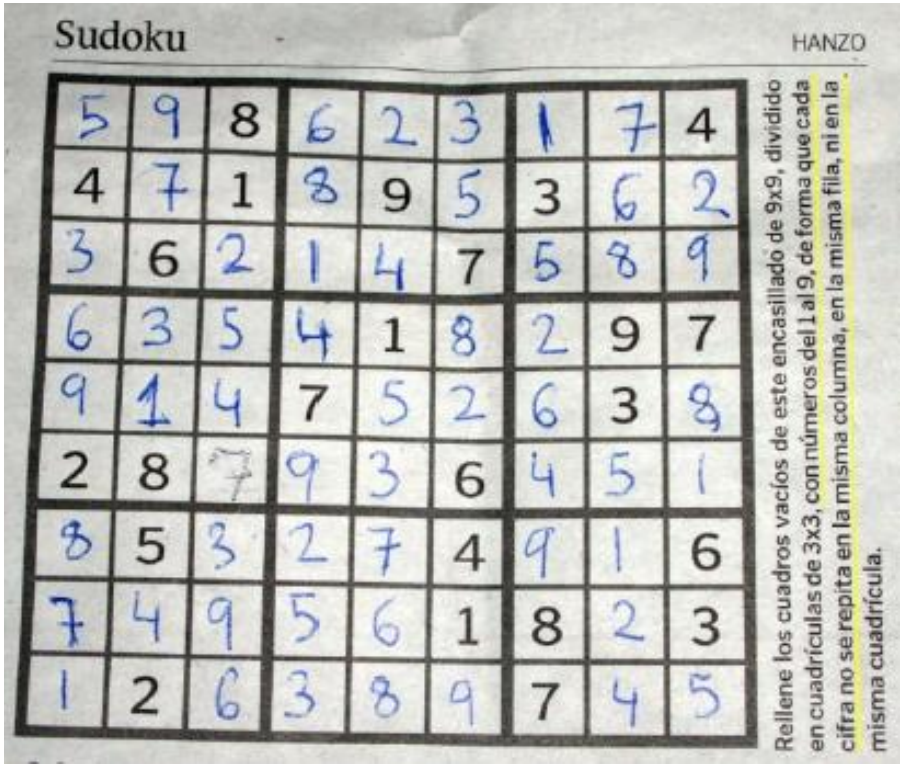
$y_{11(1)}$	$y_{12(3)}$	$y_{13(2)}$
$y_{21(3)}$	$y_{22(2)}$	$y_{23(1)}$
$y_{31(2)}$	$y_{32(1)}$	$y_{33(3)}$

Con I = 4 tendríamos 576 posibles diseños

Análisis de la varianza con tres factores

Cuadrados latinos: diseño

A veces es más sencillo con números en las celdas en vez de letras : SUDOKU



Sudoku

HANZO

5	9	8	6	2	3	1	7	4
4	7	1	8	9	5	3	6	2
3	6	2	1	4	7	5	8	9
6	3	5	4	1	8	2	9	7
9	1	4	7	5	2	6	3	8
2	8	7	9	3	6	4	5	1
8	5	3	2	7	4	9	1	6
7	4	9	5	6	1	8	2	3
1	2	6	3	8	9	7	4	5

Rellene los cuadros vacíos de este encasillado de 9x9, dividido en cuadrículas de 3x3, con números del 1 al 9, de forma que cada cifra no se repita en la misma columna, en la misma fila, ni en la misma cuadrícula.

6	3	2	7	8	1	9	4	5
4	8	7	5	9	6	2	1	3
5	1	9	2	4	3	8	7	6
8	6	4	3	5	2	7	9	1
7	5	1	9	6	8	3	2	4
2	9	3	1	7	4	6	5	8
9	4	5	6	3	7	1	8	2
1	7	6	8	2	5	4	3	9
3	2	8	4	1	9	5	6	7

Dato $y_{75(3)}$

Estos SUDOKU rellenos corresponden a dos diseños de cuadrado latino con I = 9

Existen más de 10^{21} soluciones diferentes para un SUDOKU 9x9.

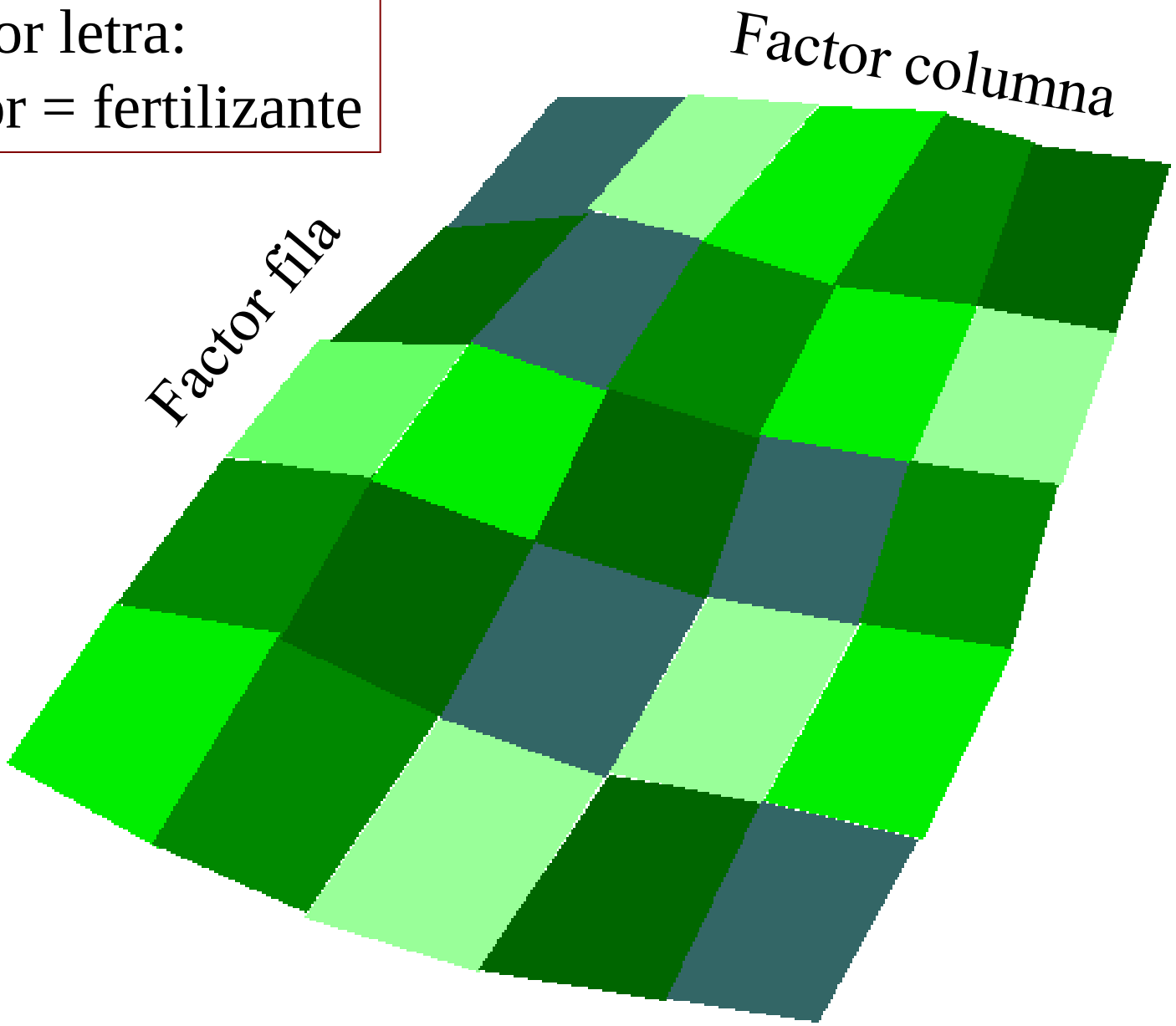
Ejemplo con I = 5

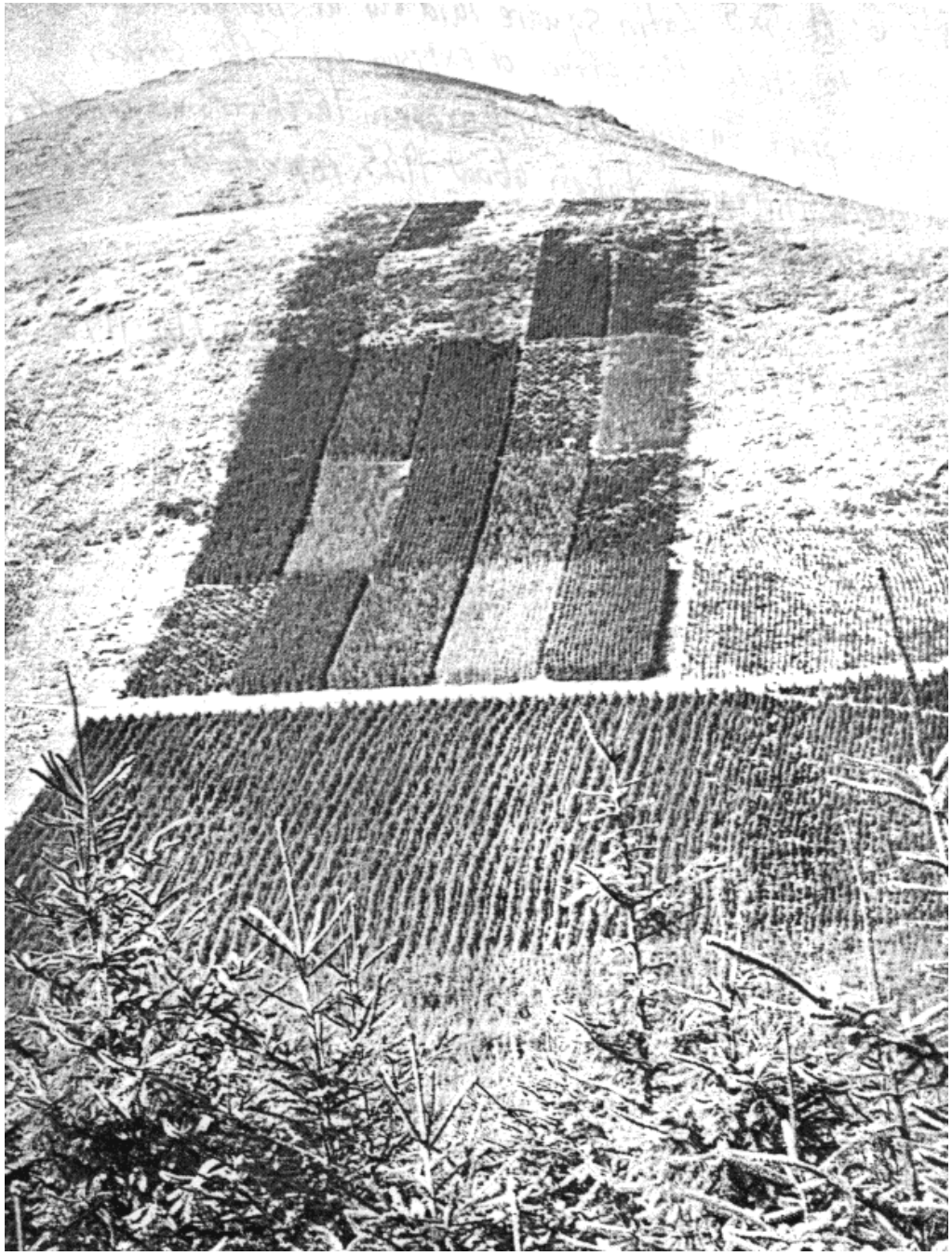
Supongamos que deseamos contrastar la eficacia de 5 fertilizantes diferentes sobre un cultivo de avena. Queremos aplicar los 5 fertilizantes, esperar a que la avena madure, recolectar y medir el resultado de la cosecha por unidad de superficie con cada fertilizante. Pero no podemos hacer los 5 experimentos en la misma tierra. Incluso terrenos contiguos pueden variar en fertilidad debido a múltiples causas (diferencias de humedad, uso previo del terreno, etc.) Dividimos el terreno experimental en una retícula de 5 x 5 rectángulos y en cada uno administramos un fertilizante (etiquetados al azar A, B, C, D, E)según el siguiente diseño de cuadrado latino:

A B C D E
B D A E C
C E D B A
D C E A B
E A B C D

www.math.sunysb.edu/.../latinI2.html

Factor letra:
Color = fertilizante





Un experimento real

A 5 x 5 Latin square laid out at Bettgelert Forest in 1929 to study the effect of exposure on Sitka spruce, Norway spruce (Abetos), Japaneses larch (Alerce), *Pinus contorta* and Beech (Haya). Photograph taken about 1945

Plate 6 from J F Box, *R.A. Fisher: The Life of a Scientist*, New York: Wiley 1978.

Análisis de la varianza con tres factores

Cuadrados latinos: estimación de los parámetros

Parámetros desconocidos del modelo (3I - 1) :

$$\mu, \alpha_1, \dots, \alpha_I, \beta_1, \dots, \beta_I, \gamma_1, \dots, \gamma_I, \sigma$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{I} Y_{ij(k)} = \bar{Y}_{...}$$

$$\hat{\alpha}_i = \frac{1}{I} \sum_{j=1}^{I} Y_{ij(k)} - \bar{Y}_{...} = \bar{Y}_{i..} - \bar{Y}_{...}$$

$$\hat{\beta}_j = \frac{1}{I} \sum_{i=1}^{I} Y_{ij(k)} - \bar{Y}_{...} = \bar{Y}_{.j.} - \bar{Y}_{...}$$

$$\hat{\gamma}_k = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{I} Y_{ij(k)} - \bar{Y}_{...} = \bar{Y}_{..k} - \bar{Y}_{...}$$

Tabla ANOVA

Fuente	s.c.	g.l.	varianzas	Test F	$p-$v.
Efecto fila	$I\sum_i \hat{\alpha}_i^2 = \text{SCE}(\alpha)$	I-1	$\hat{\boldsymbol{S}}^2_{\alpha} = \frac{\text{SCE}(\alpha)}{\text{I-1}}$	$\boldsymbol{F}_{\alpha} = \frac{\hat{S}^2_{\alpha}}{\hat{S}^2_{R}}$	¿?
Efecto columna	$I\sum_j \hat{\beta}_j^2 = \text{SCE}(\beta)$	I-1	$\hat{\boldsymbol{S}}^2_{\beta} = \frac{\text{SCE}(\beta)}{\text{I-1}}$	$\boldsymbol{F}_{\beta} = \frac{\hat{S}^2_{\beta}}{\hat{S}^2_{R}}$	¿?
Efecto letra	$I\sum_k \hat{\gamma}_k^2 = \text{SCE}(\gamma)$	I-1	$\hat{\boldsymbol{S}}^2_{\gamma} = \frac{\text{SCE}(\gamma)}{\text{I-1}}$	$\boldsymbol{F}_{\gamma} = \frac{\hat{S}^2_{\gamma}}{\hat{S}^2_{R}}$	¿?
Residual	$\sum_i\sum_j\sum_k e^2_{ij(k)} = \text{SCR}$	(I-1)(I-2)	$\hat{\boldsymbol{S}}^2_{R} = \frac{\text{SCR}}{\text{(I-1)(I-2)}}$		
Total	$\sum_i\sum_j\sum_k (Y_{ij(k)} - \bar{Y}_{...})^2$	I^2-1			

H_0 : todos los $\alpha_i = 0$, H_1 : algún $\alpha_i \neq 0$, $R = \left\{F_{\alpha} > F_{I-1,(I-1)(I-2),\alpha'}\right\}$

H_0 : todos los $\beta_j = 0$, H_1 : algún $\beta_j \neq 0$, $R = \left\{F_{\beta} > F_{I-1,(I-1)(I-2),\alpha'}\right\}$

H_0 : todos los $\gamma_k = 0$, H_1 : algún $\gamma_k \neq 0$, $R = \left\{F_{\gamma} > F_{I-1,(I-1)(I-2),\alpha'}\right\}$

Ejemplo 4

Ejemplo Se dispone de tres combustibles (a,b,c) y se trata de estudiar si el tipo de combustible influye en la reducción de emisiones de óxido de nitrógeno. Consideramos tres vehículos distintos (I,II,III) y tres conductores.

Datos:

	Conductor			$\bar{Y}_{i\cdot\cdot}$	$\hat{\alpha}_i$	
I	21 a	26 c	20 b	22.4	2.4	$\hat{\gamma}_1 =$ 21-20=1
II	23 b	26 a	20 c	23	3	$\hat{\gamma}_2 =$ 18.6-20=-1.4
III	15 c	13 b	16 a	14.6	-5.4	$\hat{\gamma}_3 =$ 20.4-20=0.4
$\bar{y}_{\cdot j\cdot}$	19.7	21.7	18.5			
$\hat{\beta}_j$	-0.3	1.7	-1.4			

Anova

Variable dependiente: emisiones

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
vehículo	128,667	2	64,333	6,226	,138
conductor	14,000	2	7,000	,677	,596
combustible	8,667	2	4,333	,419	,705
Error	20,667	2	10,333		
Total	172,000	8			

Ejemplo 5

En un estudio sobre percepción espacial con tres métodos diferentes de visión, se seleccionaron tres habitaciones, en cada habitación se pidió a 6 personas con las mismas características (edad, formación, sexo, etc.) que estimasen la medida de una de las tres dimensiones (longitud, anchura, altura) de la habitación utilizando uno de los tres métodos siguientes:

Visión real

las personas, sin nada, se pueden mover por la habitación

Visión con monitor de televisión

las personas ven a través de un monitor de televisión, situado fuera, que les permite ver la habitación desde distintos ángulos

Visión virtual

a las personas se les coloca un dispositivo de visión virtual con el que pueden moverse por la habitación

Las 48 personas se asignaron al azar a la habitación, a la dimensión y al método. La variable respuesta es el cociente entre las medidas reales y las estimadas por cada grupo de seis personas.

Factor fila Habitación (I, II, III)
Factor columna: Dimensión (L longitud W anchura, A altura)
Factor letra: Método (a = real, b = monitor, c = virtual)

Medidas reales (pies) Habitación \ Dimensión	L	W	A
I	23	18	14
II	48	19	14
III	47	28	20

Diseño Habitación \ Dimensión	L	W	A
I	a	b	c
II	b	c	a
III	c	a	b

Datos	L	W	A	Media fila
I	1.03 a	0.97 b	0.83 c	0.943
II	0.91 b	0.78 c	1.08 a	0.923
III	0.80 c	1.00 a	1.02 b	0.940
Media columna	0.913	0.917	0.977	Media total 0.936

Método	real	monitor	virtual
Media	1.037	0.967	0.803

Tabla ANOVA

Variable dependiente: estimación relativa

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
habitación	,001	2	,000	,301	,769
dimensión	,008	2	,004	3,330	,231
método	,086	2	,043	37,583	,026
Error	,002	2	,001		
Total	,097	8			

¿Conclusiones?