Anova unifactorial I

Grado de Biología sanitaria

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

28 de noviembre de 2017

Contexto del problema

Relacionar dos variables

1. Variable explicativa (independiente)
2. Variable respuesta (dependiente)

variable respuesta ~variable explicativa

(11) **C ~ C** **Regresión lineal.**	(14) **F ~ C** **Regresión Logística. o multinomial.**
(12) **C ~ F** **Anova.**	(13) **F ~ F** **Contraste χ^2.**

1. $C \sim C$. en crabs, FL y CL
2. $F \sim C$. Sobrevivir o no frente a ingesta (mg) de cierta sustancia
3. $C \sim F$. Descenso T^a (en C^o) con distintos antitérmicos (niveles)
4. $F \sim F$. Ser creyente/no creyente frente a ser hombre/mujer

Ahora nos centramos en el caso **C ~ F**

El problema

Analizaremos la relación entre una variable cualitativa y una cuantitativa

$$C \text{ (respuesta)} \sim F \text{ (explicativa)}$$

Ejemplo: Se toman muestras de aire en 6 estaciones urbanas (cada una de tamaño 30) y se mide la concentración mdia de ozono en cada estación

- ¿Son las concrentraciones medias iguales?
- En caso contrario, ¿es posible ordenarlas?

En este caso:

- Variable explicativa (factor): estación en que se recoge la muestra.
  Niveles: cada una de las estaciones
- Variable respuesta: concentración de ozono

`capítulo 11 del libro`

El problema

Ejemplo: Cortesía del Hospital Ramón y Cajal Se quiere evaluar la eficacia de distintas dosis de un fármaco contra la hipertensión arterial y comparárla con la de una dieta sin sal.
Se seleccionan al azar 25 hipertensos y se distribuyen aleatoriamente en 5 grupos.

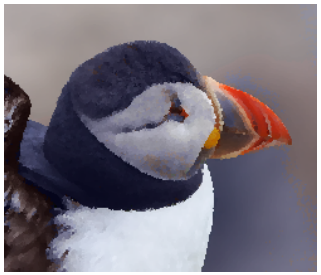
1. control (no tratamiento)
2. una dieta pobre en sal,
3. una dieta sin sal,
4. el fármaco a una dosis determinada y
5. el mismo fármaco a otra dosis.

Tras el experimento, ¿hay **diferencias significativas** en tensión arterial media de cada grupo?

En este caso:

- Variable explicativa (factor): tipo de tratamiento seguido (incluido el no-tratamiento). Niveles: cada uno de los tratamientos
- Variable respuesta: tensión arterial (en las unidades adecuadas)

El problema



Se quiere determinar cuál de los remedios

Alirón plus Vuelagra Plumiprofeno Elevantolín

es mejor para mejorar aleteos por minuto en los frailecillos.

Se toman 4 m.a. independientes de 100 frailecillos, y cada una se trata con un remedio diferente. Los resultados, en aleteos por minuto de cada individuo, están an la siguiente tabla

	Aliron	Elevantolin	Plumiprofeno	Vuelagra
1	76.65	88.66	87.14	76.74
2	79.36	78.12	82.34	74.72
3	71.83	81.74	94.06	68.61
4	73.24	89.11	88.12	72.84
5	79.73	82.90	84.47	75.83
6	74.50	80.84	83.11	66.81

Recuerda que este **no** es el mejor formato para los datos. una referencia y los datos de los frailecillos

El problema

La pregunta que queremos responder es

¿Qué tratamiento es mejor?

La responderemos en dos etapas:

1.- ¿Son todos los tratamientos igual de efectivos? Si μ_i es el número de aleteos medio con cada tratamiento

$$H_0 : \{\mu_1 = \mu_2 = \mu_3 = \mu_4\}$$

frente a

$$H_1 : \{\text{alguna de las medias es diferente de las demás}\}$$

Se trata de la comparación simultánea de 3 o más medias

Hacer comparaciones 2 a 2 dispara la probabilidad de error de tipo I
PIZARRA

2 .- Si no son todas las medias iguales: **ordenarlas**

Un poco de notación

- Consideramos un factor con k niveles $t_1, t_2, \cdots, t_k$ (tratamientos)
- Los individuos de cada nivel representan k poblaciones **independientes**

$$X_1, X_2, \cdots, X_k$$

- Hay n_j datos para el tratamiento t_j. Si $n_j \neq n_i$ *Anova no equilibrado*.

Nivel del tratamiento (j de 1 a k)

	t_1	t_2	t_3	$\cdots$	t_k
Respuestas (i de 1 a n_j)	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1k}
	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2k}
	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$	$\cdots$	$x_{n_k k}$

Trabajaremos con **experimentos equilibrados** ($n_i = n_j, \forall i, j$)

- Llamaremos $X(i, j) = x_{ij}$ individuo i del nivel/tratamiento j
- $N = n_1 + n_2 + \cdots + n_k$

Un poco de notación

Consideramos, además,

- La media total:

$$\bar{X} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{N}$$

- La media de cada nivel (tratamiento):

$$\bar{X}_{\bullet j} = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j}$$

Ejemplo: con los datos obtenidos para los 400 frailecillos:

$$\begin{cases} \bar{X}_{\cdot 1} = 78.40, & \text{respuesta media muestral para } t_1, \text{ Alirón Plus} \\ \bar{X}_{\cdot 2} = 72.10, & \text{respuesta media muestral para } t_2, \text{ Vuelagra} \\ \bar{X}_{\cdot 3} = 84.40, & \text{respuesta media muestral para } t_3, \text{ Plumiprofeno} \\ \bar{X}_{\cdot 4} = 80.40, & \text{respuesta media muestral para } t_4, \text{ Elevantolín} \end{cases}$$

y la media muestral total es

$$\bar{X} = 78.82$$

La idea

Al comparar el efecto del tratamiento en cada individuo con el resultado medio

$$x_{ij} - \bar{X}$$

intervienen (al menos) dos aspectos para cada individuo

- el tratamiento concreto (nivel del factor) recibido (modelo)
- las características individuales (azar)

cuantificar el efecto del azar y el del modelo

Podemos escribir

$$x_{ij} - \bar{X} = \left(x_{ij} - \bar{X}_{\bullet j}\right) + \left(\bar{X}_{\bullet j} - \bar{X}\right)$$

resulta

- $\left(\bar{X}_{\bullet j} - \bar{X}\right)$ repuesta media al tratamiento frente a la respuesta media total: modelo
- $\left(x_{ij} - \bar{X}_{\bullet j}\right)$ respuesta de cada individuo frente a respuesta media al tratamiento (depende del individuo): azar. Este término se llama residuo

La identidad Anova

Identidad de la suma de cuadrados para Anova

$$\underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{X})^2}_{SST} = \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{X}_{\cdot j})^2}_{SS_{\text{residual}}} + \underbrace{\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j}-\bar{X})^2}_{SS_{\text{modelo}}} \qquad (1)$$

Es decir $SST = SS_{\text{residual}} + SS_{\text{modelo}}$.

- SST es la dispersión total (no se distinguen niveles, población única)
- SS_{residual} es la dispersión debida al **azar o ruido**. Se suele llamar **dispersión dentro de los grupos** o intra-grupo, porque se debe a las circunstancias individuales de cada aplicación de un nivel del tratamiento.
- SS_{modelo} dispersión atribuida al hecho de utlizar k tratamientos distintos. Es la **dispersión entre grupos**. También se dice que es la parte de la dispersión o de la varianza **explicada por el modelo**

¿qué sucede si $SS_{\text{residual}} = 0$?

Hay que cuantificar el peso de cada sumando

$$SST = SS_{\text{residual}} + SS_{\text{modelo}},$$

se dan los mismos problemas (sensibilidad a unidades, a cambios de escala, ...) que en la recta de regresión lineal. Solución: dividir entre SS_{residual}

$$\frac{SST}{SS_{\text{residual}}} = 1 + \frac{SS_{\text{modelo}}}{SS_{\text{residual}}}$$

Manipulando ese término, se llega a que

- Si las **muestras siguen una distribución normal**,
- Todas tiene **varianzas iguales (también se dice homocedasticidad)**

$$X_1 \sim N(\mu_1, \sigma), \quad X_2 \sim N(\mu_2, \sigma), \quad \cdots, X_k \sim N(\mu_k, \sigma)$$

el cociente se comporta como una F de Fisher-Snedecor

Distribución muestral de los componentes del Anova unifactorial para el caso de un modelo equilibrado.

Supongamos que la hipótesis nula

$$H_0 = \{\mu = \mu_1 = \mu_2 = \cdots = \mu_k\}$$

es cierta, que el diseño es equilibrado con k niveles del tratamiento ($n_1 = n_2 = \cdots = n_k$). Entonces:

$$\Xi = \frac{\frac{SS_{\text{modelo}}}{k-1}}{\frac{SS_{\text{residual}}}{N-k}} = \frac{n \cdot \frac{\sum_{j=1}^{k}(\bar{X}_{\cdot j} - \bar{X})^2}{k-1}}{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_{\cdot j})^2}{N-k}} \sim F_{k-1;N-k}$$

donde $F_{k-1;N-k}$ es la distribución de Fisher-Snedecor con $k-1$ y $N-k$ grados de libertad, N es el total de observaciones (como el diseño equilibrado es, $N = k \cdot n$) El p-valor del contraste es

$$P\left(F_{k-1;N-k} > \Xi\right)$$

La tabla Anova

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico	**p-valor**
SS_{modelo}	$n \cdot \sum_{j=1}^{k} (\bar{X}_{\cdot j} - \bar{X})^2$	$k-1$	$\frac{n \cdot \sum_{j=1}^{k} (\bar{X}_{\cdot j} - \bar{X})^2}{k-1}$	Ξ	$\mathbf{P}(\mathbf{F} > \ddot{})$
SS_{residual}	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{\cdot j})^2$	$N-k$	$\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{\cdot j})^2}{n-k}$		

Ejemplo: continuación del ejemplo de los frailecillos

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico	**p-valor**
SS_{modelo}	7896.76	4-1	2632.25	149.24	0.00001
SS_{residual}	6984.41	400-4	17.6373		

podemos rechazar la hipótesis nula

Terminología

El modelo descrito es

Anova equilibrado, unifactorial, completamente aleatorio y de efectos fijos

1. **equilibrado**: todas las muestras tiene el mismo tamaño
2. **unifactorial**: sólo tenemos en cuenta cómo depende X del tratamiento aplicado sin contar otras variables: edad, género, dieta, ...
3. **Completamente aleatorio** porque los pacientes son asignados de forma aleatoria a cada grupo, sin agruparlos de ninguna manera
4. **Efectos fijos** porque hemos seleccionado los tratamientos (niveles) que queremos analizar sin elegirlos al azar de entre un posible conjunto más amplio de tratamientos

Existen modelos Anova más avanzados: por ejemplo

- los Anova de doble o triple vía (Anova de dos o tres factores)
- diseños no equilibrados

Ver apéndice *Más allá de este libro*