HMMA 307 : Advanced Linear Modeling

Chapter 3 : ANOVA

COIFFIER OPHELIE GAIZI IBRAHIM LEFORT TANGUY

https://github.com/opheliecoiffier/CM_Anova

Université de Montpellier

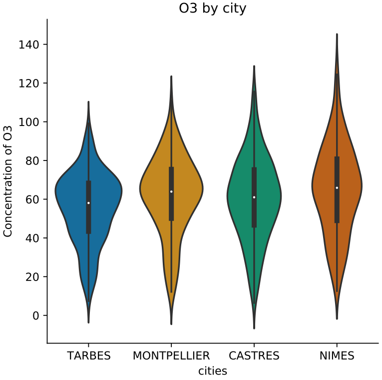
Statistical model for the ANOVA

ANOVA with the constraint $\sum \alpha_i^* = 0$

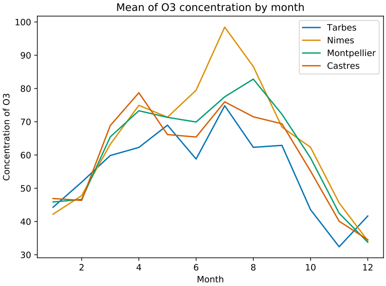
ANOVA with the constraint $\sum_{i=1}^{I} n_i \alpha_i = 0$

Non parametric alternative: permutation test

Comparison of the pollution between four cities



(a) Violin plot to compare the concentration of ozone between four cities in Occitanie.



(b) Mean of O3 by month for four cities.

Statistical model

Model equation

$$y_{ij} = \mu_i^* + \varepsilon_{ij}$$

- $\varepsilon_{ij} \overset{\textit{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ is the noise and $\mathrm{cov}(\epsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 \delta_{ii'} \delta_{jj'}$
- y_{ij} is the j^{th} measure for the modality
- $\bar{y}_n$ is the average of y *i.e.*,

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}; i \in [\![1, I]\!].$$

Statistical model

Model equation

$$y_{ij} = \mu_i^* + \varepsilon_{ij}$$

- $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ is the noise and $\operatorname{cov}(\epsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 \delta_{ii'} \delta_{jj'}$
- y_{ij} is the j^{th} measure for the modality
- $\bar{y}_n$ is the average of y *i.e.*,

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}; i \in [\![1, I]\!].$$

- We sometimes write : $\mu_i^* = \mu^* + \alpha_i^*$ to show the global mean effect and the specific effect of each feature.

Results from ANOVA and normality hypothesis

```
poll = ols('valeur_originale ~ C(nom_com)',data=df).fit()
sm.stats.anova_lm(poll, typ=2)
_, (__, ___, r) = sp.stats.probplot(poll.resid, fit=True)
```

Table: Results from the ANOVA on the O_3 concentration by cities.

	sum_sq	df	PR(>F)
C(nom_com)	16471.58	3	$3.86e^{-08}$

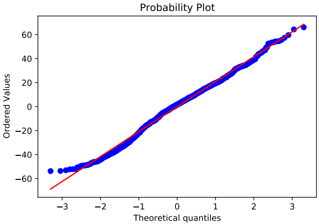


Figure: Check residues normality assumption

Rem: If we have an estimator for μ^* and α_i^* for all $i = 1, \dots I$, noted $\hat{\mu}$ and $\hat{\alpha}$:

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

and

$$(\hat{\mu}_1, \dots, \hat{\mu}_I) \in \underset{(\mu_1, \dots, \mu_I) \in \mathbb{R}^I}{\arg\min} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \tag{1}$$

Rem: If we have an estimator for μ^* and α_i^* for all $i = 1, \dots I$, noted $\hat{\mu}$ and $\hat{\alpha}$:

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

and

$$(\hat{\mu}_1, \dots, \hat{\mu}_I) \in \underset{(\mu_1, \dots, \mu_I) \in \mathbb{R}^I}{\arg\min} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \qquad (1)$$

Thanks to the separability principle:

$$\min_{(x_1, \dots, x_d)} f(x_1, \dots, x_d) \Longleftrightarrow \min_{x_j} g_j(x_j), \;\; j = 1, \dots, d.$$

we have

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{i,:}.$$

ANOVA : case of a modeling with : $\sum \alpha_i^* = 0$

Notice that if we change $\mu^* \longrightarrow \mu^* + \delta$ and $\alpha_i^* \longrightarrow \alpha_i^* - \delta$ then :
$\mu_i^* = (\mu^* + \delta) + (\alpha_i^* - \delta)$

- **hypothesis** : $\sum_{i=1}^{I} \alpha_i^* = 0$ *i.e.,* $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$

ANOVA : case of a modeling with : $\sum \alpha_i^* = 0$

Notice that if we change $\mu^* \longrightarrow \mu^* + \delta$ and $\alpha_i^* \longrightarrow \alpha_i^* - \delta$ then :
$\mu_i^* = (\mu^* + \delta) + (\alpha_i^* - \delta)$

- **hypothesis** : $\sum_{i=1}^{I} \alpha_i^* = 0$ *i.e.*, $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$
- **associated estimator** :

$$\underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I}{\arg\min} \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_1}(y_{ij} - \mu - \alpha_i)^2$$

ANOVA : case of a modeling with : $\sum \alpha_i^* = 0$

Notice that if we change $\mu^* \longrightarrow \mu^* + \delta$ and $\alpha_i^* \longrightarrow \alpha_i^* - \delta$ then :
$\mu_i^* = (\mu^* + \delta) + (\alpha_i^* - \delta)$

- **hypothesis** : $\sum_{i=1}^{I} \alpha_i^* = 0$ *i.e.*, $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$
- **associated estimator** :
$$\underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I}{\arg\min} \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_1}(y_{ij} - \mu - \alpha_i)^2$$
- **Lagrangian** :
$$\mathcal{L}(\mu, \alpha, \lambda) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \mu - \alpha_i)^2 + \lambda \sum_{i=1}^{I} \alpha_i$$

Resolution of the optimization system

$$\nabla \mathcal{L}(\hat{\mu}, \hat{\alpha}, \hat{\lambda}) = 0$$

$$\begin{cases} \sum_{i=1}^{I} \hat{\alpha}_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \hat{\mu}} = 0 \\ \frac{\partial \mathcal{L}}{\partial \hat{\alpha}_{i_0}} = 0 \ \forall i_0 \end{cases} \iff \begin{cases} \sum_{i=1}^{I} \hat{\alpha}_i = 0 \\ n\hat{\mu} + \sum_{i=1}^{I} n_i \hat{\alpha}_i - n\bar{y}_n = 0 \\ n_{i_0}\hat{\mu} + n_{i_0}\hat{\alpha_{i_0}} = n_{i_0}\bar{y}_{i_0,:} - \hat{\lambda} \end{cases}$$

$$\iff \begin{cases} \displaystyle\sum_{i=1}^{I} \hat{\alpha}_i = 0 \\ \hat{\mu} + \frac{1}{n} \sum_{i=1}^{I} n_i \hat{\alpha}_i = \bar{y}_n \\ n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:}) + \hat{\lambda} = 0 \end{cases}$$

Resolution of the optimization system

We have : $\sum_{i_0=1}^{I} n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:}) + I\hat{\lambda} = 0$, so for $i_0 = 1, \cdots, I$,
so we get

Resolution of the optimization system

We have : $\sum_{i_0=1}^{I} n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:}) + I\hat{\lambda} = 0$, so for $i_0 = 1, \cdots, I$, so we get

$$
\begin{aligned}
& \sum_{i_0=1}^{I} n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0}) + I\hat{\lambda} = 0 \\
\Longleftrightarrow \quad & n\hat{\mu} + \sum_{i_0=1}^{I} n_{i_0}\hat{\alpha}_{i_0} - \sum_{i_0=1}^{I} n_{i_0}\bar{y}_{i_0,:} + I\hat{\lambda} = 0 \\
\Longleftrightarrow \quad & n\hat{\mu} + \sum_{i_0=1}^{I} n_{i_0}\hat{\alpha}_{i_0} - n\bar{y}_n + I\hat{\lambda} = 0 \\
\Longleftrightarrow \quad & I\hat{\lambda} = 0 \Leftrightarrow \hat{\lambda} = 0
\end{aligned}
$$

Results:

- $\hat{\alpha}_{i_0} + \hat{\mu} = \bar{y}_{i_0,:}$
- $\hat{\mu} = \frac{1}{I} \sum_{i_0=1}^{I} \bar{y}_{i_0,:}$

Meaning that

$$\hat{\alpha}_{i_0} = \bar{y}_{i_0,:} - \frac{1}{I} \sum_{i_0=1}^{I} \bar{y}_{i_0,:}.$$

Notice:

- $\hat{\mu} \neq \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_n$
- It might be different if there are i, i' such that: $n_i \neq n_{i'}$

The weighted sum of the individual effects is zero

- **hypothesis** :

$$\sum_{i=1}^{I} n_i \alpha_i = 0$$

- **associated estimator** :

$$\underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I}{\arg\min} \ \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\mu-\alpha_i)^2$$

- **Lagrangian** :

$$\mathcal{L}(\mu,\alpha,\lambda) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\mu-\alpha_i)^2 + \lambda\sum_{i=1}^{I} n_i\alpha_i$$

Resolution of the optimization system

$$\nabla \mathcal{L}(\hat{\mu}, \hat{\alpha}, \hat{\lambda}) = 0$$

$$\left\{ \begin{array}{l} \sum_{i=1}^{I} n_i \hat{\alpha}_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \mu} = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha_{i_0}} = 0 \ \forall i_0 \end{array} \right. \iff \left\{ \begin{array}{l} \sum_{i=1}^{I} n_i \hat{\alpha}_i = 0 \\ n\hat{\mu} + \sum_{i=1}^{I} n_i \hat{\alpha}_i - n\bar{y}_n = 0 \\ \hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:} + \hat{\lambda} = 0, \forall i_0 \end{array} \right.$$

$$\iff \left\{ \begin{array}{l} \sum_{i=1}^{I} n_i \hat{\alpha}_i = 0 \\ \hat{\mu} = \bar{y}_n \\ \hat{\alpha}_{i_0} = \bar{y}_{i_0,:} - \hat{\lambda} - \bar{y}_n, \forall i_0 \end{array} \right.$$

Results :

- We multiply the third line of the equation by n_{i_0} then we add them up for i_0 in 1 to I. We finally obtain $\hat{\lambda} = 0$,
- $\hat{\mu} = \bar{y}_n$

Meaning that:

$$\hat{\alpha}_{i_0} = \bar{y}_{i_0,:} - \bar{y}_n.$$

Notice :

The next case to study will be:

$$\alpha_{i_0} = 0$$

Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,

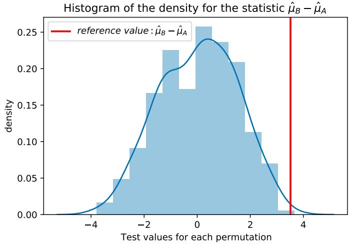


Figure: $\mu_A^* = 3,\ \mu_B^* = 7$, we reject the equality.

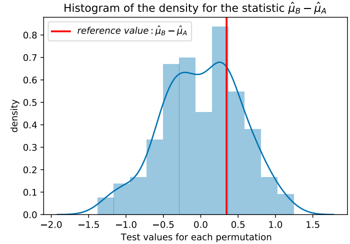


Figure: $\mu_A^* = 2,\ \mu_B^* = 2.5$, we don't reject the equality.

Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- $H0$: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),

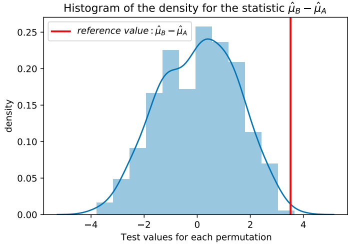


Figure: $\mu_A^* = 3, \ \mu_B^* = 7$, we reject the equality.

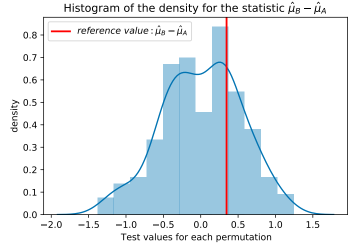


Figure: $\mu_A^* = 2, \ \mu_B^* = 2.5$, we don't reject the equality.

Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- $H0$: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,

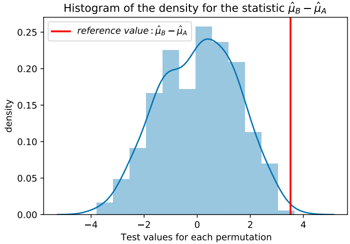


Figure: $\mu_A^* = 3, \ \mu_B^* = 7$, we reject the equality.

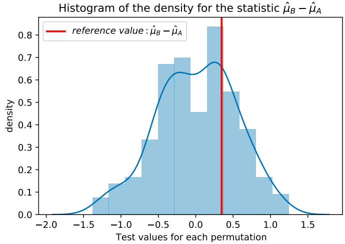


Figure: $\mu_A^* = 2, \ \mu_B^* = 2.5$, we don't reject the equality.

Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- $H0$: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,
- Get the reference statistic: $\hat{\mu}_B - \hat{\mu}_A$,

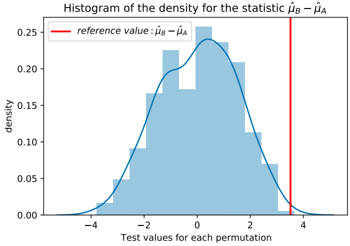


Figure: $\mu_A^* = 3, \ \mu_B^* = 7$, we reject the equality.

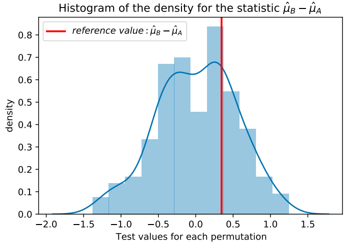


Figure: $\mu_A^* = 2, \ \mu_B^* = 2.5$, we don't reject the equality.

Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- $H0$: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,
- Get the reference statistic: $\hat{\mu}_B - \hat{\mu}_A$,
- shuffle the groups and recalculate the test statistic J times,

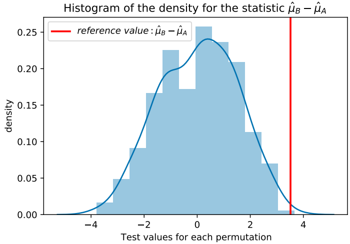


Figure: $\mu_A^* = 3,\ \mu_B^* = 7$, we reject the equality.

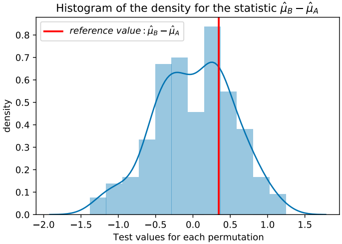


Figure: $\mu_A^* = 2,\ \mu_B^* = 2.5$, we don't reject the equality.

Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- $H0$: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,
- Get the reference statistic: $\hat{\mu}_B - \hat{\mu}_A$,
- shuffle the groups and recalculate the test statistic J times,
- p-value is the number of statistics over the reference divided by J.

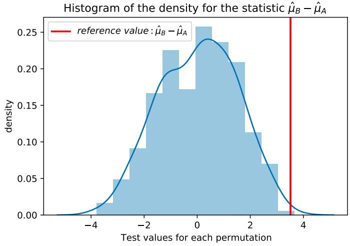


Figure: $\mu_A^* = 3,\ \mu_B^* = 7$, we reject the equality.

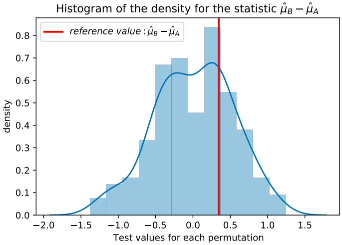


Figure: $\mu_A^* = 2,\ \mu_B^* = 2.5$, we don't reject the equality.

Bibliography

- Salmon, Joseph. *Modèle linéaire avancé : Anova*. 2019. URL: http://josephsalmon.eu/enseignement/Montpellier/HMMA307/Anova.pdf.
- Wilber, Jared. *Monte-Carlo method (permutation test)*. 2019. URL: https://www.jwilber.me/permutationtest/.