Analyse de la variance (ANOVA) à un facteur

Mathilde Boissel

22/10/2021

Table of Contents

Méthode

Test ANOVA

Hypothèses stochastiques

On étudie un caractère représenté par une variable $X \sim \mathcal{N}(\mu, \sigma^2)$. Les échantillons sont issus d'une population normale (gaussienne) : on parle de test **paramétrique.**
N.B.: l'ANOVA non paramétrique, avec le Test de Kruskal–Wallis, ne sera pas abordé dans ce cours.

On suppose que la population se divise en p sous-population $\mathcal{P}_1, \ldots, \mathcal{P}_p$.
Ainsi, $\forall i \in \{1, .., p\}$, la variable X considérée dans $\mathcal{P}_i$, est une variable

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

La moyenne μ, et a fortiori μ_i, sont inconnus.
Les variances conditionnelles (variances dans chaque sous-population) sont identiques : **homoscédasticité.**
Les données sont constituées de, $\forall i \in \{1, .., n_i\}$, la valeur de X_i pour chacun des n_i individus d'un échantillon de $\mathcal{P}_i$. Ces valeurs sont notées $x_{i,1}, \ldots, x_{i,n_i}$ pour X_i.
Les individus sont tous différents; les échantillons sont indépendants.

Enjeu

L'enjeu d'un test ANOVA est d'affirmer, avec un faible risque de se tromper, que $\mathcal{P}_1, \ldots, \mathcal{P}_p$ ne sont pas homogènes quant à μ.
Pour ce faire, on compare les moyennes $\mu_1, \ldots, \mu_p$ via un test statistique.

Notion de facteur

$\forall i \in \{1, .., p\}$, $\mathcal{P}_i$ peut être associée à une modalité A_i d'un caractère A appelé facteur.
Comparer $\mathcal{P}_1, \ldots, \mathcal{P}_p$ quant à μ revient à étudier l'influence du facteur A (caractérisé par ses p modalités $A_1, \ldots, A_p$) sur μ.
μ est ici la moyenne d'une variable (numerique) d'intérêt X.
Les données de X peuvent donc être mises sous la forme :

A_1	A_2	A_i	A_p
$x_{1,1}$	$x_{1,2}$	$x_{1,i}$	$x_{1,p}$
$x_{2,1}$	$x_{2,2}$	$x_{2,i}$	$x_{1,p}$
…	…	$x_{j,i}$	…
…	…	…	…
…	$x_{n_2,2}$	…	$x_{n_p,p}$

Hypothèses

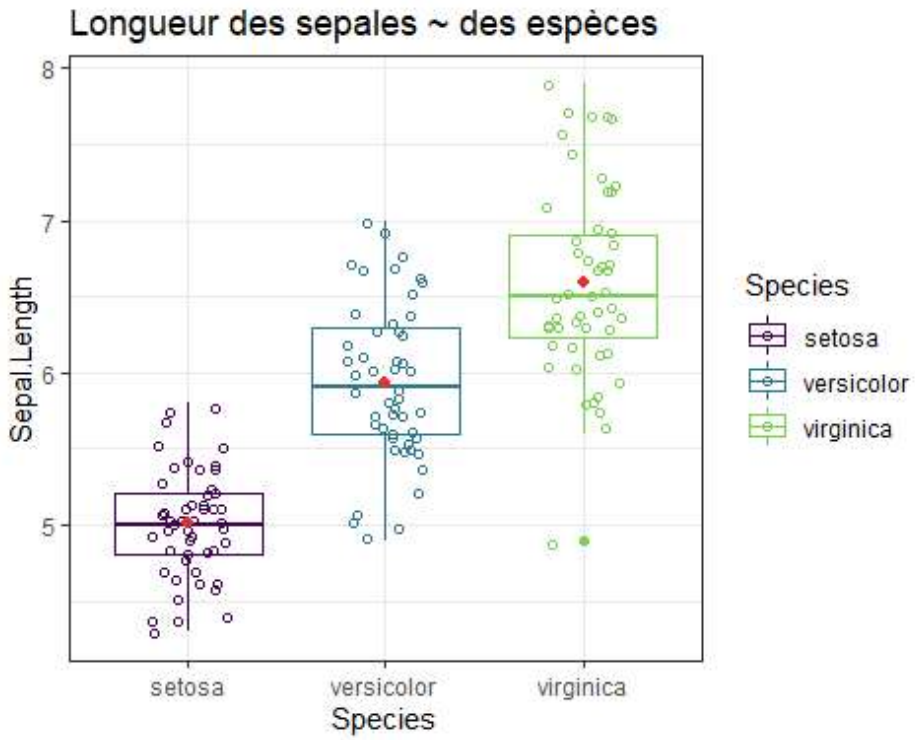
Les hypothèses associées au test ANOVA sont :

$\boldsymbol{H_0: \mu_1 = \ldots = \mu_p = \mu}$ **(*A* n'influe pas sur *X*)**
contre
$\boldsymbol{H_1: \exists j, \mu_j \neq \mu}$ **(il existe au moins 2 moyennes différentes, c-à-d *A* influe sur *X*).**

En d'autres mots, l'hypothèse nulle indique que la moyenne de la variable dépendante (X) est la même quelque soit les groupes (A_i) définis par le facteur (A). Sous H_0, le facteur A n 'a aucune influence sur la variable dépendante X.

Illustrations

Des représentations graphiques des données peuvent aider à appréhender la solution.



Objectif

On veut décider du rejet de H_0, au risque $\alpha/100, \alpha \in]0,1[$.

Test ANOVA

- Test statistique : On utilise alors le test ANOVA, lequel suppose que $\sigma_1^2, \ldots = \sigma_p^2$.
  Sa construction repose sur la loi de Fisher $\mathcal{F}(v_1, v_2)$.

- Calculs : Le test ANOVA se met en œuvre en calculant :
  - n : l'effectif total :
$$\sum_{i=1}^{p} n_i$$
  - $\overline{x}$: la moyenne totale :
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{i,j}$$
  - $\overline{x}_i, i \in \{1,..,p\}$: la moyenne de $x_{i,1},\ldots,x_{i,n_i}$:
$$\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j}$$
  - sce = sommes des carrés des écarts :
$$sce_T = sce_F + sce_R,$$
  - ddl = degrés de liberté :
$$ddl_T = ddl_F + ddl_R,$$
  - cm = carrés moyens cm_F et cm_R,
  - le f_{obs} défini par
$$f_{obs} = \frac{cm_F}{cm_R}.$$
  - le réel $f_\alpha(\upsilon_1, \upsilon_2)$ vérifiant
$$\mathbb{P}(F \geq f_\alpha(\upsilon_1, \upsilon_2)) = \alpha,$$

où $F \sim \mathcal{F}(\upsilon_1, \upsilon_2)$, $(\upsilon_1, \upsilon_2) = (ddl_F, ddl_R) = (p-1, n-p)$.
Si $\alpha = 0.05$, ce réel est évaluable dans la table ANNEXE 6 (Loi de Fisher II).

  - Le **tableau ANOVA** récapitule tout ceci :

	sce	ddl	cm	f_{obs}
Total	$sce_T = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{i,j} - \overline{x})^2$	$ddl_T = n-1$		
Factoriel	$sce_F = \sum_{i=1}^{p} n_i(\overline{x}_i - \overline{x})^2$	$ddl_F = p-1$	$cm_F = \frac{sce_F}{ddl_F}$	$f_{obs} = \frac{cm_F}{cm_R}$
Résiduel	$sce_R = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{i,j} - \overline{x}_i)^2$	$ddl_R = n-p$	$cm_R = \frac{sce_R}{ddl_R}$	

Tableau ANOVA

- Règles de décision : La règle de décision associée au Si $f_{obs} \geq f_{\alpha}(\upsilon_1, \upsilon_2)$,
  Alors on rejette H_0.

- p-valeurs : La p-valeur associée au test ANOVA est

$$p - valeur = \mathbb{P}(F \geq f_{obs}).$$

On peut alors déterminer le degré de significativité du rejet de H_0.
Par exemple, si $p - valeur = \mathbb{P}(F \geq f_{obs}) < 0.001$, alors le rejet de H_0 est hautement significatif (peut-être symbolisé par ***).

Diagnostique

Hypothèses stochastiques

Dans le même contexte décrit pour le test ANOVA.

Enjeu

On s'intéresse ici aux diagnostiques qui permettent d'attester que les résultats de l'anova sont valides. Quatre éléments sont à contrôler. Cela peut se faire graphiquement.

Hypothèses

On considère les hypothèses :

- **"Residuals vs Fitted"** et "**Constant Leverage: Residuals vs Factor Levels".**

  Les résidus sont indépendants. Les résidus ne doivent pas être corrélés entre eux. De la même façon, les résidus ne doivent pas être corrélés au facteur étudié. On peut faire le test de Dubin-Watson pour vérifier l'autocorrélation des résidus mais souvent un contrôle graphique suffit.

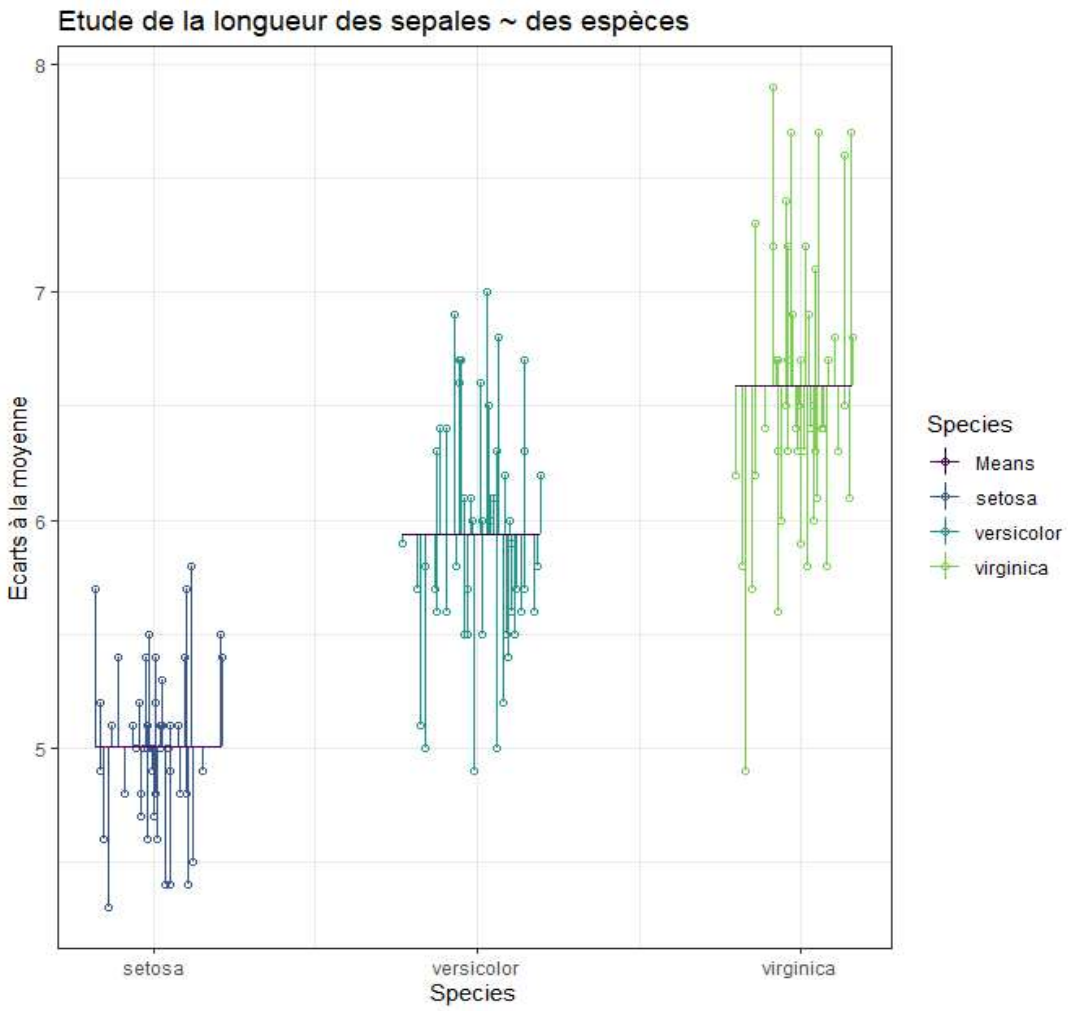
- **"Normal Q-Q".**

  Les résidus suivent une loi normale de moyenne 0. Pour vérifier cette hypothèse on peut faire un test de normalité comme le test de Shapiro-Wilk mais on préfère vérifier cela graphiquement avec un diagramme Quantile-Quantile (i.e. QQ-plot, graphique dans lequel les quantiles de deux distributions sont tracés l'un par rapport à l'autre).

- **"Scale-Location".**

  L'homogénéité des variances. Les résidus relatifs aux différentes modalités sont homogènes (ils ont globalement la même dispersion), autrement dit leur variance est constante. On peut vérifier cela graphiquement en représentant les résidus standardisés en fonction des valeurs prédites (les moyennes des différents traitements). En cas de doute on pourra aussi valider cette hypothèse avec un test statistique (Cochran, Bartlett, Levene…).

Illustrations



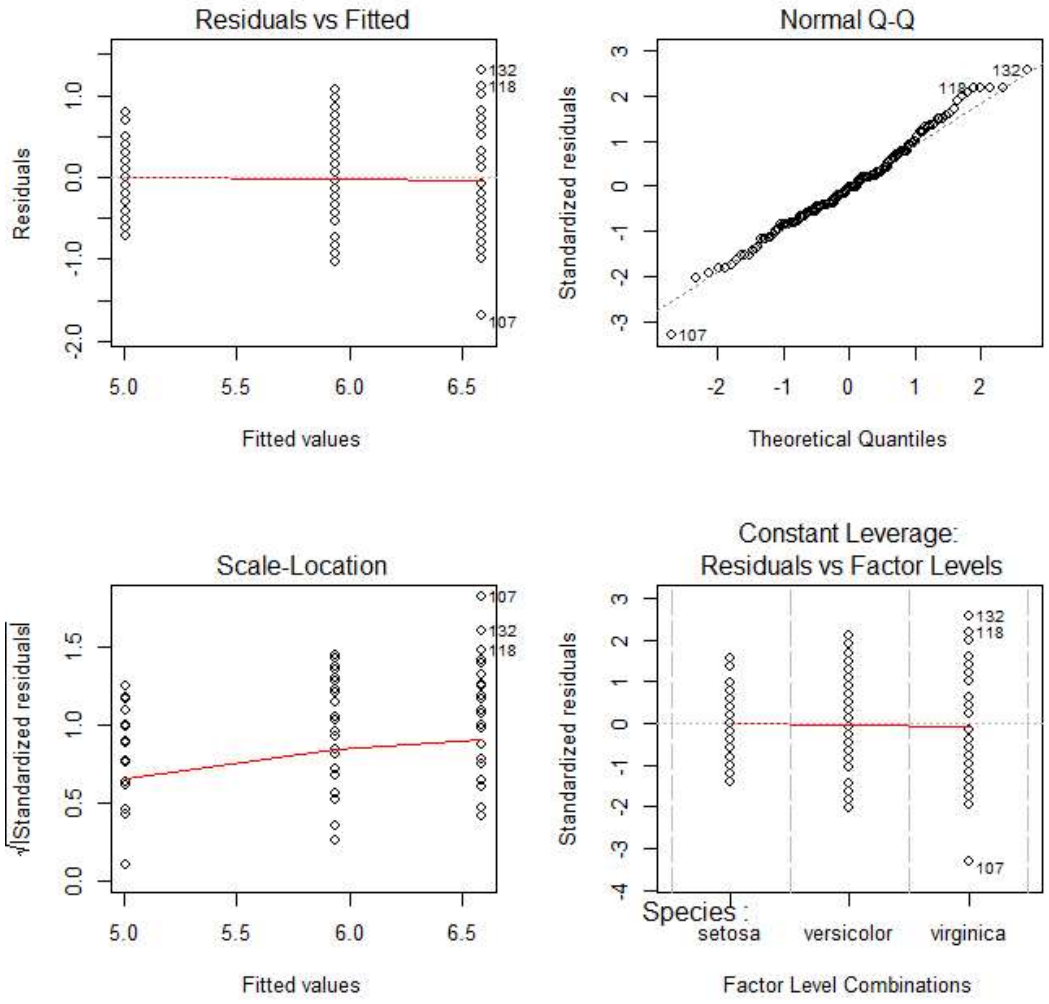

```
## run the anova in R
anova <- aov(formula = Sepal.Length~Species, data = iris)
## call print method by default
anova

## Call:
##    aov(formula = Sepal.Length ~ Species, data = iris)
##
## Terms:
##                   Species Residuals
## Sum of Squares  63.21213  38.95620
## Deg. of Freedom        2       147
##
## Residual standard error: 0.5147894
## Estimated effects may be unbalanced
```

```r
## call summary method
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## print 4 diag plots
par(mfrow=c(2, 2)); plot(anova)
```

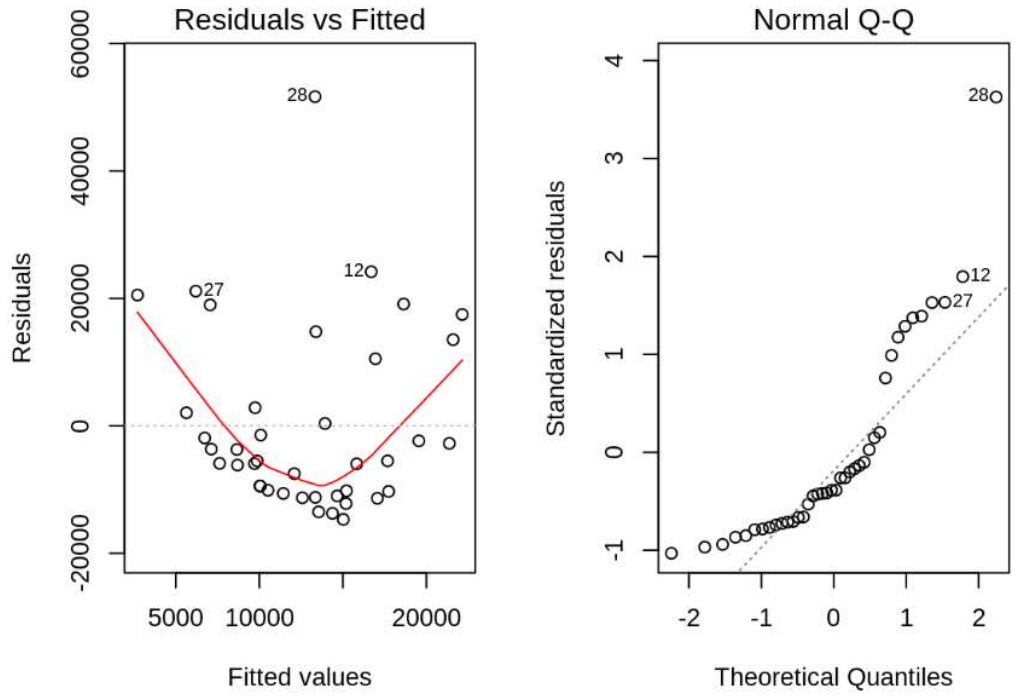


Exemple de “mauvais” diagnostique

Pour bien se rendre compte que ces précédents exemples sont signes de bon diagnostic, on peut les confronter à des graphiques qui devraient vous alarmer, signe de “mauvais” diagnostique.

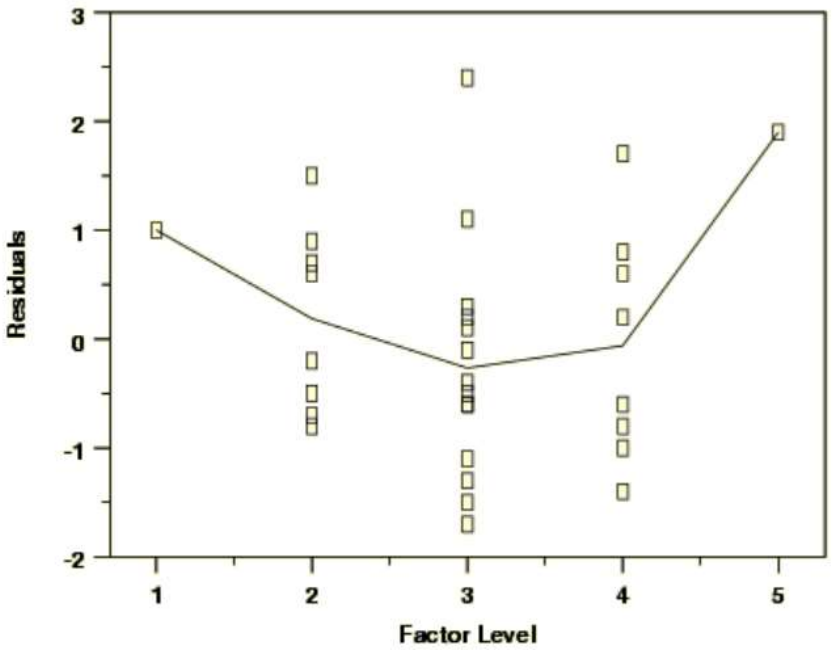
Dans le graph **“Residuals vs Fitted“**, la ligne rouge n’est pas une droite horizontale proche de Y=0.

Dans le graph **“Normal Q-Q“**, les points observés s’éloignent franchement de la droite théorique. (On pourrait aussi voir ce problème en faisant le graphique de densité des résidus, alors nous n’aurions certainement pas “une cloche” comme attendu avec le loi Normale).



Source : R Cookbook

Dans le graph **"Residuals vs Factor Levels"** ci-dessous, la valeur des résidus n'est pas rassurante quand on regarde les différents niveaux de facteur.



Source : itl.nist.gov

Dans ces cas-là, il faut s'interroger sur la nécessité d'appliquer une transformation (racine carré, log, etc) à notre prédicteur, changer de modèle, ou envisager les tests non-paramétriques.

Test de comparaison de variance

Hypothèses stochastiques

Dans le même contexte décrit pour le test ANOVA.
On s'intéresse ici à ce point en particulier : Les variances conditionnelles (variances dans chaque sous-population) sont identiques. On parle d'**homoscédasticité.**

Enjeu

L'enjeu d'un test d'homogénéité pour σ^2 est d'affirmer, avec un faible risque de se tromper, que $\mathcal{P}_1, \ldots, \mathcal{P}_p$ ne sont pas homogènes quant à σ^2.
Pour ce faire, on compare les variances $\sigma_1^2, \ldots, \sigma_p^2$ via un test statistique.

Hypothèses

On considère les hypothèses :

$\boldsymbol{H_0: \sigma_1^2 = \ldots = \sigma_p^2 = \sigma}$
contre
$\boldsymbol{H_1: \exists j, \sigma_j \neq \sigma}$ **(il existe au moins 2 variances différentes)**

Objectif

On peut décider du rejet de H_0, au risque $\alpha/100, \alpha \in]0,1[$.

Dans ce cas précis, on ne veut pas rejeter H_0 au risque 5%, puisqu'on souhaite vérifier l'égalité des variances. Par convention, on admet que $\sigma_1^2 = \ldots = \sigma_p^2$.

Test de Cochran

- Test statistique : Si $n_1 = \ldots = n_p$, on utilise le test de Cochran.
  N.B. : si $n_i \neq n_j$ voir test de Bartlett.

- Calculs : Le test de Cochran se met en œuvre en calculant :

  - si, $i \in \{1, .., p\}$: l'écart-type corrigé de $x_{i,1}, \ldots, x_{i,n_i}$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2,$$

  - le c_{obs} défini par

$$c_{obs} = \frac{max_{i \in \{1,..,p\}} s_i^2}{\sum_{i=1}^{p} s_i^2}$$

  - le réel $c(m, p)$ avec $m = n_1$ (l'effectif commun).
    Si $\alpha = 0.05$, ce réel est évaluable dans la table ANNEXE 7 (Valeurs de Cochran).

- Règles de décision : La règle de décision associées au test de Cochran est :
  Si $c_{obs} \geq c(m,p)$,
  Alors on rejette H_0.
  Par exemple si $c_{obs} < c(m,p)$ et $\alpha = 0.05$, alors on ne rejette pas H_0 ; On admet que $\sigma_1^2 = \ldots = \sigma_p^2$.

Test de Bartlett

- Test statistique : Si $min(n_1, \ldots, n_p) \geq 4$, on utilise le test de Bartlett.

- Calculs : Le test de Bartlett se met en œuvre en calculant :

  - si, $i \in \{1, .., p\}$: l'écart-type corrigé de $x_{i,1}, \ldots, x_{i,n_i}$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2,$$

  - le χ_{obs}^2 défini par

$$\chi_{obs}^2 = \frac{(n-p)ln(s^2) - \sum_{i=1}^{p}(n_i - 1)ln(s_i^2)}{1 + \frac{1}{3(p-1)}((\sum_{i=1}^{p}\frac{1}{(n_i-1)}) - \frac{1}{(n-p)})}$$

  - le réel $\mathbb{P}(K \geq \chi_\alpha^2(\upsilon)) = \alpha$ avec $K \sim \chi^2(\upsilon), \upsilon = p - 1$.
    Si $\alpha = 0.05$, ce réel est évaluable dans la table ANNEXE 4 (Loi du Chi-deux).

- Règles de décision : La règle de décision associées au test de Bartlett est :
  Si $\chi_{obs}^2 \geq \chi_\alpha^2(\upsilon)$,
  Alors on rejette H_0.

Test de comparaison de moyennes

Hypothèses stochastiques

Dans le même contexte décrit pour le test ANOVA.
On s'intéresse ici à la suite de l'étude : la réalisation d'un test "post-hoc".

Enjeu

Si le test ANOVA indique qu'au moins deux moyennes différent, il est intéressant d'étudier la différence de deux d'entre elles.

Hypothèses

Soit $(k, l) \in \{1, .., p\}$ avec $k \neq l$. On considère le test statistique :

$\boldsymbol{H_0 : \mu_k = \mu_l}$
contre
$\boldsymbol{H_0 : \mu_k \neq \mu_l}$

Test de Bonferroni

- Test statistique : On utilise le test de Bonferroni, lequel offre plus de précision que des t-Test 2 à 2 car il prend en compte toutes les données dans sa construction (avec la présence du cm_R).
  Ce test repose sur la loi de Student $\mathcal{T}(\upsilon)$.

- Calculs : Le test de Bonferroni se met en œuvre en calculant :

  - $\overline{x}, i \in \{k, l\}$: moyenne de $x_{i,1}, \ldots, x_{i,n_i}$,
  - $s_R = \sqrt{cm_R}$ (évaluable dans le tableau ANOVA),
  - le t_{obs} défini par

$$t_{obs} = \frac{\overline{x}_k - \overline{x}_l}{s_R \sqrt{\frac{1}{n_k} + \frac{1}{n_l}}},$$

  - le réel $t_\alpha^{**}(\upsilon)$ vérifiant

$$\mathbb{P}(|T| \geq t_\alpha^{**}(\upsilon)) = \frac{2\alpha}{p(p-1)},$$

  où $T \sim \mathcal{T}(\upsilon), \upsilon = n - p$.
  Ce réel est dans la table ANNEXE 3 (Loi de Student).

- Règles de décision : La règle de décision associée au test de Bonferroni est :
  Si $t_{obs} \geq t_\alpha^{**}(\upsilon)$,
  Alors on rejette H_0.

- p-valeurs : La p-valeur associée au test de Bonferroni est

$$p - valeur = \mathbb{P}(|T| \geq |t_{obs}|).$$

Autres tests de comparaison de moyennes

Il existe de nombreux autres tests post-hoc pour la comparaison de moyenne 2 à 2, entre autres :

- Le test de Tukey HSD (Honest Significant Differences),
- Le test de la petite différence significative (LSD) de Fisher,
- Le test Student de Newman-Keuls,
- Dunnett
- …

Ce sera le role du statisticien de se documenter sur les particularités de ces tests et de choisir le plus pertinent selon l'études et les données.

Récap'

ANOVA à un facteur

Soient A un facteur à p modalités, $A_1, \dots, A_p$, et X un caractère quantitatif. Pour tout $i \in \{1, \dots, p\}$, X sous A_i est une *var* $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Les paramètres sont inconnus et on suppose que

$$\sigma_1^2 = \dots = \sigma_p^2.$$

On observe la valeur de X_i pour chacun des n_i individus d'un échantillon. Les échantillons sont indépendants.
On considère les hypothèses :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p \ (A \text{ n'influe pas sur } X) \qquad \text{contre} \qquad H_1 : \text{"il existe au moins 2 moyennes différentes } (A \text{ influe sur } X)\text{"}$$

Pour pouvoir décider du rejet de H_0 au risque $100\alpha\%$, $\alpha \in]0,1[$,

- on calcule (si besoin est) $n = \sum_{i=1}^p n_i$, $\overline{x} = \frac{1}{n}\sum_{i=1}^p \sum_{j=1}^{n_i} x_{i,j}$, $\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j}$.
- on dresse le tableau ANOVA à un facteur :

	sce (SS)	ddl (DF)	cm (MS)	f_{obs} (F)
Total	$\text{sce}_T = \sum_{i=1}^p \sum_{j=1}^{n_i} (x_{i,j} - \overline{x})^2$	$\text{ddl}_T = n - 1$		
Factoriel	$\text{sce}_F = \sum_{i=1}^p n_i (\overline{x}_i - \overline{x})^2$	$\text{ddl}_F = p - 1$	$\text{cm}_F = \frac{\text{sce}_F}{\text{ddl}_F}$	$f_{obs} = \frac{\text{cm}_F}{\text{cm}_R}$
Residuel	$\text{sce}_R = \sum_{i=1}^p \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2$	$\text{ddl}_R = n - p$	$\text{cm}_R = \frac{\text{sce}_R}{\text{ddl}_R}$	

- on calcule le réel $f_\alpha(\nu_1, \nu_2)$ tel que

$$\mathbb{P}(F \geq f_\alpha(\nu_1, \nu_2)) = \alpha,$$

où $F \sim \mathcal{F}(\nu_1, \nu_2)$, $(\nu_1, \nu_2) = (\text{ddl}_F, \text{ddl}_R) = (p - 1, n - p)$.

Si

$$f_{obs} \geq f_\alpha(\nu_1, \nu_2),$$

alors on rejette H_0.

Ou alors : p-valeur= $\mathbb{P}(F \geq f_{obs}) \leq \alpha \Rightarrow$ on rejette H_0 au risque $100\alpha\%$

ANOVA à un facteur : compléments

On reprend les notations de "ANOVA à un facteur". Outil : $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2$.

Égalité des variances : On considère des hypothèses :

$$H_0 : \sigma_1^2 = \dots = \sigma_p^2 \qquad \text{contre} \qquad H_1 : \text{"il existe au moins 2 variances différentes"}$$

Tests	Stat. test obs.	Valeurs	Rejet de H_0 si
Bartlett (si $\min(n_1, \dots, n_p) \geq 4$)	$\chi^2_{obs} = \frac{(n-p)\ln(s^2) - \sum_{i=1}^p (n_i - 1)\ln(s_i^2)}{1 + \frac{1}{3(p-1)}\left(\left(\sum_{i=1}^p \frac{1}{n_i - 1}\right) - \frac{1}{n-p}\right)}$	$\mathbb{P}(K \geq \chi^2_\alpha(\nu)) = \alpha$, $K \sim \chi^2(\nu), \nu = p - 1$	$\chi^2_{obs} \geq \chi^2_\alpha(\nu)$
Cochran (si $n_1 = \dots = n_p = m$)	$c_{obs} = \frac{\max_{i \in \{1,\dots,p\}} s_i^2}{\sum_{i=1}^p s_i^2}$	$c(m,p)$ (voir table correspondante)	$c_{obs} \geq c(m,p)$

Comparaison de 2 moyennes : test de Bonferroni : Soit $(k, \ell) \in \{1, \dots, p\}^2$ avec $k \neq \ell$. On considère les hypothèses :

$$H_0 : \mu_k = \mu_\ell \qquad \text{contre} \qquad H_1 : \mu_k \neq \mu_\ell.$$

Tests	Stat. test obs.	Valeurs	Rejet de H_0 si
Bonferroni	$t_{obs} = \frac{\overline{x}_k - \overline{x}_\ell}{s\sqrt{\frac{1}{n_k} + \frac{1}{n_\ell}}}$	$\mathbb{P}(\lvert T\rvert \geq t^{**}_\alpha(\nu)) = \frac{2\alpha}{p(p-1)}$, $T \sim \mathcal{T}(\nu)$, $\nu = n - p$	$\lvert t_{obs}\rvert \geq t^{**}_\alpha(\nu)$