Razón de correlación y ANOVA

Hamdi Raïssi

PUCV- Instituto de estadística

hamdi.raissi@pucv.cl

A muchas veces queremos explicar una variable cuantitativa con una variable cualitativa (nominal o ordinal).

1- Ciencias sociales (diferencias entre generos *(hombre/mujer : nominal, categoría social :ordinal...)*.
2- Diseño de experimentos *(test del efecto de un remedio o fertilizantes...).*
3- .... etc

En los diseños de experimentos el análisis de la ANOVA es importante.

Tenemos p muestras de tamaños $n_1, \ldots, n_p$ correspondientes cada uno a un nivel diferente de un factor A. El tamaño total es $n = n_1 + \cdots + n_p$.

Supuesto

En lo que sigue suponemos que tenemos la misma varianza en el grupo de observaciones.

El factor puede ser una dosis de insecticida,de remedio, hombre/mujer ...etc...

Cada muestra sería el résultado de la aplicación de un nivel de un factor A.

$\Rightarrow$ Queremos saber si existe un nivel del factor A que tendría un efecto sobre la población.

Comentario : Podemos hacer otras preguntas como la optimización de las dosis,...etc.

Hipotesis que queremos probar.

Probamos la igualdad de las p eperanzas

$$\begin{cases} H_0 : & m_1 = m_2 = \cdots = m_p. \\ H_1 : & \exists \quad i, j / m_i \neq m_j. \end{cases}$$

El modelo.

Anotamos X_{ik} la observation k de la muestra i, $i \in \{1, \ldots, p\}$ y $k \in \{1, \ldots, n_p\}$. Suponemos que

$$X_{ik} = m_i + \epsilon_{ik},$$

donde m_i es el promedio (teorico no observado) de la muestra i y ϵ_{ik}, el error.

Supuesto Gaussiano

$\epsilon_{ik} \sim N(0, \sigma^2)$ avec $\sigma > 0$.

El modelo.

El modelo en su forma matricial :

$$\begin{pmatrix} X_{11} \\ \vdots \\ X_{1n_1} \\ X_{21} \\ \vdots \\ X_{2n_2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots \\ 1 & \vdots & \vdots & \dots \\ \vdots & 1 & 0 & \cdots \\ 1 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \dots \\ 1 & \vdots & 1 & \cdots \\ 1 & 0 & 0 & \cdots \end{pmatrix} \begin{pmatrix} m_p \\ c_1 \\ \vdots \\ c_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{pn_p} \end{pmatrix}.$$

Nota : Tenemos p modalidades pero $p-1$ parámetros más el promedio m_p. Los c_i se intepretan como diferencias entre el promedio m_p y los promedios m_i : $c_i = m_i - m_p$.

Ejemplo con un factor que tiene 2 modalidades.

Sigue tratamiento/no sigue, Tratamiento fuerte/no fuerte, hombre/mujer, cargo empresa alto/bajo, etc..... Tamaños de muestras n_1 y n_2.

$$\begin{pmatrix} X_{11} \\ \vdots \\ X_{1n_1} \\ X_{21} \\ \vdots \\ X_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \begin{pmatrix} m_2 \\ c_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}.$$

Descomposición de la varianza.

Dado que no conocemos los m_i y m, tenemos que estimar los con $\bar{X}_i = n_i^{-1} \sum_{k=1}^{n_i} X_{ik}$ et $\bar{X} = n^{-1} \sum_{i=1}^{p} \sum_{k=1}^{n_i} X_{ik}$. Podemos anotar que

$$X_{ik} - \bar{X} = (\bar{X}_i - \bar{X}) + (X_{ik} - \bar{X}_i).$$

Tomando el cuadrado y sumando las observaciones, obtenemos :

$$\underbrace{\sum_{i=1}^{p} \sum_{k=1}^{n_i} (X_{ik} - \bar{X})^2}_{S_T^2} = \underbrace{\sum_{i=1}^{p} n_i (\bar{X}_i - \bar{X})^2}_{S_F^2} + \underbrace{\sum_{i=1}^{p} \sum_{k=1}^{n_i} (X_{ik} - \bar{X}_i)^2}_{S_R^2}.$$

Test.

Bajo H_0 y los supuestos gaussianidad y de homoscedasticidad, tenemos

$$\frac{S_T}{\sigma^2} \sim \chi^2_{n-1}, \qquad \frac{S_F}{\sigma^2} \sim \chi^2_{p-1}, \qquad \frac{S_R}{\sigma^2} \sim \chi^2_{n-p}.$$

⋆ Entonces tenemos :

$$T := \frac{S_F^2/p-1}{S_R^2/n-p} \sim F_{p-1,n-p}.$$

⋆ Toma de desición :
Si $T > F_{p-1,n-p,1-\alpha}$, rechazamos H_0 al nivel α.

Test de Fisher-Análisis de la varianza

Fuente	GDL	SC	SCP	Est. de Fisher
Modelo	1	SF	SF	F=SF/$\mathfrak{s}^2$
Error	n-2	SR	$\mathfrak{s}^2$=SR/(n-2)	
Total	n-1	ST	ST/(n-1)	

- GDL=degrees of freedom, SC=suma de los cuadrados, SCP=Suma de los cuadrados en promedio.
- Recuerde : Sean U y V dos variables aleatorias independientes chi-cuadradas χ^2_p y χ^2_q. Entonces la variable aleatoria :

$$Z = \frac{U/p}{V/q},$$

sigue una ley Fisher $F_{p,q}$.

Razón de correlación

- Objetivo : Medir la relación entre una variable cualitativa y una variable cuantitavia.
- Se define de la manera siguiente :

$$\hat{\eta}^2 = \frac{SF}{ST}.$$

- $\hat{\eta}^2$ es entre 0 y 1.
- Si estamos cerca 0 no hay relación fuerte.
- Si estamos cerca 1, existe relación fuerte.

Aplicación con `R` : ver el archivo recordatorio-correlacion-anova-R.txt.
Tarea : aplicar lo que hemos visto al conjunto de datos "`chickwts`" del paquete "`datasets`".

ANOVA con dos factores.

Sea X_{ijk} la observación k teniendo la modalidad i por el factor A y j por el factor B. ($i \in \{1, \ldots, p\}$ y $j \in \{1, \ldots, q\}$)
Sea $p \times q$ muestras de tamaño n correspondientes a las diferentes combinasiones de las modalidades de dos factores A y B.

- $\Rightarrow$ $\bar{X}_{i.}$ representa el promedio de las observaciones de modalidad i.
- $\Rightarrow$ $\bar{X}_{ij}$ representa el promedio de las observaciones de modalidades i y j por los factores A y B.

ANOVA con dos factores.

De la misma manera que por la descomposición de la varianza donde tenemos 1 factor, obtenemos por 2 factores :

$$\begin{aligned}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(X_{ijk}-\bar{X})^2 &= nq\sum_{i=1}^{p}(\bar{X}_{i.}-\bar{X})^2+np\sum_{j=1}^{q}(\bar{X}_{.j}-\bar{X})^2\\ &+ n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{X}_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X})^2\\ &+ \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(X_{ijk}-\bar{X}_{ij})^2.\end{aligned}$$

$$SC_T^2 = SC_A^2 + SC_B^2 + SC_{AB}^2 + SC_R^2.$$

ANOVA con dos factores.

El termino SC^2_{AB} mide la interacción entre los dos factores :

Este termino es cero cuando las variaciones del primer factor son independientes de las modalidades del segundo factor :

$$\bar{X}_{ij} - \bar{X}_{.j} = \bar{X}_{i.} - \bar{X}$$

y al revés...

- Si la interacción es cero, entonces se dice que el modelo de análisis de la varianza es additivo.

ANOVA con dos factores.

Se puede mostrar :

$\frac{SC_A^2/p-1}{SC_R/pq(n-1)} \sim F_{p-1,pq(n-1)} \Rightarrow$ Test del efecto del factor A.

$\frac{SC_B^2/q-1}{SC_R/pq(n-1)} \sim F_{q-1,pq(n-1)} \Rightarrow$ Test del efecto del factor B.

$\frac{SC_{AB}^2/(p-1)(q-1)}{SC_R/pq(n-1)} \sim F_{(p-1)(q-1),pq(n-1)} \Rightarrow$ Test de la interacción entre los dos factores.

ANOVA con tres factores (empieza a ser dificil interpretar las interacciones... !)

De la misma manera que la ANOVA con uno o dos factores podemos escribir en el caso donde tenemos tres factores A,B y C :

$$SC_T^2 = SC_A^2 + SC_B^2 + SC_C^2 + SC_{AB}^2 + SC_{AC}^2 + SC_{BC}^2 + SC_{ABC}^2 + SC_R^2.$$

$\Rightarrow$ Podemos probar los efectos de los diferentes factores y sus interacciones.