Université de Strasbourg

UFR de Mathématiques

et d'Informatique

Han-Ping LI

Année 2019/2020

L3

Statistique

Etude de cas

## Chapitre IV Tests d'hypothèses

Exemple 1 : L'étiquette d'une bouteille de 75 cl de jus d'orange (d'une certaine marque) indique que le jus d'orange contient en moyenne, au plus un gramme de matière grasse. On prélève $n = 30$ bouteilles de la même marque, en trouve

0.99,1.19,1.03,1.10,0.97,0.79,0.87,1.46,1.02,0.95,
1.09,0.85,1.18,0.81,0.96,1.22,0.72,1.13,1.23,1.05,
1.36,1.32,1.21,1.02,1.36,0.97,1.21,1.06,1.31,1.01

Après une étude statistique (test de la normalité), on peut raisonnablement supposer que l'échantillon est issu d'une population d'une loi normalement distribuée $\mathcal{N}(\mu, \sigma^2)$. On souhaite tester

$\mathbf{H}_0 : \mu \leq 1$ contre $\mathbf{H}_1 \ : \mu > 1$.
On se donne $\alpha = 0.05$, ,
On utilise $T = \frac{\sqrt{n}(\overline{X}-1)}{S_c}$ qui suit, lorsque $\mu = 1$ , d'une loi de Student à (n-1) degrés de liberté $\mathcal{T}_{n-1}$.

$$\mathbb{P}(T > C_q) = 0.05 \Longrightarrow C_q = \texttt{qt}(0.95, \texttt{df} = 29) = 1.699127$$

l'intervalle de rejet $= ]1.699, \infty[$.

$$\text{on décidera donc de} \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ t > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } \ t \leq C_q. \end{cases}$$

Comme $\overline{x} = 1.081333$, $s_c = 0.1834297$, on a

$$t = \frac{\sqrt{n}(\overline{x}-1)}{S_c} = \frac{\sqrt{30}(1.081-1)}{0.183} = 2.4286 > 1.699!$$

on rejette donc $\mathbf{H}_0$.

Une hypothèse statistique est une affirmation ou sa forme négative concernant les valeurs du paramètre $\theta$ inconnu.

On souhaite étudier si les données de l'échantillon recueilli sont compatibles ou non avec une hypothèse faite sur la famille des lois de la population.

**Hypothèse nulle $\mathbf{H}_0$** : il s'agit d'une proposition des valeurs particulières constituée d'un ensemble fermé sur le paramètre inconnu ;

**Hypothèse alternative $\mathbf{H}_1$** une autre proposition qui s'oppose à l'hypothèse $\mathbf{H}_0$.

**Pour Fisher, un des fondateurs de la statistique, c'est l'hypothèse alternative $\mathbf{H}_1$ que l'on veut défendre**.

Exemples :

• soit $\mathbf{H}_0 : \theta = \theta_0$ contre $\mathbf{H}_1 : \theta \neq \theta_0$ (bilatéral)

• soit $\mathbf{H}_0$ : $\theta \leq \theta_0$ contre $\mathbf{H}_1$ : $\theta > \theta_0$ (unilatéral à droite)

• soit $\mathbf{H}_0$ : $\theta \geq \theta_0$ contre $\mathbf{H}_1$ : $\theta < \theta_0$ (unilatéral gauche)

**Un test statistique** est une démarche qui a pour but de fournir une règle de décision permettant, en utilisant une bonne statistique de l'échantillon, de rejeter ou conserver l'hypothèse nulle avec une faible marges du risque lorsque celle-ci est rejetée.

Pour tester si l'hypothèse nulle formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celles selon

l'hypothèse nulle est trop important pour être uniquement imputable au hasard de l'échantillonnage.

La construction de la région de rejet consiste à déterminer entre quelles valeurs peut varier la statistique du test selon l'hypothèse nulle, sur la seule considération du hasard de l'échantillonnage.

La prise de décision n'est jamais parfaite et des erreurs peuvent être commises.

| **nature** *désision* | $\mathbf{H}_0$ **est vraie** | $\mathbf{H}_0$ **est fausse** |
|---|---|---|
| *conserver* $\mathbf{H}_0$ | décision correcte | erreur de 2ème espèce |
| *rejeter* $\mathbf{H}_0$ | erreur de 1ère espèce | décision correcte |

• **erreur de 1ère espèce** = l'erreur de rejeter $\mathbf{H}_0$ alors que $\mathbf{H}_0$ est vraie (càd sous $\mathbf{H}_0$) ;

• **erreur de 2ème espèce** = l'erreur de conserver $\mathbf{H}_0$ alors que $\mathbf{H}_0$ est fausse (càd sous $\mathbf{H}_1$) ;

• **le risque de 1ère espèce** : $r_1 = \mathbb{P}(\text{rejeter } \mathbf{H}_0 | \mathbf{H}_0 \text{ est vraie})$.

• **le risque de 2ème espèce** : $r_2 = \mathbb{P}(\text{conserver } \mathbf{H}_0 | \mathbf{H}_0 \text{ est fausse})$

L'hypothèse $\mathbf{H}_0$ peut être vraie de manière différente. Il y a en général une infinité de possibilités. Si on souhaite tester $\mathbf{H}_0$ $\theta \leq 10$ contre $\mathbf{H}_1$ : $\theta > 10$, alors $\theta = 5; 7, ..., 10$ sont toutes les valeurs vérifiant la proposition de $\mathbf{H}_0$, donc il y aura un risque de première espèce $r_1(5), r_1(7), ..., r_1(10)$ correspondant à chacune de ces valeurs.

•**le niveau du test (ou le seuil de significativité)** = $\max\{r_1(\theta) : \theta$ vérifie la proposition de $\mathbf{H}_0\}$.

On introduit aussi

• **les puissances du test** = $\mathbb{P}$(rejeter $\mathbf{H}_0$ à raison$|\mathbf{H}_0$ est fausse).

L'hypothèse $\mathbf{H}_0$ peut être fausse de manière très différente. Il y a en général une infinité de possibilités. Si

on souhaite tester $\mathbf{H}_0$ $\theta \leq 10$ contre $\mathbf{H}_1$ : $\theta > 10$, alors $\theta = 11, 12, ..., 100$ sont toutes les valeurs vérifiant la proposition de $\mathbf{H}_1$, donc il y aura un risque de deuxième espèce $r_2(11), r_2(12), ..., r_2(100)$ correspondant à chacune de ces valeurs. Ce sont des risques que l'on contrôle assez mal.

On cherche donc parmi tous les tests de niveau $\alpha$, celui qui minimise uniformément les risques de deuxième espèce (si possible), par conséquent, maximise uniformément les puissances.

Cette démarche comporte 5 étapes :

1. formulation de l'hypothèse nulle et l'hypothèse alternative

2 choix d'un seuil $\alpha$ qui contrôle la probabilité de rejeter $\mathbf{H}_0$ à tort .

3 Choix d'une statistique de test $T(X_1, \ldots, X_n)$

i ) qui est très sensible aux hypothèses : elle se comporte très différemment selon que l'hypothèse est vraie ou fausse ;

ii) dont la loi est connue ou approximativement connue

4 Détermination d'une région de rejet (une règle de décision )

5 Sur la base d'une réalisation de $T$, on porte un jugement sur l'hypothèse testée.

## § 1. Comparaison de $\mu$ avec une référence $\mu_0$ où $\mu$ est l' espérance d'une loi normale $\mathcal{N}(\mu, \sigma^2)$.

On rappelle que la vraie valeur de $\mu$ est inconnue, notée toujours par $\mu$.

### 1.1 test unilatéral à droite

1) On souhaite tester $\mathbf{H}_0 \ : \mu \leq \mu_0$ contre $\mathbf{H}_1 \ : \mu > \mu_0$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$$

(*déduite du rapport des vraisemblances maximales* )

i) la loi de T, lorsque $\mu = \mu_0$, est une loi de Student ;

ii) elle est sensible aux hypothèses :

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S_c} + \frac{\sqrt{n}(\mu - \mu_0)}{S_c}$$

dont le premier terme, étant exactement une v.a. de loi de Student, a un comportement neutre ; dont le second terme a un comportement très différent selon que $\mathbf{H}_0$ est vraie ou fausse. En effet

$$\frac{\sqrt{n}(\mu-\mu_0)}{S_c}\begin{cases} > 0 & \text{si } \mu > \mu_0 \text{ (sous } \mathbf{H}_1\text{)}; \\ \leq 0 & \text{si } \mu \leq \mu_0 \text{ (sous } \mathbf{H}_0\text{)}. \end{cases}$$

Autrement dit,

• la statistique $T$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $T$ aura tendance à prendre des valeurs **faibles** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(T > C_q\Big) = \alpha$ où $T$ suit une loi de Student à (n-1) degrés de liberté. Ainsi

$$C_q = \mathrm{qt}(1-\alpha, \mathrm{n}-1)$$

5) On calcule la valeur $t = T(x_1, \ldots, x_n)$

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \; t > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } \; t \leq C_q. \end{cases}$$

**Remarque :**

Pourquoi la statistique $T$ ? On sait que le meilleur estimateur de $\mu$ est $\overline{X}$ mais on ne peut pas l'utiliser pour le test puisque

$\mathcal{L}(\overline{X}) = \mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ qui dépend de deux paramètres que on ne connait pas ;

$\mathcal{L}\left(\dfrac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}\right) = \mathcal{N}(0,1)$ sous $\mathbf{H}_0$ qui dépend encore d'un paramètre ;

$\mathcal{L}\left(\dfrac{\sqrt{n}(\overline{X}-\mu_0)}{S_c}\right) = \mathcal{T}_{(n-1)}$ sous $\mathbf{H}_0$, qui est enfin indépendant de paramètre.

La déduction de $T$ est guidée en fait par le rapport des vraisemblances maximales.

Exemple : $n = 15, \alpha = 0.05$ alors
$C_q = \mathtt{qt}(0.95, \mathtt{df} = 14) = 1.76131.$

Exemple 1 : L'étiquette d'une bouteille de 75 cl de jus d'orange (d'une certaine marque) indique que le

jus d'orange contient en moyenne, au plus un gramme de matière grasse. On prélève $n = 30$ bouteilles de la même marque, en trouve

0.99,1.19,1.03,1.10,0.97,0.79,0.87,1.46,1.02,0.95,
1.09,0.85,1.18,0.81,0.96,1.22,0.72,1.13,1.23,1.05,
1.36,1.32,1.21,1.02,1.36,0.97,1.21,1.06,1.31,1.01

Après une étude statistique (test de la normalité), on peut raisonnablement supposer que l'échantillon est issu d'une population d'une loi normalement distribuée $\mathcal{N}(\mu, \sigma^2)$.
On souhaite tester

$\mathbf{H}_0 : \mu \leq 1$ contre $\mathbf{H}_1 \ : \mu > 1$.
On se donne $\alpha = 0.05$, ,

On utilise $T = \frac{\sqrt{n}(\overline{X}-1)}{S_c}$ qui suit, lorsque $\mu = 1$ , $\mathcal{T}_{n-1}$

$$\mathbb{P}(T > C_q) = 0.05 \Longrightarrow C_q = \texttt{qt}(0.95, \texttt{df} = 29) = 1.699127.$$

La région de rejet $= ]1.699, \infty[$.

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ t > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } \ t \leq C_q. \end{cases}$$

Comme $\overline{x} = 1.081333$, $s_c = 0.1834297$, on a

$$t = \frac{\sqrt{n}(\overline{x}-1)}{S_c} = \frac{\sqrt{30}(1.081-1)}{0.183} = 2.4286 > 1.699!$$

on rejette donc $\mathbf{H}_0$.

p-valeur $= \mathbf{P}(T \geq 2.4286) = 0,01079438.$

**La p-valeur d'un test**.

— Dans la pratique, les logiciels renvoient toujours un nombre $p \in ]0,1[$, appelé **p-valeur** du test. Il s'agit de **la probabilité** d'observer une valeur (une réalisation) $t$ de la statistique de test $T$ par exemple, **au moins aussi défavorable** que $t$ pour l'hypothèse nulle.

Formellement, si $T = t_0$, si la région de rejet est de la forme $\{|T| > C\}$, alors

$$\text{p-valeur} = \mathbf{P}(|T| \geq |t_0|, \text{ si } \mathbf{H}_0 \text{ est vraie}).$$

Si la p-valeur est supérieure ou égale au seuil, alors $\mathbf{H}_0$( continue à bénéficier de la présomption de vérité, elle sera alors conservée. Si la p-valeur est inférieure au seuil, il représente approximativement alors **le risque**

**de première espèce** $r_1$, c'est-à-dire que la probabilité de rejeter l'hypothèse nulle à tort.

On a donc

"rejet de $\mathbf{H}_0$" $\Longleftrightarrow$ "$t \in$ région de rejet " $\Longleftrightarrow$ "p-valeur $< \alpha$"

Plus p-valeur est petite, plus la probabilité de commettre l'erreur de première espèce est faible. Si on utilise un seuil $\alpha$, alors vous pouvez rejeter l'hypothèse nulle si et seulement si la p-valeur est inférieure à $\alpha$.

## 1.2 test unilatéral à gauche

1) On souhaite tester $\mathbf{H}_0 \ : \mu \geq \mu_0$ contre $\mathbf{H}_1 \ : \mu < \mu_0$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$$

i) la loi de T, lorsque $\mu = \mu_0$ est une loi de Student ;

ii) elle est sensible aux hypothèses :

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S_c} + \frac{\sqrt{n}(\mu - \mu_0)}{S_c}$$

dont le premier terme, étant exactement une v.a. de loi de Student, a un comportement neutre ; dont le second terme a un comportement très différent selon que $\mathbf{H}_0$ est vraie ou fausse. En effet

$$\frac{\sqrt{n}(\mu - \mu_0)}{S_c} \begin{cases} < 0 & \text{si } \mu < \mu_0 \text{ (sous } \mathbf{H}_1\text{)}; \\ \geq 0 & \text{si } \mu \geq \mu_0 \text{ (sous } \mathbf{H}_0\text{)}. \end{cases}$$

Autrement dit,

• la statistique $T$ aura tendance à prendre des valeurs plus **faibles** lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $T$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = -C_q < 0$ tel que $\mathbb{P}\Big(T < -C_q\Big) = \alpha$ où $T$ suit une loi de Student à (n-1) degrés de liberté. Ainsi

$$C_q = -\mathrm{qt}(1-\alpha, \mathrm{n}-1) = \mathrm{qt}(\alpha, \mathrm{n}-1).$$

5) On calcule la valeur $t = T(x_1, \ldots, x_n)$

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ t < -C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } \ t \geq -C_q. \end{cases}$$

Exemple : $n = 20, \alpha = 0.05$ alors

$C_q = \texttt{qt}(0.05, \texttt{df} = 19) = -\texttt{qt}(0.95, \texttt{df} = 19) = -1.729133.$

Exemple 2 : l'université a reçu un envoi en masse de $n = 400$ mails de xx@xxxx.fr. On souhaite savoir si xx@xxxx.fr est un spammeur ( c' à d que le score $>$ 2500).

$\mathbf{H}_0 : \mu \geq 2500$ contre $\mathbf{H}_1 \ : \mu < 2500$

On suppose que les $X_i$ suivent une loi normale $\mathcal{N}(\mu, \sigma^2)$.

On se donne $\alpha = 0.05$, ,

On utilise $T = \frac{\sqrt{n}(\overline{X} - 2500)}{S_c}$ qui suit, lorsque $\mu = 2500$, une loi $\mathcal{T}_{n-1}$

$\mathbb{P}(T < C_q) = 0.05 \Longrightarrow$
$C_q = -\texttt{qt}(0.95, \texttt{df} = 399) = -1.648682$
l'intervalle de rejet $= ]-\infty, -1.65[$.

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ t < C_q, \\ \\ \text{conserver } \mathbf{H}_0 & \text{si } \ t \geq C_q. \end{cases}$$

Après avoir calculé les scores sur ces 400 mails, on trouve $\overline{x} = 2505$ et $s_c^2 = 3293$, on a

$$t = \frac{\sqrt{n}(\overline{x} - 2500)}{s_c} = \frac{\sqrt{400}(2505 - 2500)}{\sqrt{3293}} = 1.742626 > -1.65.$$

On conserve donc $\mathbf{H}_0$.

**Remarques** : Acceptation/non rejet. En général, un test significatif amène à rejeter une hypothèse mais un test non significatif n'amène jamais à accepter d'emblée une hypothèse à cause du risque de 2ème espèce.

L'hypothèse $\mathbf{H}_0$, dite hypothèse nulle, joue un rôle particulier : le but du test est de réunir suffisamment de preuves au sein des données pour démontrer qu'elle est fausse. Si c'est le cas, l'hypothèse nulle est rejetée. Dans le cas contraire, on ne rejette pas $\mathbf{H}_0$, et on dit sobrement la conserver par « faute de preuves pour la contredire », c'est-à-dire que les données ne sont pas incompatibles avec cette hypothèse... ce qui ne veut pas dire qu'elle est vraie !

Imaginez que vous perdiez un bouton de chemise. Si vous en trouvez un par hasard, vous pouvez faire l'hypothèse $\mathbf{H}_0$ « le bouton trouvé est mon bouton perdu

». Vous pouvez faire des tests (sur la taille, la couleur, la forme etc). Si l'un de ces tests est négatif, alors vous rejetterez l'hypothèse $\mathbf{H}_0$. Mais tous les tests positifs ne pourront rarement prouver que l'hypothèse $\mathbf{H}_0$ est vraie (au maximum, ils créeront une présomption de vérité pour $\mathbf{H}_0$ mais sans certitude).

Le choix de $\mathbf{H}_0$ est parfois dicté par le bon sens ; par exemple imaginons un diagnostic pour une maladie grave : – déclarer malade d'un patient sain entraîne des traitements désagréables. – déclarer sain d'un patient malade entraîne des conséquences plus graves. Dans ce cas mieux vaut poser : $\mathbf{H}_0$ « le patient est malade » puisque l'on peut contrôler le risque de première espèce, qui correspond à l'erreur de la plus lourde de conséquences.

$\mathbf{H}_0 : \mu \leq 2500$ contre $\mathbf{H}_1 \ : \mu > 2500$

On suppose que les $X_i$ suivent une loi normale $\mathcal{N}(\mu, \sigma^2)$.

On se donne $\alpha = 0.05$, ,

On utilise $T = \frac{\sqrt{n}(\overline{X} - 2500)}{S_c}$ qui suit, lorsque $\mu = 2500$, une loi $\mathcal{T}_{n-1}$

$\mathbb{P}(T > C_q) = 0.05 \Longrightarrow$
$C_q = \texttt{qt}(0.95, \texttt{df} = 399) = 1.648682$

l'intervalle de rejet $= ]1.65, \infty[$.

$$\text{on décidera donc de} \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } t > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } t \leq C_q. \end{cases}$$

Après avoir calculé les scores sur ces 400 mails, on trouve $\overline{x} = 2505$ et $s_c^2 = 3293$, on a

$$t = \frac{\sqrt{n}(\overline{x} - 2500)}{s_c} = \frac{\sqrt{400}(2505 - 2500)}{\sqrt{3293}} = 1.742626 > 1.65.$$

On rejette donc $\mathbf{H}_0$ au seuil de significativité de 5%.

## 1.3 test bilatéral

1) On souhaite tester $\mathbf{H}_0$ : $\mu = \mu_0$ contre $\mathbf{H}_1$ : $\mu \neq \mu_0$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$$

i) la loi de T, lorsque est vraie est une loi de Student ;

ii) Elle est sensible aux hypothèses :

Si on note la vraie valeur du paramètre par $\mu^*$

$$T = \frac{\sqrt{n}(\overline{X} - \mu^*)}{S_c} + \frac{\sqrt{n}(\mu^* - \mu_0)}{S_c}$$

dont le premier terme, étant une v.a. de loi de Student, a un comportement neutre ; dont le second terme a un comportement très différent selon que $\mathbf{H}_0$ est vraie ou fausse. En effet

$$\frac{\sqrt{n}(\mu - \mu_0)}{S_c} \begin{cases} > 0 & \text{si } \mu > \mu_0 \text{ (sous } \mathbf{H}_1\text{)}, \\ < 0 & \text{si } \mu < \mu_0 \text{ (sous } \mathbf{H}_1\text{)}; \\ = 0 & \text{si } \mu = \mu_0 \text{ (sous } \mathbf{H}_0\text{)}. \end{cases}$$

Autrement dit,

• la statistique $T$ aura tendance à prendre des valeurs soit plus **élevées** soit plus **faibles** lorsque $\mathbf{H}_0$ est fausse

• la statistique $T$ aura tendance à prendre des valeurs **modérées** lorsque $\mathbf{H}_0$ est vraie

4) Par conséquent, on va déterminer les deux constances $C_1 = -C_q$ et $C_2 = C_q$ tel que $\mathbb{P}\Big(C_1 < T(x) < C_2\Big) = 1-\alpha$ où $T$ suit une loi de Student à (n-1) degrés de liberté. Ainsi

$$C_q = \mathrm{qt}(1 - \alpha/2, \mathrm{n} - 1.$$

5) On calcule la valeur $t = T(x_1, \ldots, x_n)$ et

on décidera donc de $\begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ t \notin [-C_q, C_q], \\ \text{conserver } \mathbf{H}_0 & \text{si } \ t \in [-C_q, C_q]. \end{cases}$

Exemple : $n = 24, \alpha = 0.01$ alors
$C_q = \texttt{qt}(0.995, \texttt{df} = 23) = 2.807336$, et $-C_q = \texttt{qt}(0.005, \texttt{df} = 23) = -2.807336$.

## § 2. Comparaison de $\sigma^2$ avec une référence $\sigma_0^2$ où $\sigma^2$ est la variance théorique d'une loi normale $\mathcal{N}(\mu, \sigma^2)$

On rappelle que la vraie valeur de $\sigma^2$ est inconnue, notée toujours par $\sigma^2$.

### 2.1 test unilatéral à droite

1) On souhaite tester $\mathbf{H}_0 : \sigma^2 \leq \sigma_0^2$ contre $\mathbf{H}_1 : \sigma^2 > \sigma_0^2$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$K^2 = \frac{(n-1)S_c^2}{\sigma_0^2}$$

déduite du rapport des vraisemblances maximales

i) la loi de $K^2$, lorsque $\sigma^2 = \sigma_0^2$, est une loi de Khi-deux ;

ii) elle est sensible aux hypothèses :

$$K^2 = \frac{(n-1)S_c^2}{\sigma^2} \times \frac{\sigma^2}{\sigma_0^2}$$

dont le premier facteur positif, étant exactement une v.a. de loi de khi-deux, a un comportement neutre ; dont le second facteur a un comportement très différent selon que $\mathbf{H}_0$ est vraie ou fausse. En effet

$$\left(\frac{\sigma^2}{\sigma_0^2}\right) \quad \begin{cases} > 1 & \text{si } \sigma^2 > \sigma_0^2 \text{ (sous } \mathbf{H}_1\text{)}; \\ \\ \leq 1 & \text{si } \sigma^2 \leq \sigma_0^2 \text{ (sous } \mathbf{H}_0\text{)}. \end{cases}$$

Autrement dit,

• la statistique $K^2$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $K^2$ aura tendance à prendre des valeurs **faibles** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(K^2 > C_q\Big) = \alpha$ où $K$ suit une loi de Khi-deux à (n-1) degrés de liberté. Ainsi

$$C_q = \text{qchisq}(1-\alpha, \text{n}-1).$$

5) On calcule la valeur $k^2 = K^2(x_1, \ldots, x_n)$ et

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } k^2 > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } k^2 \leq C_q. \end{cases}$$

Exemple : $n = 15, \alpha = 0.05$ alors $C_q = \texttt{qchisq}(0.95, \texttt{df} = 14) = 23.68479$.

## 2.2 test unilatéral à gauche

1) On souhaite tester $\mathbf{H}_0 : \sigma^2 \geq \sigma_0^2$ contre $\mathbf{H}_1 : \sigma^2 < \sigma_0^2$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$K^2 = \frac{(n-1)S_c^2}{\sigma_0^2}$$

i) la loi de $K^2$, lorsque $\sigma^2 = \sigma_0^2$, est une loi de Khi-deux ;

ii) elle est sensible aux hypothèses :

$$K^2 = \frac{(n-1)S_c^2}{\sigma^2} \times \frac{\sigma^2}{\sigma_0^2}$$

dont le premier facteur positif, étant exactement une v.a. de loi de khi-deux, a un comportement neutre ; dont le second facteur a un comportement très différent selon que $\mathbf{H}_0$ est vraie ou fausse. En effet

$$\left(\frac{\sigma^2}{\sigma_0^2}\right) \quad \begin{cases} < 1 & \text{si } \sigma^2 < \sigma_0^2 \text{ (sous } \mathbf{H}_1\text{)}; \\ \geq 1 & \text{si } \sigma^2 \geq \sigma_0^2 \text{ (sous } \mathbf{H}_0\text{)}. \end{cases}$$

Autrement dit,

• la statistique $K^2$ aura tendance à prendre des valeurs plus **faibles** lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $K^2$ aura tendance à prendre des valeurs **élevés** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(K^2 < C_q\Big) = \alpha$ où $K$ suit une loi de Khi-deux à (n-1) degrés de liberté. Ainsi

$$C_q = -\text{qchisq}(1-\alpha, \text{n}-1) = \text{qt}(\alpha, \text{n}-1).$$

5) On calcule la valeur $k^2 = K^2(x_1, \ldots, x_n)$ et

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ k^2 < C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } \ k^2 \geq C_q. \end{cases}$$

Exemple : $n = 24, \alpha = 0.01$ alors $C_q = \texttt{qchisq}(0.01, 23) = 10.19572$.

## 2.3 test bilatéral

1) On souhaite tester $\mathbf{H}_0 : \sigma^2 = \sigma_0^2$ contre $\mathbf{H}_1 : \sigma^2 \neq \sigma_0^2$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$K^2 = \frac{(n-1)S_c^2}{\sigma_0^2}$$

i) la loi de $K^2$, lorsque $\sigma^2 = \sigma_0^2$, est une loi de Khi-deux ;

ii) elle est sensible aux hypothèses :

$$K^2 = \frac{(n-1)S_c^2}{\sigma^2} \times \frac{\sigma^2}{\sigma_0^2}$$

dont le premier facteur positif, étant exactement une v.a. de loi de khi-deux, a un comportement neutre ; dont le second facteur a un comportement très différent selon que $\mathbf{H}_0$ est vraie ou fausse. En effet

$$\left(\frac{\sigma^2}{\sigma_0^2}\right) \begin{cases} < 1 & \text{si } \sigma^2 < \sigma_0^2 \text{ (sous } \mathbf{H}_1); \\ > 1 & \text{si } \sigma^2 > \sigma_0^2 \text{ (sous } \mathbf{H}_1); \\ = 1 & \text{si } \sigma^2 = \sigma_0^2 \text{ (sous } \mathbf{H}_0). \end{cases}$$

Autrement dit,

• la statistique $K^2$ aura tendance à prendre des valeurs soit plus **faibles** , soit plus **élevéess** , lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $K^2$ prendra des valeurs **modérés** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C_1$ et $C_2$ tel que $\mathbb{P}\Big(K^2 < C_1\Big) = \mathbb{P}\Big(K^2 > C_2\Big) = \alpha/2$ où $K$ suit une loi de Khi-deux à (n-1) degrés de liberté. Ainsi

$$C_1 = \mathsf{qchisq}(\alpha/2, \mathsf{n}-1), \quad \mathtt{C}_2 = \mathsf{qchisq}(1-\alpha/2, \mathsf{n}-1).$$

5) On calcule la valeur $k^2 = K^2(x_1, \ldots, x_n)$ et

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } \ k^2 \notin [C_1, C_2], \\ \text{conserver } \mathbf{H}_0 & \text{si } \ k^2 \in [C_1, C_2] \end{cases}$$

Exemple : $n = 20, \alpha = 0.05$ alors

$C_1 = \mathtt{qchisq}(0.025, \mathtt{df} = 19) = 8.906516$ ,

$C_2 = \mathtt{qchisq}(0.975, \mathtt{df} = 19) = 32.85233.$

## § 3. Comparaison de $p$ avec une référence $p_0$ où $p$ est la probabilité du "succès" (ou la proportion théorique) : $p = \mathrm{P}(X_i = 1)$

Soit une population très grande où la proportion d'individus possédant le caractère A est égale à $p$.

Nous nous proposons de comparer la proportion théorique $p$ avec une valeur de référence $p_0$. Nous disposons pour ce faire d'un échantillon $(X_1, \ldots, X_n)$ avec $X_i \in \{0, 1\}$ et $\overline{X} = \frac{\text{nb de 1}}{n}$ représente donc la fréquence des 1 parmi n.

Quelle statistique à utiliser pour ce test ? Bien que $\overline{X}$ soit le meilleur estimateur de la proportion $p$, elle ne

peut pas être directement utilisée puisque sa loi dépend du paramètre inconnu.

On suppose que l'on dispose d'un grand échantillon ($n \geq 30$) et que « p n'est pas trop petit » (de manière que l'on ait $np_0 \geq 5$ et $n(1-p_0) \geq 5$).

On utilise :

$$Z' = \frac{\sqrt{n}\Big((\overline{X} - p_0)\Big)}{\sqrt{p_0(1-\ p_0)}}$$

qui suit, lorsque $p = p_0$, asymptotiquement une loi $\mathcal{N}(0,1)$.

Pour une meilleure approximation, on préfère plutôt la version suivante en appliquant la correction de continuité de Yates :

$$Z = \text{sign}\left(\overline{X} - p_0\right) \frac{\sqrt{n}\left(|\overline{X} - p_0| - \min\left(\frac{1}{2n}, |\overline{X} - p_0|\right)\right)}{\sqrt{p_0(1 - p_0)}}$$

$$= \begin{cases} \dfrac{\sqrt{n}\left((\overline{X} - p_0) - \frac{1}{2n}\right)}{\sqrt{p_0(1 - p_0)}} & \text{si} \quad \overline{X} - p_0 > \frac{1}{2n} \\ \dfrac{\sqrt{n}\left((\overline{X} - p_0) + \frac{1}{2n}\right)}{\sqrt{p_0(1 - p_0)}} & \text{si} \quad \overline{X} - p_0 < -\frac{1}{2n} \\ 0 & \text{si} \quad -\frac{1}{2n} \leq \overline{X} - p_0 \leq \frac{1}{2n} \end{cases}$$

$Z$ suit asymptotiquement une loi $\mathcal{N}(0,1)$, où

$$c = \min\left(\frac{1}{2n}, |\overline{X} - p_0|\right)$$

est le terme de correction de Yates.

### 3.1 test unilatéral à droite

1) On souhaite tester $\mathbf{H}_0 \ : p \leq p_0$ contre $\mathbf{H}_1 \ : p > p_0$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$Z = \frac{\sqrt{n}\left((\overline{X} - p_0) \pm \frac{1}{2n}\right)}{\sqrt{p_0(1 - \ p_0)}}.$$

(*déduite du rapport des vraisemblances maximales plus la correction de Yates* )

dont la loi de $Z$, lorsque $p = p_0$, est approximativement une loi de normale $\mathcal{N}(0, 1)$ ;

ii) elle est sensible aux hypothèses :

• la statistique $Z$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $Z$ aura tendance à prendre des valeurs **faibles** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(Z > C_q\Big) \approx \alpha$ où $Z$ suit approximativement une loi normale $\mathcal{N}(0, 1)$ ;

Ainsi $C_q =$ `qnorm(1-`$\alpha$`)`

`5) On calcule la valeur` $z = Z(x_1, \ldots, x_n)$

$$\text{on décidera donc de} \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } t > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } t \leq C_q. \end{cases}$$

## 3.2 test unilatéral à gauche

`1) On souhaite tester` $\mathbf{H}_0 : p \geq p_0$ `contre` $\mathbf{H}_1 : p < p_0$

`2) On se donne un seuil` $\alpha > 0$.

3) On utilise la statistique :

$$Z = \frac{\sqrt{n}\left((\overline{X} - p_0) \pm \frac{1}{2n}\right)}{\sqrt{p_0(1 - \ p_0)}}.$$

(*déduite du rapport des vraisemblances maximales*)

i) la loi de Z, lorsque $p = p_0$, est une loi asymptotiquement $\mathcal{N}(0, 1)$ ;

ii) elle est sensible aux hypothèses :

• la statistique $Z$ aura tendance à prendre des valeurs plus **faibles** lorsque $\mathbf{H}_0$ est fausse;

• la statistique $Z$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(Z < C_q\Big) \approx \alpha$ où $Z$ suit approximativement une loi normale $\mathcal{N}(0,1)$ ;

Ainsi $C_q =$ - qnorm(1-$\alpha$)

5) On calcule la valeur $z = Z(x_1, \ldots, x_n)$,

$$\text{on décidera donc de } \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } t < C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } t \geq C_q. \end{cases}$$

## 3.3 test bilatéral

1) On souhaite tester $\mathbf{H}_0 \ : p = p_0$ contre $\mathbf{H}_1 \ : p \neq p_0$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$Z = \frac{\sqrt{n}\left((\overline{X} - p_0) \pm \frac{1}{2n}\right)}{\sqrt{p_0(1 - \ p_0)}}.$$

(*déduite du rapport des vraisemblances maximales* )

dont la loi de $Z$ , lorsque $p = p_0$, est approximativement une loi de normale $\mathcal{N}(0, 1)$ ;

ii) elle est sensible aux hypothèses :

• la statistique $Z$ aura tendance à prendre des valeurs soit plus **élevées** soit **faibles** lorsque $\mathrm{H}_0$ est fausse;

• la statistique $Z$ aura tendance à prendre des valeurs plus **modérées** lorsque $\mathrm{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(|Z| > C_q\Big) \approx \alpha$ où $Z$ suit approximativement une loi normale $\mathcal{N}(0,1)$ ;

Ainsi $C_q =$ qnorm(1-$\alpha/2$)

5) On calcule la valeur $z = Z(x_1, \ldots, x_n)$

on décidera donc de $\begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } |z| > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } |z| \leq C_q. \end{cases}$

## Exemple 1

Selon l'INSEE, 14,2 % (noté $p_0$) des ménages en 2015 vivaient dans la pauvreté. Le maire d'une grande agglomération a affirmé que la proportion de pauvreté $p$ chez lui était bien inférieure à ce chiffre, en s'appuyant sur le résultat d'un échantillon $(x_1, \ldots, x_n)$ de taille n=1650 (ménages) avec seulement $\sum_{i=1}^{n} x_i = 220$ vivant dans la pauvreté en 2015 puisque $\overline{x} = 220/1650 \approx 0.1333 < 0.142$.

Un bureau d'étude souhaite vérifier l'affirmation du maire par un test statistique au seuil alpha = 0,01.

*Pour pouvoir appuyer l'affirmation du maire, avec un risque majoré, le bureau doit effectuer :*

## § 4. Comparaison de $p_1$ et $p_2$, deux proportions théoriques de deux échantillons indépendants

Il y a de nombreux cas (comme échéances électorales, efficacité d'un nouveaux médicament etc) où nous devons décider si l'écart observé entre deux proportions d'échantillon est significatif ou non.

Nous disposons pour ce faire deux échantillons indépendants $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$ avec $X_i \in \{0, 1\}$, telles que $p_1 = \mathbf{P}(X_i = 1)$ , et $Y_j \in \{0, 1\}$ avec $p_2 = \mathbf{P}(Y_j = 1)$. Comme les variables $X_i$ et $Y_j$ ne prennent que des valeurs 1 ou 0, nous avons affaire aux deux lois de Bernoulli (donc non normales).

Puisque $X_i = X_i^2$ et $Y_j = Y_j^2$, il est évident que Les moyennes d'échantillon $\overline{X}$ et $\overline{Y}$ représentent la fréquence

de 1 de deux échantillons et que les deux variances d'échantillons sont donnés par

$$\begin{aligned} S^2_{x,c} &= \frac{1}{(n_1-1)} \sum_{i=1}^{n_1} (X_i - \overline{X})^2 \\ &= \frac{1}{(n_1-1)} \Big( \sum_{i=1}^{n_1} X_i^2 - n_1 \overline{X}^2 \Big) \\ &= \frac{1}{(n_1-1)} \Big( n_1 \overline{X} - n_1 \overline{X}^2 \Big) \\ &= \frac{n_1}{(n_1-1)} \overline{X}(1 - \overline{X} \approx \overline{X}(1 - \overline{X}.) \end{aligned}$$

Nous nous proposons de comparer deux proportions théoriques inconnues $p_1$ et $p_2$, Nous disposons pour ce faire de deux échantillons **indépendants** de taille $n_1$ pour $p_1$ que l'on estime par $\overline{X}$ et de taille $n_2$ pour $p_2$ que l'on estime par $\overline{Y}$.

Sous $\mathbf{H}_0$, $p_1 = p_2 = p$, on a

$$\begin{aligned}
\overline{X} &\approx \mathcal{N}(p, \tfrac{p(1-p)}{n_1}) \\
\overline{Y} &\approx \mathcal{N}(p, \tfrac{p(1-p)}{n_2}) \\
\overline{X} - \overline{Y} &\approx \mathcal{N}(0, \tfrac{p(1-p)}{n_1} + \tfrac{p(1-p)}{n_2}) \\
\\
\frac{\overline{X} - \overline{Y}}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} &\approx \mathcal{N}(0,1)
\end{aligned}$$

En remplaçant le paramètre $p$ par son meilleur estimateur $\widehat{p} = \dfrac{n_1\overline{X} + n_2\overline{Y}}{n_1 + n_2}$ dans le dénominateur de la dernière expression, on obtient notre statistique du test suivant :

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Lorsque les tailles de deux échantillons $n_1, n_2$ sont assez grands telles que $n_1, n_2 \geq 30$ et que $n_1\overline{X} \geq 10$, $n(1 - \overline{X}) \geq 10$, $n_1\overline{Y} \geq 10$, $n(1 - \overline{Y}) \geq 10$

On a une loi asymptotique lorsque $p_1 = p_2$ :

$$\mathcal{L}(Z) \approx \mathcal{N}(0, 1).$$

Ainsi,

## 4.1 test unilatéral à droite

1) On souhaite tester $\mathbf{H}_0 : p_1 \leq p_2$ contre $\mathbf{H}_1 : p_1 > p_2$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$Z = \frac{(\overline{X} - \overline{Y})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(*déduite du rapport des vraisemblances maximales* )

dont la loi, lorsque $p_1 = p_2$, est approximativement une loi de normale $\mathcal{N}(0, 1)$ ;

ii) elle est sensible aux hypothèses :

• la statistique $Z$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est fausse ;

• la statistique $Z$ aura tendance à prendre des valeurs **faibles** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(Z > C_q\Big) \approx \alpha$ où $Z$ suit approximativement une loi normale $\mathcal{N}(0,1)$ ;

Ainsi $C_q =$ `qnorm(1-`$\alpha$`)`

5) On calcule la valeur $z = Z(x_1, \ldots, x_n)$

on décidera donc de $\begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } t > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } t \leq C_q. \end{cases}$

## 4.2 test unilatéral à gauche

1) On souhaite tester $\mathbf{H}_0 : p_1 \geq p_2$ contre $\mathbf{H}_1 : p_1 < p_2$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$Z = \frac{(\overline{X} - \overline{Y})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(*déduite du rapport des vraisemblances maximales*)

i) la loi de Z, lorsque $p_1 = p_2$, est une loi asymptotiquement $\mathcal{N}(0,1)$ ;

ii) elle est sensible aux hypothèses :

• la statistique $Z$ aura tendance à prendre des valeurs plus **faibles** lorsque $\mathbf{H}_0$ est fausse;

• la statistique $Z$ aura tendance à prendre des valeurs plus **élevées** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(Z < C_q\Big) \approx \alpha$ où $Z$ suit approximativement une loi normale $\mathcal{N}(0,1)$ ;

Ainsi $C_q =$ - qnorm(1-$\alpha$)

5) On calcule la valeur $z = Z(x_1, \ldots, x_n)$,

$$\text{on décidera donc de} \begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } t < C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } t \geq C_q. \end{cases}$$

## 4.3 test bilatéral

1) On souhaite tester $\mathbf{H}_0 : p_1 = p_2$ contre $\mathbf{H}_1 : p_1 \neq p_2$

2) On se donne un seuil $\alpha > 0$.

3) On utilise la statistique :

$$Z = \frac{(\overline{X} - \overline{Y})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(*déduite du rapport des vraisemblances maximales* )

i) dont la loi, lorsque $p_1 = p_2$, est approximativement une loi de normale $\mathcal{N}(0,1)$ ;

ii) elle est sensible aux hypothèses :

• la statistique $Z$ aura tendance à prendre des valeurs soit plus **élevées** soit **faibles** lorsque $\mathrm{H}_0$ est fausse;

• la statistique $Z$ aura tendance à prendre des valeurs plus **modérées** lorsque $\mathbf{H}_0$ est vraie.

4) Par conséquent, on va déterminer une constante $C = C_q > 0$ tel que $\mathbb{P}\Big(|Z| > C_q\Big) \approx \alpha$ où $Z$ suit approximativement une loi normale $\mathcal{N}(0,1)$ ;

Ainsi $C_q =$ qnorm(1-$\alpha/2$)

5) On calcule la valeur $z$)

on décidera donc de $\begin{cases} \text{rejeter } \mathbf{H}_0 & \text{si } |z| > C_q, \\ \text{conserver } \mathbf{H}_0 & \text{si } |z| \leq C_q. \end{cases}$

**Exercice 2 :**

*Un sociologue effectue une étude sur les enfants vivant avec une personne adulte autre que leurs parents biologiques. Il souhaite notamment comparer les deux proportions $p_1$ et $p_2$ de deux régions pour voir s'il y a une différence significative au seuil de $\alpha = 0.05$. Pour ce faire, il prélève un échantillon dans chacune de deux régions, constate que*

| l'échantillon | la taille | dont le nombre enfants vivant avec une autre personne adulte |
|---|---|---|
| $(X_1, \ldots, X_{n_1})$ | $n_1 = 5759$ | 244 |
| $(Y_1, \ldots, Y_{n_2})$ | $n_2 = 6839$ | 243 |

Quel test le sociologue doit-il effectuer ?

## § 5. Comparaison des deux espérances de deux échantillons indépendants gaussiens

Soient $(x_1, \cdots, x_{n_1})$ et $(y_1, \cdots, y_{n_2})$ réalisations de deux échantillons indépendants gaussiens

$$\left(X_1, \cdots, X_{n_1}\right) \text{ et } \left(Y_1, \cdots, Y_{n_2}\right)$$

avec $\mu_1$ l'espérance (moyenne théorique) de $X_i$ et $\mu_2$ celle de $Y_j$. Nous souhaitons tester **Hypothèses** :

$$H_0 \ : \ \mu_1 = \mu_2 \quad \text{contre} \quad H_1 \ : \ \mu_1 \neq \mu_2.$$

Avant d'appliquer ce test (**principal**), il est conseillé de commencer par trois tests préliminaires :

• Deux tests de la normalités sur les deux populations :

Si aucun des deux tests de la normalité n'est significatif, alors nous pouvons nous contenter de supposer que les deux échantillons sont issues respectivement de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$, les quatre paramètres étant inconnus.

• un test de comparaison des variances.

Intuitivement, cela se justifie : il n'est pas équivalent de comparer les moyennes de deux populations ayant à peu près la même dispersion, et les moyennes de deux

populations ayant des dispersions très différentes. Ainsi, on suivra deux procédures différentes selon le résultat du test de comparaison des variances.

Nous procédons donc un troisième test préliminaire (le test sur l'égalité de deux variances théoriques) :

$$\tilde{H_0} \ : \ \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad \tilde{H_1} \ : \ \sigma_1^2 \neq \sigma_2^2$$

Nous utilisons la statistique de Fisher suivante :

$$F = \frac{S_{x,c}^2}{S_{y,c}^2}$$

où $S_{x,c}^2 = \sum_{i=1}^{n_1} (X_i - \overline{X})^2/(n_1-1)$, $S_{y,c}^2 = \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2/(n_2-1)$.

Nous savons que la statistique $F$ suit, lorsque $H_0$ est vraie, une loi de Fisher à $(n_1 - 1, n_2 - 1)$ degrés de liberté :

$$\mathcal{L}_{H_0}(F) = \mathcal{F}_{(n_1-1,n_2-1)}.$$

Nous calculons alors les deux valeurs critiques suivantes :

$$C_1 = qf\Big(\frac{\alpha}{2}, n_1-1, n_2-1\Big), \quad C_2 = qf\Big(1-\frac{\alpha}{2}, n_1-1, n_2-1\Big)$$

et nous décidons au seuil fixé $\alpha$ de

$$\begin{cases} \text{conserver} & H_0 & \text{si } f \in [C_1, C_2] \\ \text{rejeter} & H_0 & \text{si } f \notin [C_1, C_2]. \end{cases}$$

Deux cas sont possibles :

**Si le test de l'égalité de deux variances théoriques n'est pas significatif**

Nous supposons alors que $\sigma_1^2 = \sigma_2^2$, c'est-à-dire qu'il n'y a pas de différence significative entre $\sigma_1^2$ et $\sigma_2^2$. Nous précédons comme suit :

Pour tester **Hypothèses** :

$$H_0 \ : \ \mu_1 = \mu_2 \quad \text{contre} \quad H_1 \ : \ \mu_1 \neq \mu_2.$$

Nous savons que

$$\begin{aligned}
\overline{X} &\sim \mathcal{N}(\mu_1,\ \frac{\sigma^2}{n_1}),\\
\overline{Y} &\sim \mathcal{N}\Big(\mu_2, \frac{\sigma^2}{n_2}),\\
\overline{X}-\overline{Y} &\sim \mathcal{N}(\mu_1-\mu_2, \sigma^2\Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)\Big)\\
\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma^2\Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)}} &\sim \mathcal{N}(0,1) \ \text{ si } H_0 \text{ est vraie.}
\end{aligned}$$

En remplaçant la variance théorique commune (sous $\mathbf{H}_0$) par le meilleur estimateur, nous pouvons montrer que

$$\frac{\overline{X}-\overline{Y}}{\sqrt{S_c^2\Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)}} \sim \mathcal{T}_{n_1+n_2-2}$$

avec $S_c^2 = \frac{(n_1-1)S_{x,c}^2+(n_2-1)S_{y,c}^2}{n_1+n_2-2}$ si $H_0$ est vraie.

C'est-à-dire que la statistique $T$ suit, lorsque $H_0$ est vraie, une loi de Student à $n_1+n_2-2$ degrés de liberté :

$$\mathcal{L}_{H_0}(T) = \mathcal{T}_{n_1+n_2-2}.$$

Nous calculons alors la valeur critique suivante :

$$C = qt\Big(1-\frac{\alpha}{2}, n_1+n_2-2\Big)$$

et nous décidons au seuil fixé $\alpha$ de

$$\begin{cases} \text{conserver} & H_0 & \text{si } t \in [-C, C] \\ \text{rejeter} & H_0 & \text{si } t \notin [-C, C]. \end{cases}$$

**Si le test de l'égalité de deux variances théoriques est significatif**

Nous pensons alors que $\sigma_1$ et $\sigma_2$ sont arbitraires. Nous précédons comme suit :

Pour tester **Hypothèses** :

$$H_0 \ : \ \mu_1 = \mu_2 \quad \text{contre} \quad H_1 \ : \ \mu_1 \neq \mu_2,$$

Nous savons que

$$\begin{aligned}
\overline{X} \qquad & \sim \mathcal{N}(\mu_1, \ \frac{\sigma_1^2}{n_1}), \\
\overline{Y} \qquad & \sim \mathcal{N}\Big(\mu_2, \frac{\sigma_2^2}{n_2}), \\
\overline{X} - \overline{Y} \quad & \sim \mathcal{N}(\mu_1 - \mu_2, \Big(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\Big)\Big) \\
\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad & \sim \mathcal{N}(0, 1) \ \text{ si } H_0 \text{ est vraie.}
\end{aligned}$$

En remplaçant les deux variances théoriques par leurs estimateurs, nous pouvons utiliser la statistique de **Welsh** suivante :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_{x,c}^2}{n_1} + \frac{S_{y,c}^2}{n_2}}}$$

Nous savons que la statistique $T$ suit, lorsque $H_0$ est vraie, approximativement une loi de Student à $\nu$ degrés de liberté :

$$\mathcal{L}_{H_0}(T) = \mathcal{T}_\nu.$$

avec $\nu$ le nombre entier le plus proche de

$$\frac{\left(\frac{s_{x,c}^2}{n_1} + \frac{s_{y,c}^2}{n_2}\right)^2}{\left(\frac{s_{x,c}^4}{n_1^2(n_1-1)} + \frac{s_{y,c}^4}{n_2^2(n_2-1)}\right)}$$

Nous calculons alors la valeur critique suivante :

$$C = qt\Big(1 - \frac{\alpha}{2}, \nu\Big)$$

et nous décidons au seuil fixé $\alpha$ de

$$\begin{cases} \text{conserver} & H_0 & \text{si } t \in [-C, C] \\ \text{rejeter} & H_0 & \text{si } t \notin [-C, C]. \end{cases}$$

**Remarque** . Ce test est bilatéral ; il est le plus puissant que nous pouvons mettre en œuvre. De manière analogue, nous construisons des tests unilatéraux, et la valeur critique $C = qt\Big(1 - \frac{\alpha}{2}, \nu)$ sera remplacée par $C_1 = qt\Big(1-\alpha, \nu)$ $(C_2 = qt\Big(\alpha, \nu) = -qt\Big(1-\alpha, \nu)$ respectivement) et nous rejetons $H_0$ si $t \notin ]-\infty, C_1]$ ( $t \notin [C_2, \infty[$ respectivement).

**Comparaison de deux échantillons indépendants de grandes tailles** Si un des deux tests de la normalité est significatif, et si les tailles $n_1 > 30$ et $n_2 > 30$, alors nous pouvons procéder comme si nous avions deux échantillons indépendants et normaux. Bien que les lois de $X_i$ et $Y_j$ ne soient pas nécessairement normales, mais lorsque les tailles sont grandes, les lois de $\sqrt{n_1}(\overline{X} - \mu_1)$ et $\sqrt{n_2}(\overline{Y} - \mu_2)$ sont approximativement normales grâce au Théorème Central Limite.

Nous utilisons la statistique suivante :

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_{x,c}^2}{n_1} + \frac{S_{y,c}^2}{n_2}}}$$

Nous savons que la statistique $Z$ suit, lorsque $H_0$ est vraie, **approximativement** une loi normale centrée et réduite :

$$\mathcal{L}_{H_0}(Z) \approx \mathcal{N}(0,1).$$

*Remarque* : Nous pouvons constater que nous avons pris la même statistique $T$ utilisée dans le cas où les deux échantillons indépendants sont normaux mais sans la contrainte sur l'égalité de deux variance théoriques. Nous calculons alors la valeur critique suivante :

$$C = qnorm\Big(1 - \frac{\alpha}{2}\Big)$$

et nous décidons au seuil fixé $\alpha$ de

$$\begin{cases} \text{conserver} & H_0 & \text{si } z \in [-C, C] \\ \text{rejeter} & H_0 & \text{si } z \notin [-C, C]. \end{cases}$$