Sir Ronald Aylmer Fisher (1890 UK, 1962 Australia)

Figura: El Anova fue desarrollado por el genetista R. A. Fisher en los años 1920 y 1930 (de Wikipedia)

Clase 26 de Bioestadística

Análisis de varianza (ANOVA)

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Análisis de varianza

Análisis de varianza

- El análisis de varianza (comunmente conocido como ANOVA por sus siglas en inglés), es un caso particular del modelo lineal
- En este caso las X_i son 0 o 1, indicando si un parámetro es necesario en el modelo (1) o no (0)

Ejemplo

- Supongamos que se quiere comparar el rendimiento de un cultivo (medido en kilogramos por hectárea) en tres tipos de suelo:
  - arenoso ($i = 1$),
  - arcilloso ($i = 2$),
  - limoso ($i = 3$)
- Tenemos (por ejemplo) 10 muestras de distintas parcelas de cada uno.
- En total tenemos 30 muestras

- La variable rendimiento Y puede depender del tipo de suelo usado, y de la parcela en cuestión,
- Tenemos entonces datos
  - $Y_{1,1}, \ldots, Y_{1,10}$ rendimientos para el tipo de suelo arenoso, para cada una de las 10 parcelas,
  - Tenemos $Y_{2,1}, \ldots, Y_{2,10}$ para el tipo de suelo arcilloso
  - Tenemos $Y_{3,1}, \ldots, Y_{3,10}$ para el tipo de suelo limoso.
- Supondremos que podemos modelar las $Y_{i,1}, \ldots, Y_{i,10}$ como

$$Y_{i,j} = \mu + \alpha_i + e_{i,j} \quad \text{para } j = 1, \ldots, 10,$$

  donde μ es una constante.
- Suponemos que los errores $e_{i,j}$ son independientes idénticamente distribuidos con distribución normal, todos ellos con media 0 y varianza σ^2.

- La hipótesis que podríamos querer contrastar es

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

contra

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

- Si *no* se rechaza H_0 tenemos indicios de que el rendimiento medio, en cada tipo de suelo, es similar, ya que suponemos $E(e_{i,j}) = 0$ y por lo tanto $E(Y_{i,j}) = \mu$.
- Es decir *no* hay diferencias en el rendimiento medio, según el tipo de suelo, si se rechaza hay evidencia estadística de que el suelo influye en el rendimiento medio.

Ensayos clínicos

- Otro ejemplo importante es el testeo del efecto de un medicamento
- Por ejemplo, el efecto de cierto medicamento sobre alguna variable medible Y (presión arterial, nivel de colesterol, etc).
- En este caso rechazar H_0 estaría significando que alguno de los medicamentos influye sobre el valor promedio de la variable Y que se estudia.
- Este puede ser un resultado deseado por la farmaceútica (eficacia del medicamento) como indeseado (efecto secundario)

Comentarios

- Es importante mencionar que en lo que haremos la hipótesis de que
  - los errores tienen distribución normal
  - los errores son independientes,

  es la base teórica del razonamiento siguiente.
- Veremos unicamente el análisis de varianza de una vía o factor: en el caso del suelo, no consideramos otras características que puedan influir en el rendimiento, más que el tipo de suelo.

Comentarios

Consideremos el modelo

$$Y_{i,j} = \mu + \alpha_i + e_{i,j} \quad (i = 1, \ldots, k, \quad j = 1, \ldots, n_i)$$

- Del factor $i = 1, \ldots, k$ (tipo de suelo, tipo de medicamento, etc), tenemos n_i datos,
- En total tenemos

$$n = \sum_{i=1}^{k} n_i$$

datos.

- La hipótesis que queremos contrastar es

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

contra la alternativa

$$H_1 : \text{para algún } i, \alpha_i \neq 0.$$

Comentarios

- Para definir el estadístico calculamos primero el promedio de la clase i:

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$$

- Luego calculamos el promedio general:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{i,j}$$

El estadístico es un cociente:

- En el numerador va la diferencia entre los promedios por clase y el promedio general:

$$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{Y_i} - \overline{Y})^2$$

  ponderado por n_i, la cantidad de mediciones de cada factor, es decir, la variabilidad inter-clases.

- En el denominador tenemos la variabilidad interna de las clases:

$$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{i,j} - \overline{Y_i})^2.$$

  o variabilidad intra-clases.

- El estadístico compara la variabilidad inter-clases con respecto a la variabilidad intra-clases

$$F_{k-1,n-k} = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{Y_i} - \overline{Y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{i,j} - \overline{Y_i})^2}. \quad (1)$$

- La idea es que si los factores tienen distinto promedio, el numerador debería superar al denominador y el estadístico ser ”grande”

La distribución de F

- Como siempre que construímos un estadístico, tenemos una intuición sobre su comportamiento en la hipótesis nula y en la alternativa
- En este caso, si todos los factores tienen igual distribución (los α_i son cero) el numerador y el denominador tenderían a ser parejos
- Por el contrario, si algún α_i es distinto de cero, el numerador tiende a superar al denominador
- En otros términos, si F es cercano a cero no rechazamos H_0, y si es grande, la rechazamos
- El problema, como siempre, es saber que quiere decir grande
- Y allí viene en nuestra ayuda la teoría

- Se asumen errores normales $\mathcal{N}(0, \sigma^2)$ e independientes
- Bajo estos supuestos
  - el numerador tiene distribución Chi cuadrado con $k - 1$ grados de libertad
  - el denominador tiene distribución Chi cuadrado con $n - k$ grados de libertad

  Los grados de libertad son la cantidad de sumandos menos los promedios que aparecen
  - en el numerador aparecen k sumandos y un promedio $\overline{Y}$
  - en el denominador aparecen n sumandos y k promedios $\overline{Y}$
- El cociente resulta con distribución F de Fisher-Snedecor con $k - 1$ y $n - k$ grados de libertad.

En R

En R, con `df1` y `df2` los grados de libertad, tenemos la distribución F con los comandos

- `df(x, df1, df2)` da la densidad de F en x
- `pf(q, df1, df2)` da la distribución de F en x
- `qf(p, df1, df2)` da el cuantil p
- `rf(n, df1, df2)` simula n variables F

Para la región crítica del test usamos entonces

```
qf(0.95, k-1,n-k)
```

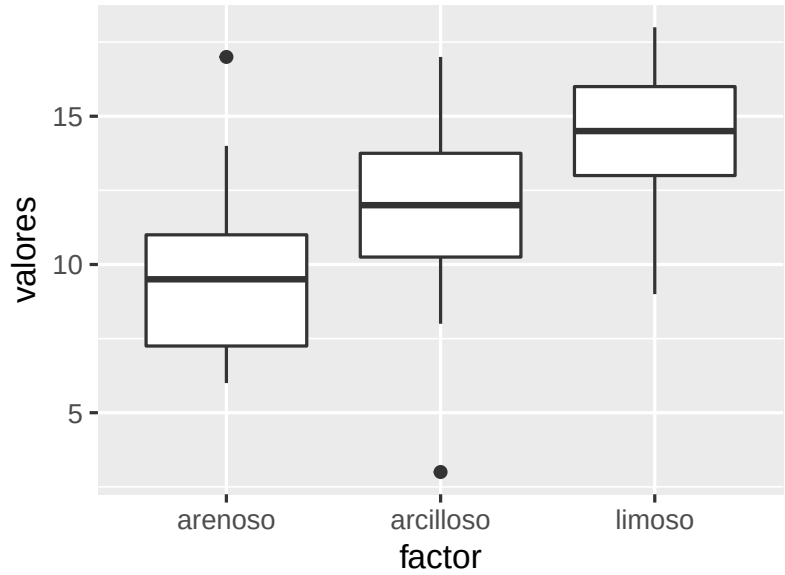
Ejemplo numérico

Vamos a ponerle números al ejemplo inicial. Supongamos que nuestros datos son[1]

- `arenoso<-c(6,10,8,6,14,17,9,11,7,11)`
- `arcilloso<-c(17,15,3,11,14,12,12,8,10,13)`
- `limoso<-c(13,16,9,12,15,16,17,13,18,14)`

[1] Tomado de las notas de A. Cholaquidis

Primero visualizamos los datos



Tenemos tres boxplots con los datos de cada factor (tipo de suelo)

Cálculo de F

Ahora calculamos el estadístico F paso por paso.
Comenzamos con el numerador

$$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$$

- Calculo los tres promedios de los rendimientos de cada suelo:

```
> x1<-mean(arenoso)
> x1
[1] 9.9
> x2<-mean(arcilloso)
> x2
[1] 11.5
> x3<-mean(limoso)
> x3
[1] 14.3
```

Para el numerador además preciso el promedio general

```
> total<-c(arenoso,limoso,arcilloso)
> xt<-mean(total)
> xt
[1] 11.9
```

Como la cantidad de datos $n_i = 10$ para los tres factores, el numerador de la F es:

```
> num<-10*((x1-xt)^2+(x2-xt)^2+(x3-xt)^2)/(3-1)
> num
[1] 49.6
```

Esta es la variabilidad inter-clases (between)

El denominador

$$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{i,j}-\overline{Y_i})^2.$$

Primero calculamos las tres sumas en azul:

```
> d1<-sum((arenoso-x1)^2)
> d1
[1] 112.9
> d2<-sum((arcilloso-x2)^2)
> d2
[1] 138.5
> d3<-sum((limoso-x3)^2)
> d3
[1] 64.1
```

Ahora calculamos el denominador

```
> den<-(d1+d2+d3)/(30-3)
> den
[1] 11.68519
```

Conclusión

- Tenemos nuestro estadístico F

```
> F<-num/den
> F
[1] 4.244691
> round(F,digits = 2)
[1] 4.24
```

- Este número: ¿es grande o pequeño?
- El estadístico tiene distribución F con $k - 1 = 2$ y $n - k = 27$ grados de libertad.
- Con una significancia de $\alpha = 0{,}05$ calculo el cuantil correspondiente:

```
> qf(0.95,2,27)
[1] 3.354131
```

- ¿Cuál es la conclusión? ¿Rechazo la hipótesis nula?

Sí, porque F es mayor que el cuantil

El R incluye el comando `aov` para hacer analisis de varianza. Primero hay que poner todos los datos en un `data.frame`

```
> datos<-data.frame(cbind(arenoso,arcilloso,limoso))
> datos
   arenoso arcilloso limoso
1        6        17     13
2       10        15     16
3        8         3      9
4        6        11     12
5       14        14     15
6       17        12     16
7        9        12     17
8       11         8     13
9        7        10     18
10      11        13     14
```

Sin embargo, para aplicar `aov` hay que ordenarlos todos en una pila (stack)

```
> datos2<-stack(datos)
> colnames(datos2)<-c("valores","factor")
> datos2
   valores    factor
1        6   arenoso
2       10   arenoso
3        8   arenoso
4        6   arenoso
5       14   arenoso
6       17   arenoso
7        9   arenoso
8       11   arenoso
9        7   arenoso
10      11   arenoso
11      17 arcilloso
12      15 arcilloso
13       3 arcilloso
14      11 arcilloso
15      14 arcilloso
16      12 arcilloso
17      12 arcilloso
```

Ahora aplicamos el comando `aov` y le pedimos el `summary`:

```
> summary(aov(valores~factor,data=datos2))
            Df Sum Sq Mean Sq F value Pr(>F)
factor       2   99.2   49.60   4.245  0.025 *
Residuals   27  315.5   11.69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

- R calcula todo lo que calculamos antes
- ¿Dónde está el *p*-valor?