Cours 10 – Tests de conformité
Test de comparaison de 2 moyennes.

Eya ZOUGAR *

Institut National des Sciences appliqués-INSA

Génie mathématiques GM3
Monday 3rd April, 2023

*Basé sur le cours de Bruno PORTIER

1. Introduction.

Dans ce cours, on s'intéresse à:

- Le test de conformité d'une moyenne;
- Le test de conformité d'une variance;
- Le test de comparaison de deux moyennes dans le cadre de deux échantilons gaussiens indépendants.

2. Test de conformité d'une moyenne (Test de Student).

2.1 Le problème.

On considère n données réelles $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable quantitative.

On se place dans le cadre de l'échantillon gaussien, c'est à dire que l'on suppose que les données $x_1, \ldots, x_n$ sont les réalisations de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, les paramètres m et σ^2 étant supposés inconnus.

On veut savoir au risque α si la moyenne théorique m des données est différente ou non d'une valeur m_0 donnée a priori (valeur de référence ou supposée).

Pour cela, on teste, au risque α, l'hypothèse nulle

$$H_0 : \ll m = m_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll m \neq m_0 \gg$$

2.2. Choix de la statistique de test.

Pour tester $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m \neq m_0 \gg$, il nous faut maintenant choisir une statistique de test.
On sait que pour tout $m \in \mathbb{R}$,

$$\frac{\sqrt{n}(\overline{X}_n - m)}{S} \sim T_{n-1} \quad \text{avec} \quad \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ et } S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

Sous H_0, le paramètre m est égal à m_0.

2.2. Choix de la statistique de test.

Pour tester $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m \neq m_0 \gg$, il nous faut maintenant choisir une statistique de test.
On sait que pour tout $m \in \mathbb{R}$,

$$\frac{\sqrt{n}(\overline{X}_n - m)}{S} \sim T_{n-1} \quad \text{avec} \quad \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ et } S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

Sous H_0, le paramètre m est égal à m_0.

- Donc, sous H_0, la variable Z définie par

$$Z = \frac{\sqrt{n}(\overline{X}_n - m_0)}{S}$$

suit une loi de Student à $n-1$ ddl, et on note $Z \underset{H_0}{\sim} T_{n-1}$.

2.2. Choix de la statistique de test.

Pour tester $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m \neq m_0 \gg$, il nous faut maintenant choisir une statistique de test.
On sait que pour tout $m \in \mathbb{R}$,

$$\frac{\sqrt{n}(\overline{X}_n - m)}{S} \sim T_{n-1} \quad \text{avec} \quad \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ et } S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

Sous H_0, le paramètre m est égal à m_0.

- Donc, sous H_0, la variable Z définie par

$$Z = \frac{\sqrt{n}(\overline{X}_n - m_0)}{S}$$

suit une loi de Student à $n-1$ ddl, et on note $Z \underset{H_0}{\sim} T_{n-1}$.

- En revanche, sous H_1, Z ne suit plus une T_{n-1} puisque on à la décomposition suivante :

$$Z = \underbrace{\frac{\sqrt{n}(\overline{X}_n - m)}{S}}_{\sim T_{n-1}} + \underbrace{\frac{\sqrt{n}(m - m_0)}{S}}_{\neq 0 \text{ sous } H_1}$$

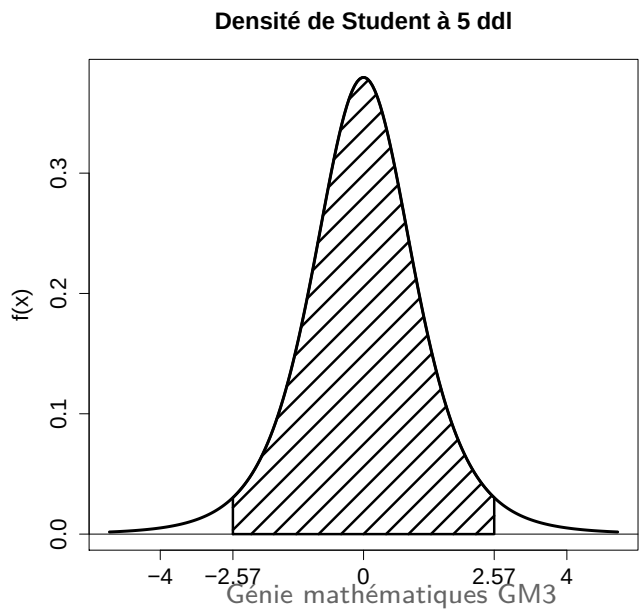
2.3. Construction de la zone de rejet du test.

L'idée va être de rejeter l'hypothèse nulle H_0 lorsque l'on observera une valeur de la statistique de test faisant partie des valeurs les moins probables (petites ou grandes) d'une loi de Student à $(n-1)$ ddl.

La loi de Student étant centrée et symétrique, on cherche une zone de rejet de l'hypothèse nulle H_0 au risque α de la forme $\{|Z| > t\} = \{Z < -t\} \bigcup \{Z > t\}$ avec $t > 0$ tel que:

$$\mathbb{P}_{H_0}\left[|Z| > t\right] = \alpha \iff \mathbb{P}_{H_0}\left[|Z| \leq t\right] = 1 - \alpha$$

Le réel t est ainsi le quantile d'ordre $(1 - \alpha/2)$ d'une loi de Student à $(n-1)$ ddl, noté $t_{n-1,1-\alpha/2}$.



2.4. Construction du test de conformité.

Donc, pour tester au risque α l'hypothèse nulle $H_0 : \ll m = m_0 \gg$ contre l'hypothèse alternative $H_1 : \ll m \neq m_0 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S} \underset{H_0}{\sim} T_{n-1}$.

La zone de rejet est de la forme $\left\{|Z| > t_{n-1,1-\alpha/2}\right\}$ où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $n-1$ ddl, c'est à dire $t_{n-1,1-\alpha/2}$ est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{n-1,1-\alpha/2}\right] = \mathbb{P}\left[|T_{n-1}| \leq t_{n-1,1-\alpha/2}\right] = 1-\alpha$$

2.4. Construction du test de conformité.

Donc, pour tester au risque α l'hypothèse nulle $H_0 : \ll m = m_0 \gg$ contre l'hypothèse alternative $H_1 : \ll m \neq m_0 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S} \underset{H_0}{\sim} T_{n-1}$.

La zone de rejet est de la forme $\left\{|Z| > t_{n-1,1-\alpha/2}\right\}$ où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $n-1$ ddl, c'est à dire $t_{n-1,1-\alpha/2}$ est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{n-1,1-\alpha/2}\right] = \mathbb{P}\left[|T_{n-1}| \leq t_{n-1,1-\alpha/2}\right] = 1-\alpha$$

On calcule alors la valeur z_{obs} de Z sur les données :

$$z_{obs} = \frac{\sqrt{n}(\overline{x}_n - m_0)}{s}$$

2.4. Construction du test de conformité.

Donc, pour tester au risque α l'hypothèse nulle $H_0 : \ll m = m_0 \gg$ contre l'hypothèse alternative $H_1 : \ll m \neq m_0 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S} \underset{H_0}{\sim} T_{n-1}$.

La zone de rejet est de la forme $\left\{|Z| > t_{n-1,1-\alpha/2}\right\}$ où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $n-1$ ddl, c'est à dire $t_{n-1,1-\alpha/2}$ est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{n-1,1-\alpha/2}\right] = \mathbb{P}\left[|T_{n-1}| \leq t_{n-1,1-\alpha/2}\right] = 1-\alpha$$

On calcule alors la valeur z_{obs} de Z sur les données :

$$z_{obs} = \frac{\sqrt{n}(\overline{x}_n - m_0)}{s}$$

• Si $|z_{obs}| \leq t_{n-1,1-\alpha/2}$, alors on ne peut pas rejeter l'hypothèse nulle H_0 et on considère que la moyenne observée n'est pas significativement différente de m_0.

• Sinon, on rejette l'hypothèse nulle H_0 au risque α, et on considère que la moyenne des données est significativement différente de m_0.

2.5. Lien avec l'intervalle de confiance.

2.5. Lien avec l'intervalle de confiance.

On notera que ne pas rejeter l'hypothèse nulle H_0 au risque α est équivalent à ce que m_0 appartienne à l'intervalle de confiance de la moyenne m au niveau de confiance $(1 - \alpha)$.

2.5. Lien avec l'intervalle de confiance.

On notera que ne pas rejeter l'hypothèse nulle H_0 au risque α est équivalent à ce que m_0 appartienne à l'intervalle de confiance de la moyenne m au niveau de confiance $(1-\alpha)$.

En effet, on ne rejette pas l'hypothèse nulle H_0 si

$$
\begin{aligned}
|z_{obs}| \leq t_{n-1,1-\alpha/2} &\iff -t_{n-1,1-\alpha/2} \leq \frac{\sqrt{n}(\bar{x}_n - m_0)}{s} \leq t_{n-1,1-\alpha/2} \\
&\iff -t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \bar{x}_n - m_0 \leq t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \\
&\iff m_0 \in \left[\bar{x}_n \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right] = \mathrm{IC}_{1-\alpha}(m)
\end{aligned}
$$

2.5. Lien avec l'intervalle de confiance.

On notera que ne pas rejeter l'hypothèse nulle H_0 au risque α est équivalent à ce que m_0 appartienne à l'intervalle de confiance de la moyenne m au niveau de confiance $(1-\alpha)$.

En effet, on ne rejette pas l'hypothèse nulle H_0 si

$$
\begin{aligned}
|z_{obs}| \leq t_{n-1,1-\alpha/2} &\iff -t_{n-1,1-\alpha/2} \leq \frac{\sqrt{n}(\bar{x}_n - m_0)}{s} \leq t_{n-1,1-\alpha/2} \\
&\iff -t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \bar{x}_n - m_0 \leq t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \\
&\iff m_0 \in \left[\bar{x}_n \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right] = \mathrm{IC}_{1-\alpha}(m)
\end{aligned}
$$

Ainsi, pour décider entre H_0 et H_1 au risque α, on peut regarder si m_0 appartient ou non à l'intervalle de confiance de la moyenne théorique m au niveau de confiance $(1-\alpha)$.

2.6. Test unilatéral.

• Il est aussi possible de mettre en oeuvre **un test unilatéral.** On peut en effet tester au risque α :

- $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m < m_0 \gg$;
- Ou bien: $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m > m_0 \gg$**.**

2.6. Test unilatéral.

• Il est aussi possible de mettre en oeuvre **un test unilatéral.** On peut en effet tester au risque α :

- $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m < m_0 \gg$;
- Ou bien: $H_0 : \ll m = m_0 \gg$ contre $H_1 : \ll m > m_0 \gg$**.**

• La statistique de test reste inchangée, mais les zones de rejet sont alors définies respectivement par:

- $\{Z < t_{n-1,\alpha}\}$ avec $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$;
- $\{Z > t_{n-1,1-\alpha}\}$;

où $t_{n-1,1-\alpha}$ désigne le quantile d'ordre $(1-\alpha)$ d'une loi de Student à $(n-1)$ ddl.

2.7. Illustration sur le poids des porcs.
2.7.1. Description du problème .

A la suite d'un traitement (régime alimentaire) sur une variété de porcs, on prélève un échantillon de 5 porcs et on les pèse.

On obtient les poids suivants (en kg) : 83 ; 81 ; 84 ; 80 ; 85.

On sait que le poids moyen pour cette variété de porcs est de 87,5 kg. On suppose que le poids de cette variété de porcs est normalement distribué.

Le poids moyen des porcs traités diffère-t-il significativement de cette norme au seuil de 5 % ?

2.7.2. Test bilatéral.

On teste au risque 5% l'hypothèse nulle H_0 : $\ll m = 87,5 \gg$ contre H_1 : $\ll m \neq 87,5 \gg$.
Avec le logiciel R, on résoud le problème avec les instructions suivantes :

```
X = c(83 , 81 , 84 , 80 , 85)
t.test(X, mu= 87.5 , conf.level=0.95)
```

On obtient le résultat suivant ($t_{4,0.975} = 2,776$) :

```
> t.test(X, mu= 87.5 , conf.level=0.95)
^^IOne Sample t-test
data:  X
t = -5.2838, df = 4, p-value = 0.006154
alternative hypothesis: true mean is not equal to 87.5
95 percent confidence interval: 80.02523 85.17477
sample estimates: mean of x   82.6
```

On peut constater que $m_0 = 87,5$ n'appartient pas à l'intervalle de confiance On rejette donc H_0 au risque 5%.

2.7.3. Test unilatéral.

Le test de $H_0 : \ll m = 87,5 \gg$ contre $H_1 : \ll m < 87,5 \gg$ conduit à rejeter H_0 au risque 5% ($t_{4,0.05} = -2,131$)

```
> t.test(X, mu= 87.5 , conf.level=0.95, alternative = "less")
One Sample t-test
t = -5.2838, df = 4, p-value = 0.003077
alternative hypothesis: true mean is less than 87.5
95 percent confidence interval:  -Inf 84.57699
sample estimates: mean of x  82.6
```

2.7.3. Test unilatéral.

Le test de $H_0 : \ll m = 87,5 \gg$ contre $H_1 : \ll m < 87,5 \gg$ conduit à rejeter H_0 au risque 5% ($t_{4,0.05} = -2,131$)

```
> t.test(X, mu= 87.5 , conf.level=0.95, alternative = "less")
One Sample t-test
t = -5.2838, df = 4, p-value = 0.003077
alternative hypothesis: true mean is less than 87.5
95 percent confidence interval:  -Inf 84.57699
sample estimates: mean of x  82.6
```

Alors que le test de $H_0 : \ll m = 87,5 \gg$ contre $H_1 : \ll m > 87,5 \gg$ conduit à ne pas rejeter H_0 au risque 5% ($t_{4,0.95} = 2,131$)

```
> t.test(X, mu= 87.5 , conf.level=0.95, alternative = "greater")
One Sample t-test
t = -5.2838, df = 4, p-value = 0.9969
alternative hypothesis: true mean is greater than 87.5
95 percent confidence interval:  80.62301  Inf
sample estimates: mean of x 82.6
```

3. Test de conformité d'une variance.
3.1. Le problème.

On considère n données réelles $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable quantitative.

On se place dans le cadre de l'échantillon gaussien, c'est à dire que l'on suppose que les données $x_1, \ldots, x_n$ sont les réalisations de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, les paramètres m et σ^2 étant supposés inconnus.

On veut savoir au risque α si la variance théorique σ^2 des données est différente ou non d'une valeur σ_0^2 donnée a priori (valeur de référence ou supposée).

3. Test de conformité d'une variance.
3.1. Le problème.

On considère n données réelles $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable quantitative.

On se place dans le cadre de l'échantillon gaussien, c'est à dire que l'on suppose que les données $x_1, \ldots, x_n$ sont les réalisations de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, les paramètres m et σ^2 étant supposés inconnus.

On veut savoir au risque α si la variance théorique σ^2 des données est différente ou non d'une valeur σ_0^2 donnée a priori (valeur de référence ou supposée).

Pour cela, on teste, au risque α, l'hypothèse nulle

$$H_0 : \ll \sigma^2 = \sigma_0^2 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \sigma^2 \neq \sigma_0^2 \gg$$

3.2. Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.

3.2. Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $\sigma^2 \in \mathbb{R}_+$,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{avec} \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

Sous H_0, la variance σ^2 est égale à σ_0^2.

3.2. Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $\sigma^2 \in \mathbb{R}_+$,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{avec} \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

Sous H_0, la variance σ^2 est égale à σ_0^2.

- Donc, sous H_0, la variable Z définie par

$$Z = \frac{(n-1)S^2}{\sigma_0^2}$$

suit une loi du khi-deux à $(n-1)$ ddl, et on note $Z \underset{H_0}{\sim} \chi^2_{n-1}$.

3.2. Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $\sigma^2 \in \mathbb{R}_+$,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{avec} \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

Sous H_0, la variance σ^2 est égale à σ_0^2.

- Donc, sous H_0, la variable Z définie par

$$Z = \frac{(n-1)S^2}{\sigma_0^2}$$

suit une loi du khi-deux à $(n-1)$ ddl, et on note $Z \underset{H_0}{\sim} \chi^2_{n-1}$.

- En revanche, sous H_1, Z ne suit plus une χ^2_{n-1} puisque l'on a la décomposition suivante :

$$Z = \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{\sim\chi^2_{n-1}} \times \underbrace{\frac{\sigma^2}{\sigma_0^2}}_{\neq 1 \text{ sous } H_1}$$

3.3. Construction de la zone de rejet.

Ici, compte-tenu de la forme de la statistique de test et de l'hypothèse alternative, nous allons rejeter l'hypothèse nulle lorsque la valeur de la statistique de test fera partie des valeurs les moins probables d'un Khi-deux à $(n-1)$ ddl, petites ou grandes.

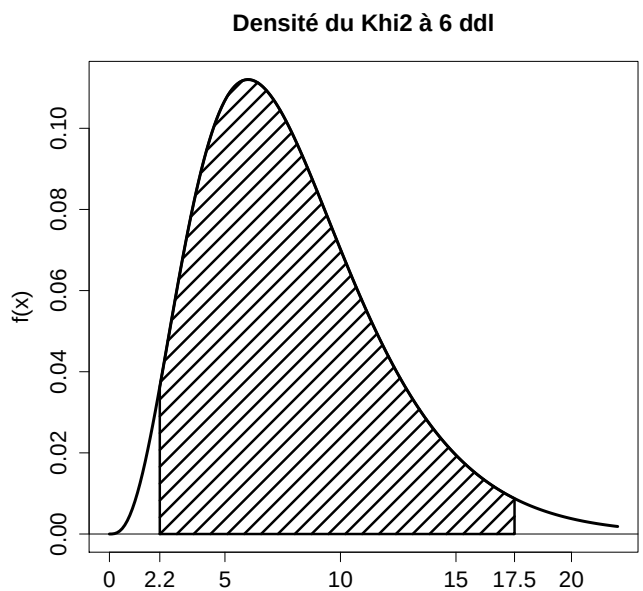
3.3. Construction de la zone de rejet.

Ici, compte-tenu de la forme de la statistique de test et de l'hypothèse alternative, nous allons rejeter l'hypothèse nulle lorsque la valeur de la statistique de test fera partie des valeurs les moins probables d'un Khi-deux à $(n-1)$ ddl, petites ou grandes.

On cherchera donc une région de rejet de l'hypothèse nulle au risque α de la forme $\{Z < k_1\} \bigcup \{Z > k_2\}$ avec $k_1 > 0$ et $k_2 > 0$ tels que

$$\mathbb{P}_{H_0}[Z < k_1] + \mathbb{P}_{H_0}[Z > k_2] = \alpha \iff \mathbb{P}_{H_0}[k_1 \leq Z \leq k_2] = 1 - \alpha$$

Les réels k_1 et k_2 sont donc respectivement les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ d'une loi du khi-deux à $(n-1)$ ddl, notés $k_{\alpha/2}$ et $k_{1-\alpha/2}$.



3.4. Construction du test de conformité.

Donc, pour tester au risque α l'hypothèse nulle $H_0 : \ll \sigma^2 = \sigma_0^2 \gg$ contre l'hypothèse alternative $H_1 : \ll \sigma^2 \neq \sigma_0^2 \gg$,

• on utilise la statistique de test $Z = \dfrac{(n-1)S^2}{\sigma_0^2} \underset{H_0}{\sim} \chi^2_{n-1}$.

• La zone de rejet est de la forme $\{Z < k_{\alpha/2}\} \bigcup \{Z > k_{1-\alpha/2}\}$ où $k_{\alpha/2}$ et $k_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $(1-\alpha/2)$ de la loi du khi-deux à $(n-1)$ ddl.

• On calcule alors la valeur z_{obs} de Z sur les données : $z_{obs} = \dfrac{(n-1)s^2}{\sigma_0^2}$,

avec s^2 la réalisation de S^2 sur les données.

Conclusions:

• Si $k_{\alpha/2} \leq z_{obs} \leq k_{1-\alpha/2}$, alors on ne peut pas rejeter l'hypothèse nulle H_0 et on considère que la variance observée n'est pas significativement différente de σ_0^2.

• Sinon, on rejette l'hypothèse nulle H_0 au risque α, et on considère que la variance des données est significativement différente de σ_0^2.

3.5. Lien avec l'intervalle de confiance pour σ^2.

On notera que ne pas rejeter l'hypothèse nulle H_0 est équivalent à ce que σ_0^2 appartienne à l'intervalle de confiance de la variance σ^2.

3.5. Lien avec l'intervalle de confiance pour σ^2.

On notera que ne pas rejeter l'hypothèse nulle H_0 est équivalent à ce que σ_0^2 appartienne à l'intervalle de confiance de la variance σ^2.

En effet, on ne rejette pas l'hypothèse nulle H_0 si

$$\begin{aligned} k_{\alpha/2} \leq z_{obs} \leq k_{1-\alpha/2} &\iff k_{\alpha/2} \leq \frac{(n-1)s^2}{\sigma_0^2} \leq k_{1-\alpha/2} \\ &\iff \frac{(n-1)s^2}{k_{1-\alpha/2}} \leq \sigma_0^2 \leq \frac{(n-1)s^2}{k_{\alpha/2}} \\ &\iff \sigma_0^2 \in \mathrm{IC}_{1-\alpha}(\sigma^2) \end{aligned}$$

ainsi, pour tester H_0 contre H_1 au risque α, on peut aussi regarder si σ_0^2 appartient ou non à l'intervalle de confiance de σ^2 au niveau de confiance $(1-\alpha)$.

4. Le test bilatéral de comparaison de deux moyennes.

4.1. Le problème

- On dispose de deux échantillons de mesures :
  - $x_1, x_2, \ldots, x_p$ qui sont p mesures indépendantes d'une variable quantitative X,
  - $y_1, y_2, \ldots, y_q$ qui sont q mesures indépendantes d'une variable quantitative Y.
- On souhaite comparer les moyennes théoriques des variables X et Y, sachant que leur variabilité est identique.

- Ces moyennes sont respectivement estimées par $\overline{x}_p$ et $\overline{y}_q$. Leur comparaison devrait donc nous permettre de répondre à la question. Cependant, en raison du phénomène des fluctuations d'échantillonnage, même si les moyennes théoriques sont égales, les moyennes empiriques $\overline{x}_p$ et $\overline{y}_q$ seront différentes.

Toute la question sera donc de savoir si la différence observée entre $\overline{x}_p$ et $\overline{y}_q$ n'est due qu'au seul fait du hasard (fluctuations d'échantillonnage) ou bien au fait que les moyennes théoriques sont réellement différentes ?

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

- Les données $x_1, x_2, \ldots, x_p$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$;

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

- Les données $x_1, x_2, \ldots, x_p$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$;
- Les données $y_1, y_2, \ldots, y_q$ sont les réalisations des variables aléatoires $Y_1, Y_2, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$;

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

- Les données $x_1, x_2, \ldots, x_p$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$;
- Les données $y_1, y_2, \ldots, y_q$ sont les réalisations des variables aléatoires $Y_1, Y_2, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$;
- Les 2 échantillons sont de même variance, c'est à dire que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

- Les données $x_1, x_2, \ldots, x_p$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$;
- Les données $y_1, y_2, \ldots, y_q$ sont les réalisations des variables aléatoires $Y_1, Y_2, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$;
- Les 2 échantillons sont de même variance, c'est à dire que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Les 2 échantillons sont indépendants.

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

- Les données $x_1, x_2, \ldots, x_p$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$;
- Les données $y_1, y_2, \ldots, y_q$ sont les réalisations des variables aléatoires $Y_1, Y_2, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$;
- Les 2 échantillons sont de même variance, c'est à dire que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Les 2 échantillons sont indépendants.

Pour savoir si les moyennes théoriques sont égales, nous allons tester l'hypothèse nulle H_0 suivante

$$H_0 : \ll \mu_1 = \mu_2 \gg \iff H_0 : \ll \mu_1 - \mu_2 = 0 \gg$$

contre l'hypothèse alternative H_1 suivante :

$$H_1 : \ll \mu_1 \neq \mu_2 \gg \iff H_1 : \ll \mu_1 - \mu_2 \neq 0 \gg$$

4.2. Le modèle probabiliste.

On fait les hypothèses fondamentales suivantes :

- Les données $x_1, x_2, \ldots, x_p$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$;
- Les données $y_1, y_2, \ldots, y_q$ sont les réalisations des variables aléatoires $Y_1, Y_2, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$;
- Les 2 échantillons sont de même variance, c'est à dire que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Les 2 échantillons sont indépendants.

Pour savoir si les moyennes théoriques sont égales, nous allons tester l'hypothèse nulle H_0 suivante

$$H_0 : \ll \mu_1 = \mu_2 \gg \iff H_0 : \ll \mu_1 - \mu_2 = 0 \gg$$

contre l'hypothèse alternative H_1 suivante :

$$H_1 : \ll \mu_1 \neq \mu_2 \gg \iff H_1 : \ll \mu_1 - \mu_2 \neq 0 \gg$$

Notons que l'hypothèse H_1 pourrait-être $H_1 : \ll \mu_1 > \mu_2 \gg$ ou bien encore $H_1 : \ll \mu_1 < \mu_2 \gg$.

4.3. Le résultat clé.

Dans le cas de l'échantillon $X_1, X_2, \ldots, X_p$ i.i.d. de loi $\mathcal{N}(\mu_1, \sigma^2)$, on sait estimer les paramètres μ_1 et σ^2 par $\overline{X}_p$ et S_X^2 définis par

$$\overline{X}_p = \frac{1}{p}\sum_{j=1}^{p} X_j \quad \text{et} \quad S_X^2 = \frac{1}{p-1}\sum_{j=1}^{p}(X_j - \overline{X}_p)^2$$

De la même manière, dans le cas de l'échantillon $Y_1, Y_2, \ldots, Y_q$ i.i.d. de loi $\mathcal{N}(\mu_2, \sigma^2)$, on sait estimer les paramètres μ_2 et σ^2 par $\overline{Y}_q$ et S_Y^2 définis par

$$\overline{Y}_q = \frac{1}{q}\sum_{j=1}^{q} Y_j \quad \text{et} \quad S_Y^2 = \frac{1}{q-1}\sum_{j=1}^{q}(Y_j - \overline{Y}_q)^2$$

Comme les deux échantillons sont indépendants, on peut facilement déduire que :

$$\frac{\overline{X}_p - \overline{Y}_q - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2}$$

avec $S^2 = \dfrac{1}{p+q-2}\left[(p-1)S_X^2 + (q-1)S_Y^2\right]$.

4.4. Choix de la statistique de test.

On sait que pour tous réels μ_1 et μ_2,

$$\frac{\overline{X}_p - \overline{Y}_q - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2}$$

4.4. Choix de la statistique de test.

On sait que pour tous réels μ_1 et μ_2,

$$\frac{\overline{X}_p - \overline{Y}_q - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2}$$

• Sous H_0, les moyennes théoriques μ_1 et μ_2 sont égales et leur différence est nulle. Donc, sous H_0, la variable Z définie par

$$Z = \frac{\sqrt{pq}\,(\overline{X}_p - \overline{Y}_q)}{S\sqrt{p+q}}$$

suit une loi de Student à $(p+q-2)$ ddl, et on note $Z \underset{H_0}{\sim} T_{p+q-2}$.

4.4. Choix de la statistique de test.

On sait que pour tous réels μ_1 et μ_2,

$$\frac{\overline{X}_p - \overline{Y}_q - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2}$$

• Sous H_0, les moyennes théoriques μ_1 et μ_2 sont égales et leur différence est nulle. Donc, sous H_0, la variable Z définie par

$$Z = \frac{\sqrt{pq}\,(\overline{X}_p - \overline{Y}_q)}{S\sqrt{p+q}}$$

suit une loi de Student à $(p+q-2)$ ddl, et on note $Z \underset{H_0}{\sim} T_{p+q-2}$.

• En revanche, sous H_1, Z ne suit plus une T_{p+q-2} puisque on a la décomposition suivante :

$$Z = \underbrace{\frac{\sqrt{pq}(\overline{X}_p - \overline{Y}_q - (\mu_1 - \mu_2))}{S\sqrt{p+q}}}_{\sim T_{p+q-2}} + \underbrace{\frac{\sqrt{pq}(\mu_1 - \mu_2)}{S\sqrt{p+q}}}_{\neq 0 \text{ sous } H_1}$$

Pour ces raisons, Z est une statistique de test admissible.

4.5 Construction de la zone de rejet du test.

Compte-tenu de la statistique de test (et sa loi sous H_0) et de l'hypothèse alternative, l'idée va être de rejeter l'hypothèse nulle H_0 lorsque l'on observera une valeur de la statistique de test faisant partie des valeurs les moins probables (petites et grandes) d'une loi de Student à $(p+q-2)$ ddl.

On cherche une zone de rejet de l'hypothèse nulle H_0 au risque α de la forme $\{Z < -t\} \bigcup \{Z > t\}$ avec $t > 0$ tel que :

$$\mathbb{P}_{H_0}\left[|Z| > t\right] = \alpha \iff \mathbb{P}_{H_0}\left[|Z| \leq t\right] = 1 - \alpha$$

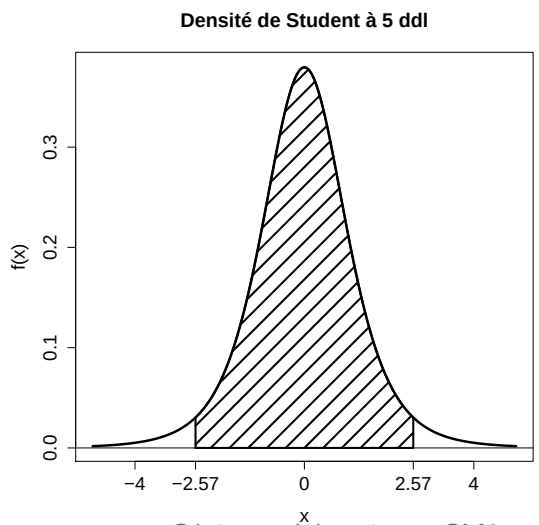
4.5 Construction de la zone de rejet du test.

Compte-tenu de la statistique de test (et sa loi sous H_0) et de l'hypothèse alternative, l'idée va être de rejeter l'hypothèse nulle H_0 lorsque l'on observera une valeur de la statistique de test faisant partie des valeurs les moins probables (petites et grandes) d'une loi de Student à $(p+q-2)$ ddl.

On cherche une zone de rejet de l'hypothèse nulle H_0 au risque α de la forme $\{Z < -t\} \bigcup \{Z > t\}$ avec $t > 0$ tel que :

$$\mathbb{P}_{H_0}\left[|Z| > t\right] = \alpha \iff \mathbb{P}_{H_0}\left[|Z| \leq t\right] = 1 - \alpha$$

Le réel t est ainsi le quantile d'ordre $(1-\alpha/2)$ d'une loi de Student à $(p+q-2)$ ddl, noté $t_{p+q-2,1-\alpha/2}$.



4.6. Construction du test d'égalité de deux moyennes.

Pour tester au risque α l'hypothèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre l'hypothèse alternative $H_1 : \ll \mu_1 \neq \mu_2 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{pq}(\overline{X}_p - \overline{Y}_q)}{S\sqrt{p+q}} \underset{H_0}{\sim} T_{p+q-2}$.

La zone de rejet est de la forme $\{|Z| > t_{p+q-2}\}$ où t_{p+q-2} est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $p+q-2$ ddl, c'est à dire t_{p+q-2} est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{p+q-2}\right] = \mathbb{P}\left[|T_{p+q-2}| \leq t_{p+q-2}\right] = 1-\alpha$$

4.6. Construction du test d'égalité de deux moyennes.

Pour tester au risque α l'hypothèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre l'hypothèse alternative $H_1 : \ll \mu_1 \neq \mu_2 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{pq}(\overline{X}_p - \overline{Y}_q)}{S\sqrt{p+q}} \underset{H_0}{\sim} T_{p+q-2}$.

La zone de rejet est de la forme $\{|Z| > t_{p+q-2}\}$ où t_{p+q-2} est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $p+q-2$ ddl, c'est à dire t_{p+q-2} est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{p+q-2}\right] = \mathbb{P}\left[|T_{p+q-2}| \leq t_{p+q-2}\right] = 1-\alpha$$

- On calcule alors la valeur z_{obs} de Z sur les données :

$$z_{obs} = \frac{\sqrt{pq}(\overline{x}_p - \overline{y}_q)}{s\sqrt{p+q}}$$

4.6. Construction du test d'égalité de deux moyennes.

Pour tester au risque α l'hypothèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre l'hypothèse alternative $H_1 : \ll \mu_1 \neq \mu_2 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{pq}(\overline{X}_p - \overline{Y}_q)}{S\sqrt{p+q}} \underset{H_0}{\sim} T_{p+q-2}$.

La zone de rejet est de la forme $\{|Z| > t_{p+q-2}\}$ où t_{p+q-2} est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $p+q-2$ ddl, c'est à dire t_{p+q-2} est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{p+q-2}\right] = \mathbb{P}\left[|T_{p+q-2}| \leq t_{p+q-2}\right] = 1-\alpha$$

- On calcule alors la valeur z_{obs} de Z sur les données :

$$z_{obs} = \frac{\sqrt{pq}(\overline{x}_p - \overline{y}_q)}{s\sqrt{p+q}}$$

- Si $|z_{obs}| \leq t_{p+q-2}$, alors on ne peut pas rejeter l'hypothèse nulle H_0 selon laquelle les 2 moyennes seraient égales.

4.6. Construction du test d'égalité de deux moyennes.

Pour tester au risque α l'hypothèse nulle H_0 : $\ll \mu_1 = \mu_2 \gg$ contre l'hypothèse alternative H_1 : $\ll \mu_1 \neq \mu_2 \gg$,

on utilise la statistique de test $Z = \dfrac{\sqrt{pq}(\overline{X}_p - \overline{Y}_q)}{S\sqrt{p+q}} \underset{H_0}{\sim} T_{p+q-2}$.

La zone de rejet est de la forme $\{|Z| > t_{p+q-2}\}$ où t_{p+q-2} est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $p+q-2$ ddl, c'est à dire t_{p+q-2} est tel que

$$\mathbb{P}_{H_0}\left[|Z| \leq t_{p+q-2}\right] = \mathbb{P}\left[|T_{p+q-2}| \leq t_{p+q-2}\right] = 1-\alpha$$

• On calcule alors la valeur z_{obs} de Z sur les données :

$$z_{obs} = \frac{\sqrt{pq}(\overline{x}_p - \overline{y}_q)}{s\sqrt{p+q}}$$

• Si $|z_{obs}| \leq t_{p+q-2}$, alors on ne peut pas rejeter l'hypothèse nulle H_0 selon laquelle les 2 moyennes seraient égales.
• Sinon, on rejette l'hypothèse nulle H_0 au risque α, et on considère que les 2 moyennes sont significativement différentes.

4.7. Test unilatéral.

Il est aussi possible de mettre en oeuvre un test unilatéral. On peut en effet tester au risque α :

- $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 < \mu_2 \gg$;
- ou bien $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 > \mu_2 \gg$.

4.7. Test unilatéral.

Il est aussi possible de mettre en oeuvre un test unilatéral. On peut en effet tester au risque α :

- $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 < \mu_2 \gg$;
- ou bien $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 > \mu_2 \gg$.

La statistique de test reste inchangée, mais les zones de rejet sont alors définies respectivement par :

4.7. Test unilatéral.

Il est aussi possible de mettre en oeuvre un test unilatéral. On peut en effet tester au risque α :

- $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 < \mu_2 \gg$;
- ou bien $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 > \mu_2 \gg$.

La statistique de test reste inchangée, mais les zones de rejet sont alors définies respectivement par :

- $\{Z < t_{p+q-2,\alpha}\}$;

4.7. Test unilatéral.

Il est aussi possible de mettre en oeuvre un test unilatéral. On peut en effet tester au risque α :

- $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 < \mu_2 \gg$;
- ou bien $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 > \mu_2 \gg$.

La statistique de test reste inchangée, mais les zones de rejet sont alors définies respectivement par :

- $\{Z < t_{p+q-2,\alpha}\}$;
- $\{Z > t_{p+q-2,1-\alpha}\}$;

où $t_{p+q-2,1-\alpha}$ ($t_{p+q-2,\alpha} = -t_{p+q-2,1-\alpha}$) désigne le quantile d'ordre $(1-\alpha)$ d'une loi de Student à $(p+q-2)$ ddl.

4.8. Lien entre test et Intervalle de confiance.

4.8. Lien entre test et Intervalle de confiance.

Tester au risque α l'hypothèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$, est en fait équivalent à étudier l'appartenance de 0 à l'intervalle de confiance du paramètre $\mu_1 - \mu_2$ au niveau de confiance $(1 - \alpha)$. Plus 0 est éloigné des bornes de cet intervalle, plus le test sera significatif.

4.8. Lien entre test et Intervalle de confiance.

Tester au risque α l'hypothèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$, est en fait équivalent à étudier l'appartenance de 0 à l'intervalle de confiance du paramètre $\mu_1 - \mu_2$ au niveau de confiance $(1-\alpha)$. Plus 0 est éloigné des bornes de cet intervalle, plus le test sera significatif.

En effet, si t désigne le quantile d'ordre $(1-\alpha/2)$ d'une Student à $p+q-2$ ddl, on a l'équivalence suivante :

$$
\begin{aligned}
|z_{obs}| \leq t \quad &\Longleftrightarrow \quad -t \leq \frac{\sqrt{pq}(\bar{x}_p - \bar{y}_q)}{s\sqrt{p+q}} \leq t \\
&\Longleftrightarrow \quad \bar{x}_p - \bar{y}_q - \frac{s\,t\sqrt{p+q}}{\sqrt{pq}} \leq 0 \leq \bar{x}_p - \bar{y}_q + \frac{s\,t\sqrt{p+q}}{\sqrt{pq}} \\
&\Longleftrightarrow \quad 0 \in \mathrm{IC}_{1-\alpha}(\mu_1 - \mu_2)
\end{aligned}
$$

Ainsi, on peut conclure sur le test en regardant si 0 appartient ou non à l'intervalle de confiance.

4.9. Discussion sur la mise en oeuvre du test.

La mise en oeuvre du test d'égalité de 2 moyennes requiert les hypothèses suivantes :

4.9. Discussion sur la mise en oeuvre du test.

La mise en oeuvre du test d'égalité de 2 moyennes requiert les hypothèses suivantes :

- Normalité des échantillons: les données doivent provenir de variables distribuées selon une loi normale. Il est possible de tester préalablement la normalité de chaque échantillon, à l'aide d'un test de normalité, Shapiro-Wilk par exemple.

4.9. Discussion sur la mise en oeuvre du test.

La mise en oeuvre du test d'égalité de 2 moyennes requiert les hypothèses suivantes :

- Normalité des échantillons: les données doivent provenir de variables distribuées selon une loi normale. Il est possible de tester préalablement la normalité de chaque échantillon, à l'aide d'un test de normalité, Shapiro-Wilk par exemple.
- Indépendance des échantillons: Les 2 échantillons doivent être indépendants : c'est le protocole expérimental qui doit l'assurer. Lorsque les données sont par exemple appariées, on utilise une autre statistique de test ;

4.9. Discussion sur la mise en oeuvre du test.

La mise en oeuvre du test d'égalité de 2 moyennes requiert les hypothèses suivantes :

- Normalité des échantillons: les données doivent provenir de variables distribuées selon une loi normale. Il est possible de tester préalablement la normalité de chaque échantillon, à l'aide d'un test de normalité, Shapiro-Wilk par exemple.
- Indépendance des échantillons: Les 2 échantillons doivent être indépendants : c'est le protocole expérimental qui doit l'assurer. Lorsque les données sont par exemple appariées, on utilise une autre statistique de test ;
- Egalité des variances: les variances des 2 échantillons doivent être égales : on peut le vérifier à l'aide d'un test de comparaison de deux variances, le test de Fisher par exemple. Lorsque les variances sont significativement différentes, on utilise un autre test pour comparer les moyennes (test de Welch, par exemple).

5. Exemple : Les Iris de Fisher.
5.1. Le jeu de données.

"Les iris de Fisher" sont des données fameuses collectées par Edgar Enderson, et proposées en 1933 par le statisticien Ronald Aylmer Fisher comme données de référence pour l'analyse discriminante et la classification.
Il s'agit de reconnaître le type d'iris (*setosa*, *virginica* et *versicolore*) à partir seulement de la longueur et de la largeur de ses pétales et sépales.

Le jeu de données est constitué de 150 individus (50 fleurs de chaque type) et 5 variables :

1. Sepal.Length (en mm) **2.** Sepal.Width (en mm) **3.** Petal.Length (en mm) **4.** Petal.Width (en mm) **5.** Species : *setosa*, *virginica* et *versicolore*.

5.2. Le problème.

On s'intéresse ici à la variable `Sepal.Length` pour les variétés d'iris `Virginica` et `Versicolor`.

On veut savoir si la longueur moyenne des sépales diffère d'une variété d'iris à l'autre.

5.2. Le problème.

On s'intéresse ici à la variable `Sepal.Length` pour les variétés d'iris `Virginica` et `Versicolor`.

On veut savoir si la longueur moyenne des sépales diffère d'une variété d'iris à l'autre.

On notera $x_1, x_2, \ldots, x_{50}$ les longueurs des sépales des iris de la variété `Virginica` et $y_1, y_2, \ldots, y_{50}$ celles de la variété `Versicolor`.

Le tableau ci-dessous regroupe les statistiques de base de chacune des séries de mesures.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std
Virginica	4.90	5.60	5.90	5.94	6.30	7.00	0.52
Versicolor	4.90	6.23	6.50	6.59	6.90	7.90	0.64

On peut constater que les moyennes observées sont différentes : la longueur moyenne des sépales des Iris Versicolor est plus grande.

On souhaite confirmer cela à l'aide d'un test de Student au risque 5%.

5.3. Hypothèses.

On supposera que pour chaque variété d'iris, la longueur d'un sépale est distribuée selon une loi normale, et que la variabilité est la même dans chaque variété d'iris.

• Autrement dit, on peut considérer que les données $x_1, x_2, \ldots, x_{50}$ sont les réalisations des variables aléatoires $X_1, X_2, \ldots, X_{50}$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma^2)$ et que les données $y_1, y_2, \ldots, y_{50}$ sont les réalisations des variables aléatoires $Y_1, Y_2, \ldots, Y_{50}$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma^2)$.

• On suppose de plus que les deux échantillons sont indépendants.

5.5. Mise en oeuvre du test d'égalité de 2 moyennes.

Pour tester au risque 5% l'hypothèse selon laquelle les longueurs moyenne des sépales sont les mêmes, on va tester l'hypotèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$.

5.5. Mise en oeuvre du test d'égalité de 2 moyennes.

Pour tester au risque 5% l'hypothèse selon laquelle les longueurs moyenne des sépales sont les mêmes, on va tester l'hypotèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$.

- Pour cela, on utilise la statistique de Student Z définie par :

$$Z = \frac{\sqrt{50 \times 50}(\overline{X}_{50} - \overline{Y}_{50})}{S\sqrt{50+50}} \underset{H_0}{\sim} T_{50+50-2}$$

5.5. Mise en oeuvre du test d'égalité de 2 moyennes.

Pour tester au risque 5% l'hypothèse selon laquelle les longueurs moyenne des sépales sont les mêmes, on va tester l'hypotèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$.

- Pour cela, on utilise la statistique de Student Z définie par :

$$Z = \frac{\sqrt{50 \times 50}(\overline{X}_{50} - \overline{Y}_{50})}{S\sqrt{50+50}} \underset{H_0}{\sim} T_{50+50-2}$$

- Le quantile d'ordre 0,975 d'une loi de Student à 98 ddl est égal à : 1,98.

5.5. Mise en oeuvre du test d'égalité de 2 moyennes.

Pour tester au risque 5% l'hypothèse selon laquelle les longueurs moyenne des sépales sont les mêmes, on va tester l'hypotèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$.

- Pour cela, on utilise la statistique de Student Z définie par :

$$Z = \frac{\sqrt{50 \times 50}(\overline{X}_{50} - \overline{Y}_{50})}{S\sqrt{50+50}} \underset{H_0}{\sim} T_{50+50-2}$$

- Le quantile d'ordre 0,975 d'une loi de Student à 98 ddl est égal à : 1,98.
- La zone de rejet de H_0 est donc de la forme : $\{|Z| > 1,98\}$.

5.5. Mise en oeuvre du test d'égalité de 2 moyennes.

Pour tester au risque 5% l'hypothèse selon laquelle les longueurs moyenne des sépales sont les mêmes, on va tester l'hypotèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$.

- Pour cela, on utilise la statistique de Student Z définie par :

$$Z = \frac{\sqrt{50 \times 50}(\overline{X}_{50} - \overline{Y}_{50})}{S\sqrt{50+50}} \underset{H_0}{\sim} T_{50+50-2}$$

- Le quantile d'ordre 0,975 d'une loi de Student à 98 ddl est égal à : 1,98.
- La zone de rejet de H_0 est donc de la forme : $\{|Z| > 1,98\}$.
- On calcule maintenant la valeur z_{obs} de Z sur les données. Le tableau des statistiques de base nous permet de calculer la valeur s de D sur les données :

$$s^2 = \left(49 \times 0,52^2 + 49 \times 0,64^2\right)/98 = 0,3354$$

5.5. Mise en oeuvre du test d'égalité de 2 moyennes.

Pour tester au risque 5% l'hypothèse selon laquelle les longueurs moyenne des sépales sont les mêmes, on va tester l'hypotèse nulle $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$.

- Pour cela, on utilise la statistique de Student Z définie par :

$$Z = \frac{\sqrt{50 \times 50}(\overline{X}_{50} - \overline{Y}_{50})}{S\sqrt{50+50}} \underset{H_0}{\sim} T_{50+50-2}$$

- Le quantile d'ordre 0,975 d'une loi de Student à 98 ddl est égal à : 1,98.
- La zone de rejet de H_0 est donc de la forme : $\{|Z| > 1,98\}$.
- On calcule maintenant la valeur z_{obs} de Z sur les données. Le tableau des statistiques de base nous permet de calculer la valeur s de D sur les données :

$$s^2 = \left(49 \times 0,52^2 + 49 \times 0,64^2\right)/98 = 0,3354$$

On en déduit donc la valeur de z_{obs} qui est égale à $z_{obs} = -5,63$.
Puisque z_{obs} est dans la zone de rejet, on rejette l'hypothèse nulle H_0 au risque 5% , on considère donc que les longueurs moyennes des sépales sont différentes d'une variété d'iris à l'autre.

5.6. Avec le logiciel R.

Avec le logiciel R, on utilise l'instruction :

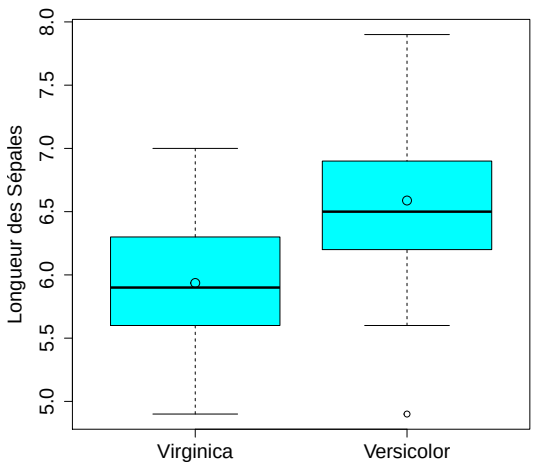
```
t.test(Sepal.Length ~Species,var.equal=TRUE,
        data=iris[51:150,])
```

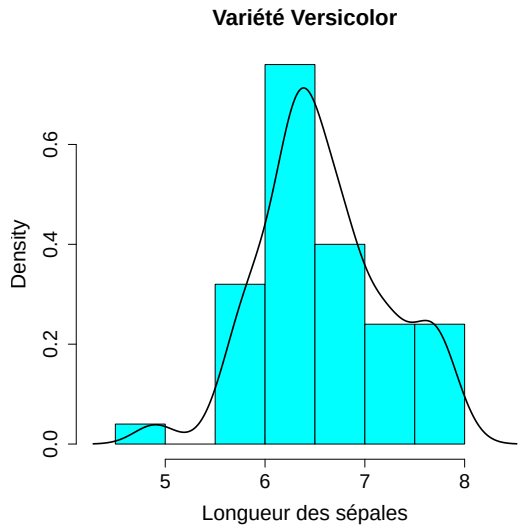
On obtient alors le résultat suivant :

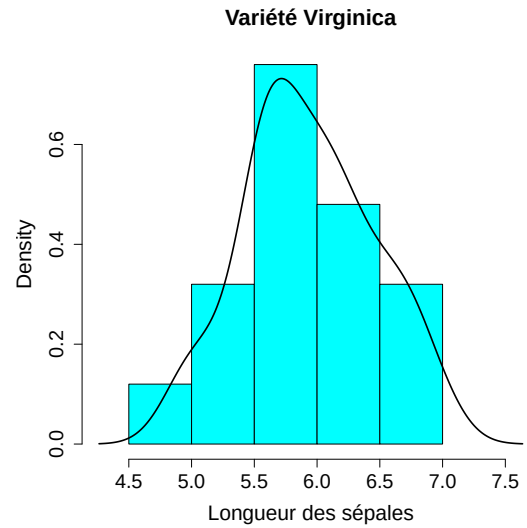
```
^^ITwo Sample t-test

data:  Sepal.Length by Species
t = -5.6292, df = 98, p-value = 1.725e-07
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
-0.8818516 -0.4221484
sample estimates:
mean in group versicolor  mean in group virginica
                   5.936                      6.588
```

5.7. Vérification empirique des hypothèses.

On trouvera ci-dessous les boîtes à moustaches des 2 séries de mesures ainsi que les histogrammes en fréquences.
Que pensez-vous des hypothèses introduites ?







5.8. Commentaires.

- Les boites à moustaches montrent des distributions assez symétriques, une variabilité assez comparable et laissent penser que les moyennes théoriques sont significativement différentes.
- Les histogrammes en fréquences laissent à penser que les deux variétés sont distribuées selon une loi normale. Ceux-ci ont en effet la forme d'une gaussienne. Ceci serait bien sûr à confirmer à l'aide d'un test statistique car les échantillons sont petits.