Introduction to ANOVA

Using python.

Iván Andrés Trujilllo Abella

Facultad de Ingenieria
Pontificia Universidad Javeriana

trujilloiv@javeriana.edu.co

Preamble

For this lesson you need remember some concepts as **Probability Distribution Function(PDF)**, **joint distribution function**, **expectation** and **variance** of random varaibles please check the following material:

- *Introduction to probability*
- *Introduction to hypothesis testing*

This note we are constructed with several references that are listed in references.

IDEA

We can study the variance among:

Importance of school

Assume that in your locality there are k schools with n_k the score of the $j-th$ student in $i-th$ school is y_{ij} students and $\bar{y}$ is global mean.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 \tag{1}$$

Important identities

...

$$y_i = \sum_{j=1}^{n_i} y_{ij} \tag{2}$$

$$\bar{y}_i = \frac{y_i}{n_i} \tag{3}$$

$$y = \sum_{i=1}^{k} \sum_{j}^{n_i} y_{ij} \tag{4}$$

$$\bar{y} = \frac{y}{\sum_{i=1}^{k} n_i} = \frac{y}{N} \tag{5}$$

Descomposition

Take in mind that y_i is the score mean in $i-th$ school.

$$
\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((y_{ij}-\bar{y}_i)^2+2(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})+(\bar{y}_i-\bar{y})^2\right)
\end{aligned} \tag{6}
$$

let $\sum_{j=1}^{n_i}$ and studying $\sum_{j=1}^{n_i} 2(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})=0$.

Descomposition

$$\sum_{j=1}^{n_i} 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = 2\sum_{j=1}^{n_i}(y_{ij}\bar{y}_i - y_{ij}\bar{y} - \bar{y}_i\bar{y}_i + \bar{y}_i\bar{y}) \tag{7}$$
$$= n_i\bar{y}_i\bar{y}_i - n_i\bar{y}_i\bar{y} - n_i\bar{y}_i\bar{y}_i + n_i\bar{y}_i\bar{y} = 0$$

...

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 \tag{8}$$

Descomposition

...

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2 \tag{9}$$

The variability is decomposed in inter and intra variability, therefore the school importance is:

...

$$\rho = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2} \tag{10}$$

ρ is R^2

...

$$y_i = \beta_0 + \beta_1 school_j \tag{11}$$

You can get the same result getting the coefficient of determination in the before model.

Table

Take in mind that $\tau_i = y_i - \bar{y}$ and $u_{ij} = y_{ij} - \bar{y}_i$.

Treatment	y_{ij}	$\bar{y}_i$	τ_i	u_{ij}	$\bar{y} + \tau_i + u_{ij}$
A	15	15	5,875	0	15
A	17	15	5,875	2	17
A	13	15	5,875	-2	13
B	9,3	8	-1,125	1,3	9,3
B	7,7	8	-1,125	-0,3	7,7
B	7	8	-1,125	-1	7
C	2,5	2	-7,125	0,5	2,5
C	1,5	2	-7,125	-0,5	1,5
$\bar{y}$	9,125				

The model

Especification

$$y_{ij} = \mu + \tau_i + u_{ij} \tag{12}$$

Where μ is the global mean, τ_i is deviation of the mean of factor i regarding with the global mean and finally u_{ij} is devation from each observation to its factor or treatment i.

Assumptions

...

- $u_{ij} \sim N(0, \sigma^2)$
- σ^2 is constant

Is important check the hypothesis given this allow us to get relaibles results.

MLE estimation

...

- The unknown parameters are $\mu_1, \mu_2, ...\mu_k$ for k treatments and σ^2.
- We are going uses the observed data and Maximun Likelihood Estimation (MLE).

MLE

The parameters of the model are considered fixed (μ and τ_i).

$$u_{ij} \sim N(0, \sigma^2) \implies y_{ij} \sim N(\mu_i, \sigma^2) \tag{13}$$

Maximun likelihood function

...

$$\mathcal{L}(\mu_1, ..., \mu_k, \sigma^2) = \prod_{i}^{k} \prod_{j}^{n_i} f(y_{ij}) \quad (14)$$

Now $f(y_{ij})$ is the PMF.

...

$$f(y_{ij}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_{ij} - \mu_i)^2}{2\sigma^2}\right) \quad (15)$$

Properties

...

- $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\lambda x_{ij} = \lambda\sum_{i=1}^{k}\sum_{j=1}^{n_i}$
- $\prod_{i=1}^{k}\prod_{j=1}^{ni}\lambda x_{ij} = \lambda^{\sum_{i=1}^{k} n_i}\prod_{i=1}^{k}\prod_{j=1}^{n_i} x_{ij}$
- $\prod_{i=1}^{k}\prod_{j=1}^{n_i} e^{\delta_{ij}} = e^{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\delta_{ij}}$

Properties

This propertie is intutive given the linearity of all terms: $(\lambda x_{11}, ..., \lambda x_{kn_k}) = \lambda(x_{11}, ..., x_{kn_k})$

...

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}\lambda x_{ij} =& \sum_{i=1}^{k}(\lambda x_{i1}, \lambda x_{12}, ..., \lambda x_{in_i}) \\
=& \sum_{i=1}^{k}\lambda\sum_{j=1}^{n_i}x_{ij} = \lambda\left(\sum_{j=1}^{n_1}x_{1j} + \sum_{j=1}^{n_2}x_{2j}+, ..., +\sum_{j=1}^{n_k}x_{kj}\right) \\
=&\lambda\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}.
\end{aligned} \tag{16}$$

Properties

...

$$
\begin{aligned}
\prod_{i=1}^{k}\prod_{j=1}^{n_i}\lambda x_{ij} &= \prod_{i=1}^{k}(\lambda x_{i1}\lambda x_{i2},\ldots,\lambda x_{in_i}) \\
&= \prod_{i=1}^{k}\lambda^{n_i}(x_{i1}x_{i2},\ldots,x_{in}) = \prod_{i=1}^{k}\lambda^{n_i}\prod_{j=1}^{n_i}x_{ij} \\
&= \lambda^{n_1}\left(\prod_{j=1}^{n_1}x_{1j}\right)\lambda^{n_2}\left(\prod_{j=1}^{n_2}x_{1j}\right),\ldots,\lambda^{n_k}\left(\prod_{j=1}^{n_k}x_{kj}\right) \\
&= \lambda^{\sum_{i=1}^{k}n_i}\prod_{j=1}^{n_1}x_{1j}\prod_{j=1}^{n_2}x_{2j},\ldots,\prod_{j=1}^{n_k}x_{kj} = \lambda^{\sum_{i=1}^{k}n_i}\prod_{i=1}^{k}\prod_{j=1}^{n_i}x_{ij}
\end{aligned} \tag{17}
$$

Properties

...

$$
\begin{aligned}
\prod_{i=1}^{k}\prod_{j=1}^{n_i} e^{\delta_{ij}} &= \prod_{i=1}^{k} e^{\delta_{i1}} e^{\delta_{i2}}, \ldots, e^{\delta_{in_i}} \\
&= \prod_{i=1}^{k} e^{\sum_{j=1}^{n_i}\delta_{ij}} = e^{\sum_{j=1}^{n_1}\delta_{1j}} e^{\sum_{j=1}^{n_2}\delta_{2j}}, \ldots, e^{\sum_{j=1}^{n_k}\delta_{kj}} \\
&= e^{\sum_{j=1}^{n_1}\delta_{1j}+\sum_{j=1}^{n_2}\delta_{2j},\ldots,\sum_{j=1}^{n_k}\delta_{kj}} \\
&= e^{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\delta_{ij}}
\end{aligned}
\tag{18}
$$

Applying the properties

...

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{k}\prod_{j}^{n_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y_{ij}-\mu_i)^2}{2\sigma^2}\right)\\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j}^{n_i}(y_{ij}-\mu_i)^2\right)\end{aligned} \tag{19}$$

Log likelihood function

...

$$\ln(\mathcal{L}) = \frac{-N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j}^{n_i}(y_{ij} - \mu_i)^2 \tag{20}$$

FCO

$$\frac{\partial \ln(\mathcal{L})}{\partial \mu_i} = 0, \forall i. \quad \frac{\partial \ln(\mathcal{L})}{\partial \sigma^2} = 0 \tag{21}$$

FCO

Hint: uses the propertie 'derivate of sum is the sum of derivatives' and rule chain.

....

$$\frac{\partial \ln(\mathcal{L})}{\partial \mu_i} = \frac{-1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)(-1) = 0 \tag{22}$$

Now to get the estimator of μ_i

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i) = 0 \tag{23}$$

Solving $\sum_j^{n_i}$

$$\sum_{i}^{k} (y_i - n_i \mu_i) = 0 \tag{24}$$

FCO

...

For each i we have:

$$y_i = n_i \mu_i \tag{25}$$

And therefore the estimator of μ_i is $\bar{y}_i$.

FCO

To find $\frac{\partial \mathcal{L}}{\partial \sigma^2} = 0$, remember that $\frac{\partial \ln f(x)}{\partial x} = \frac{f'(X)}{f(x)}$ and $\frac{\partial \frac{1}{x}}{\partial x} = \frac{-1}{x^2}$.

...

$$\frac{\partial \ln(\mathcal{L})}{\partial \sigma^2} = \frac{\partial \left(\frac{-N}{2} \ln(2\pi\sigma^2)\right)}{\partial \sigma^2} - \frac{\partial \left(\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - u_i)^2\right)}{\partial \sigma^2}$$

$$0 = \frac{-N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

$$0 = -N + \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2}{N}$$

Biased estimator of variance

Proof that is biased

$$E(\hat{\sigma}^2) \neq \sigma^2 \tag{26}$$

Model

specification

$$y_{ij} = \mu + \tau_i + u_{ij} \tag{27}$$

where:

$$\begin{aligned} \tau_i &= (\mu_i - \mu) \\ u_{ij} &= (y_{ij} - \mu_i) \end{aligned} \tag{28}$$

Note:

...

$$\mu_i = \mu + \tau_i \tag{29}$$

This equation will be important to state the hypothesis in our data.

Model

$$\begin{aligned} \hat{u}_{ij} &= (y_{ij} - \hat{\mu}_i) \\ &= (y_{ij} - \bar{y}_i) \\ &= e_{ij} \end{aligned} \tag{30}$$

Residuals

The residuals e_{ij} measure the variability no explained by the model.

Variance of residuals

...

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(e_{ij})^2}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(e_{ij} - \bar{e})^2}{N} \tag{31}$$

Proof that $\bar{e} = 0$.

Unbiased estimator

...

$$\hat{S}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (e_{ij})}{N - k} \tag{32}$$

Residuals are not independent

If you have N residuals and compute k means then there are $N - k$ independent residuals.

Proof $S \sim \chi^2$

IN construction

Quasi-variances

...

$$\hat{S}^2 = \frac{1}{N-k}\sum_{i=1}^{K}(n_i - 1)\hat{S}_i^2 \qquad (33)$$

Descomposition

...

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 \tag{34}$$

...

$$\text{Total variability(TV)} = \text{No Explained variability } + \text{Explained Variability} \tag{35}$$

Note that:

$$\hat{S}^2 = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(y_{ij}-\bar{y}_i)^2}{n-k} \tag{36}$$

-

$$E\left(\frac{VNE}{N-k}\right)=\sigma^2 \tag{37}$$

$$E\left(\frac{VE}{k-1}\right)=\sigma^2+\sum_{i=1}^{k} n_i\tau_i^2 \tag{38}$$

ANOVA

...

$$
\begin{aligned}
H_0 &: \tau_1 = \tau_2 = \ldots \tau_k = 0 \\
H_a &: \tau_i \neq 0 \quad \text{for any } i = 1, \ldots K
\end{aligned} \tag{39}
$$

How works?

Total variation = Explained variation + Unexplained variation..

- The total sum of squares (SST)
- Sum of Squared errors (SSE)
- Residual Sum of Squares (RSS)

...

$$SST = SSE + RSS \tag{40}$$

ANOVA table

Source of variation	Sum of Squares	Degrees of Freedom	Mean sum of Squares
Intra-group(Explained)	RSS	k-1	$\frac{RSS}{k-1}$
Inter-group (No-explained)	SSE	N-k	$\frac{SSE}{N-k}$
Total	SST	N-1	$\frac{SST}{N-1}$

Levene test

insight

Used to test if variance are not equal

...

$$H_0 : \sigma_i^2 = \sigma_j^2 \quad \forall i, j \tag{41}$$

Therefore if **P value** is higher than 0.05 then there are homocedasiticy. The alternative hypotehis is that atleas two differ..

Summary

Anova validation

- independence
- Normal residuals
- homocedasticity

Check normality

..

- Shapiro test
- QQ-plot

Homocedasticity

Levene test

Not assumptions

Where assumption are not met then we need uses no parametric test..

Doctorado

Austin TEXAS neurocomputation. Research areas...

https://www.cs.utexas.edu/users/ai-lab/?evolution