# Tema 5
# Análisis de la varianza

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

# Estructura de este tema

- Descripción del problema: ejemplo
- El modelo de análisis de la varianza con un factor
- Contraste de igualdad de medias. Descomposición de la variabilidad. Tabla ANOVA

# 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.

**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 encontrar diferencias significativas entre los 5 grupos. Es un problema más general que la comparación de dos medias del tema 3

Un **factor** es una variable cualitativa cuyos niveles sirven para establecer los grupos (en el ejemplo, el factor es el tipo de desayuno)

La **variable respuesta** es la variable cuantitativa que se mide en cada individuo (en el ejemplo, la respuesta es la nota)

# El modelo de análisis de la varianza con un factor

Queremos comparar $k$ muestras (en el ejemplo $k = 5$)

$$
\begin{array}{c|cccc|c|c}
\text{Muestra 1} & Y_{11} & Y_{12} & \cdots & Y_{1n_1} & \bar{Y}_{1.} & S_1 \\
\text{Muestra 2} & Y_{21} & Y_{22} & \cdots & Y_{2n_2} & \bar{Y}_{2.} & S_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\text{Muestra k} & Y_{k1} & Y_{k2} & \cdots & Y_{kn_k} & \bar{Y}_{k.} & S_k
\end{array}
$$

$\bar{Y}_{i.}$ es la media de la muestra $i$, $S_i$ es su desviación típica y $n_i$ es su tamaño

**Hipótesis:**

- Las muestras son independientes
- Proceden de $k$ poblaciones normales
- La muestra $i$ procede de una población normal de media $\mu_i$
- Las varianzas de las $k$ poblaciones son iguales. A la varianza común la llamamos $\sigma^2$.

# El contraste de igualdad de medias

El objetivo principal es estudiar si hay diferencias significativas entre las medias de los $k$ grupos

**Hipótesis nula** ($H_0$): las medias de los $k$ grupos son iguales ($\mu_1 = \cdots = \mu_k$)

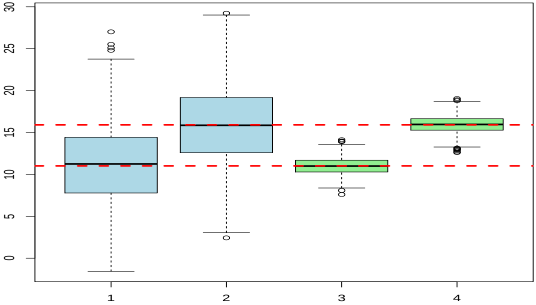
**Hipótesis alternativa** ($H_1$): no todas las medias son iguales

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

Si la variabilidad interna es muy grande, las diferencias entre los grupos pueden no ser significativas

# 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?

# Variabilidad entre y dentro de los grupos

Sea $\bar{Y}_{..}$ la media global:

$$\bar{Y}_{..} = \frac{n_1 \bar{Y}_{1.} + \cdots + n_k \bar{Y}_{k.}}{n},$$

donde $n = n_1 + \cdots + n_k$.

La **variabilidad entre los grupos** o variabilidad explicada se mide mediante:

$$\mathrm{SCEG} = \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

La **variabilidad dentro de los grupos** o variabilidad no explicada se mide mediante:

$$\mathrm{SCDG} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{Y}_{ij} - \bar{Y}_{i.})^2 = (n_1 - 1) S_1^2 + \ldots + (n_k - 1) S_k^2,$$

# Ejemplo

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 |

Por lo tanto,

$$\bar{Y}_{..} = \frac{(21 \times 7.17) + \cdots + (15 \times 6.18)}{467} = \frac{3118.98}{467} \approx 6.68$$

$$\text{SCEG} = 21(7.17 - 6.68)^2 + 125(6.84 - 6.68)^2 + \cdots + 15(6.18 - 6.68)^2 \approx 14.02$$

$$\text{SCDG} = 20 \times 1.74^2 + 124 \times 0.30^2 + \cdots + 14 \times 1.89^2 \approx 129.15$$

# Contraste

Para contrastar $H_0: \ \mu_1 = \cdots = \mu_k$, utilizamos el cociente entre SCEG y SCDG, corregido por grados de libertad.

Este cociente se llama **el estadístico F**:

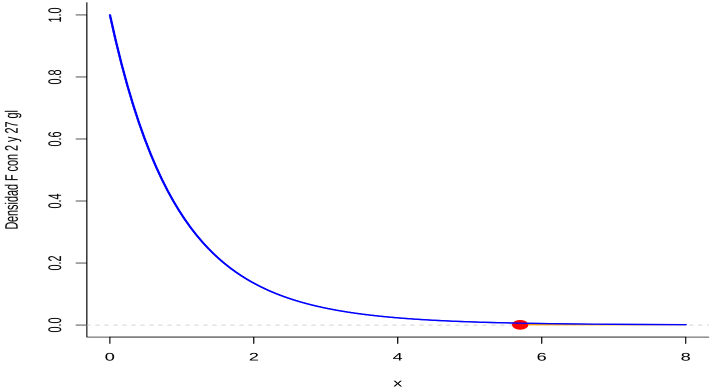
$$F = \frac{\frac{\text{SCEG}}{k-1}}{\frac{\text{SCDG}}{n-k}}$$

La hipótesis nula se rechaza a nivel $\alpha$ si $F$ es "suficientemente grande".

# La distribución $F$

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

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



# 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 $P(F_{5,10} > 5.636) = 0.01\cdot$

| | $n_1$ | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_2$ | 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 $\alpha$, el nivel de significación del contraste, es decir, la probabilidad de error al rechazar $H_0$.

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

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

# Ejemplo

Para los datos del ejemplo,

$$F = \frac{\frac{\text{SCEG}}{k-1}}{\frac{\text{SCDG}}{n-k}} = \frac{\frac{14.02}{4}}{\frac{129.15}{462}} \approx 12.54$$

$$F_{4,462,0.05} \approx 2.39$$

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.

# Tabla ANOVA

Toda la información necesaria para llevar a cabo el contraste se suele ordenar en forma de tabla:

| Fuente de variación | Suma de cuadrados | gl | cuadrados medios | estadístico |
|---|---|---|---|---|
| Entre grupos | $\sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$ | $k-1$ | $\frac{\sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2}{k-1}$ | $F$ |
| Dentro de los grupos | $\sum_{i=1}^{k}(n_i - 1)S_i^2$ | $n-k$ | $\frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{n-k}$ | |

Con los datos del ejemplo:

| Fuente de variación | Suma de cuadrados | gl | cuadrados medios | estadístico |
|---|---|---|---|---|
| Entre grupos | 14.02 | 4 | 3.50 | 12.54 |
| Dentro de los grupos | 129.15 | 462 | 0.28 | |

# Análisis de la varianza con jamovi

- Se ha considerado la cantidad de calorías y de sodio en salchichas de varias marcas de cada uno de los tipos siguientes:
  - Carne de ternera
  - Mezcla (hasta 15% de carne de pavo)
  - Carne de pavo

| Nombre variable | Descripción |
|---|---|
| `tipo` | Tipo de carne (1=ternera, 2=mezcla, 3=pavo) |
| `calorias` | Cantidad de calorías |
| `sodio` | Cantidad de sodio |

- Si $\mu_i$ es la cantidad media de sodio de las salchichas de cada tipo, vamos a contrastar $H_0 : \mu_1 = \mu_2 = \mu_3$

# ANOVA

ANOVA

| | Sum of Squares | df | Mean Square | F | p |
|---|---|---|---|---|---|
| tipo | 31739 | 2 | 15869 | 1.78 | 0.179 |
| Residuals | 455249 | 51 | 8926 | | |

[3]

## Estimated Marginal Means

### tipo

Estimated Marginal Means - tipo

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| tipo | Mean | SE | Lower | Upper |
| Ternera | 401 | 21.1 | 359 | 444 |
| Mezcla | 419 | 22.9 | 373 | 465 |
| Pavo | 459 | 22.9 | 413 | 505 |

[4]

# Cuestiones

- ¿Podemos rechazar la igualdad de las tres medias a nivel $\alpha = 0.05$?
- ¿Podemos rechazar la igualdad de las tres medias a nivel $\alpha = 0.1$?
- ¿Cuál es la fórmula del error típico y de los intervalos de confianza de la segunda tabla?

# Ejercicio: completar una tabla ANOVA

Completa la siguiente tabla ANOVA correspondiente a la comparación de 4 grupos con 4 observaciones por grupo:

| Fuente de variación | Suma de cuadrados | gl | Cuadrados medios | Test F |
|---|---|---|---|---|
| Entre grupos | | | | 5.91 |
| Dentro de los grupos | 54 | | | |
| Total | | | | |