Introduction aux tests statistiques avec R

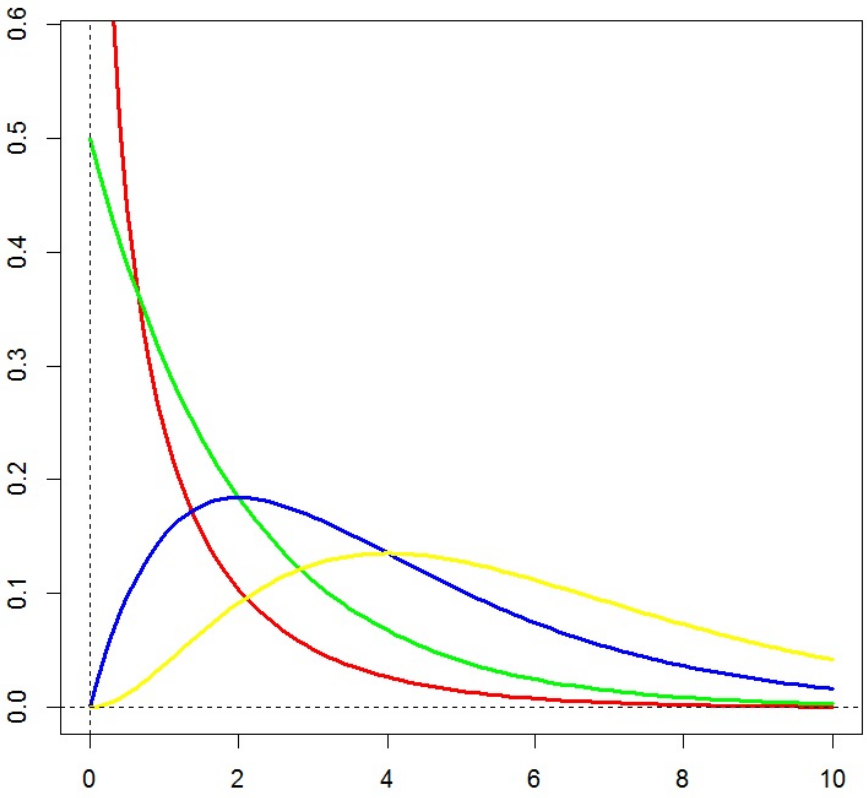
Christophe Chesneau

Université de Caen

Introduction aux tests statistiques avec R

Christophe Chesneau

http://www.math.unicaen.fr/~chesneau/



Caen, le 14 Septembre 2016

Table des matières

~ Note ~

L'objectif de ce document est de présenter les principaux tests statistiques et commandes R utilisés dans la pratique. Ce document complète certains points du livre :

`http://www.editions-ellipses.fr/product_info.php?products_id=10674`

La principale quantité utilisée sera la "p-valeur".

Contact : `christophe.chesneau@gmail.com`

Bonne lecture !

1 Notions de base

Définition

Population et individus. Une population est un ensemble fini d'objets sur lesquels une étude se porte. Ces objets sont appelés individus.

Caractère (ou variable). Un caractère est une qualité que l'on étudie chez des individus.

Échantillon. Un échantillon est un ensemble d'individus issus d'une population.

Données. Les données en notre possession sont les observations d'un caractère sur les individus d'un échantillon.

Estimation paramétrique. L'enjeu de l'estimation paramétrique est d'évaluer/estimer avec précision un paramètre inconnu émanant d'un caractère à partir des données.

Moyenne et écart-type corrigé. La moyenne et l'écart-type corrigé des données sont les principales mesures statistiques intervenant en estimation paramétrique.

En notant X un caractère numérique, n le nombre d'individus d'un échantillon et $x_1, \ldots, x_n$ les données associées, on définit :

- La moyenne de $x_1, \ldots, x_n$:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

  C'est une estimation ponctuelle de la valeur moyenne de X.

- L'écart-type corrigé de $x_1, \ldots, x_n$:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

  C'est une estimation ponctuelle de la variabilité de X autour de sa moyenne. La valeur obtenue a la même unité que X.

Exemple

Contexte :

<table>
<tr><td>Population</td><td>Ensemble des pommes d'une ferme</td></tr>
<tr><td>Individu</td><td>Pomme</td></tr>
<tr><td>Caractère</td><td>Poids d'une pomme (en grammes)</td></tr>
<tr><td>Paramètre inconnu</td><td>Poids moyen d'une pomme</td></tr>
<tr><td>Échantillon</td><td>7 pommes choisies au hasard ($n = 7$)</td></tr>
<tr><td>Données</td><td>
<table>
<tr><td>x_1</td><td>x_2</td><td>x_3</td><td>x_4</td><td>x_5</td><td>x_6</td><td>x_7</td></tr>
<tr><td>162</td><td>155</td><td>148</td><td>171</td><td>151</td><td>165</td><td>154</td></tr>
</table>
(par exemple, x_1 est le poids de la première pomme de l'échantillon, soit 162 grammes)
</td></tr>
<tr><td>Objectif</td><td>Évaluer le poids moyen inconnu d'une pomme à l'aide des données $x_1, \ldots, x_7$</td></tr>
</table>

Mesures statistiques :

Moyenne	$\overline{x} = \frac{1}{7}\sum_{i=1}^{7} x_i = 158$
Écart-type corrigé	$s = \sqrt{\frac{1}{7-1}\sum_{i=1}^{7}(x_i - \overline{x})^2} = 8.246211$

Modélisation

Loi normale. Si le caractère X représente une grandeur sujette à une somme d'erreurs mineures indépendantes, on le modélise comme une *var* $X \sim \mathcal{N}(\mu, \sigma^2)$.

Par exemple, X peut être : poids, taille, temps, distance, masse, vitesse, température, indice, score, salaire, note, quantité ou teneur. En outre, la taille en centimètres d'un homme est une *var* X suivant la loi normale $\mathcal{N}(175, 6^2)$ *(le "est" est un abus de langage ; la var X est l'application qui, à chaque homme choisi au hasard dans la population, associe sa taille exprimée en centimètres. Il est plus précis de dire : la taille en centimètres d'un homme peut être modélisée par une var X suivant la loi normale $\mathcal{N}(175, 6^2)$).* Dans ce cas, μ est la moyenne de X et σ^2 mesure la variabilité de X autour de μ.

Loi de Bernoulli. Si X prend deux valeurs : 0 ou 1, correspondant souvent à un codage binaire, on le modélise comme une *var* $X \sim \mathcal{B}(p)$.

Par exemple, $X = 1$ peut caractériser :

- le succès à une épreuve,
- la présence d'un élément caractéristique.

Le paramètre p est la probabilité que $X = 1$ se réalise, laquelle peut aussi s'interpréter en terme de proportion d'individus dans la population vérifiant $X = 1$.

Exemple.

Population	Ensemble des fromages d'une laiterie
Individu	Fromage
Caractère 1	X = Poids d'un fromage (en grammes)
Modélisation	$X \sim \mathcal{N}(\mu, \sigma^2)$
Paramètres : μ et σ^2	μ = Poids moyen d'un fromage σ^2 mesure la dispersion du poids d'un fromage autour de μ
Caractère 2	$Y = 1$ si le fromage présente un défaut de conditionnement et $Y = 0$ sinon
Modélisation	$Y \sim \mathcal{B}(p)$
Paramètre p	p = Proportion de fromages ayant un défaut de conditionnement

2 Bases des tests statistiques

Hypothèses. On oppose deux hypothèses complémentaires : H_0 et H_1,

- l'hypothèse H_0 formule ce que l'on souhaite rejeter/réfuter,
- l'hypothèse H_1 formule ce que l'on souhaite montrer.

Par exemple, si on veut montrer l'hypothèse "lot non conforme", H_0 et H_1 s'opposent sous la forme :

$$H_0 : \text{"lot conforme"} \qquad \text{contre} \qquad H_1 : \text{"lot non conforme"}.$$

Risque. Le risque est le pourcentage de chances de rejeter H_0, donc d'accepter H_1, alors que H_0 est vraie. On veut que ce risque soit aussi faible que possible.

Il s'écrit sous la forme : $100\alpha\%$, avec $\alpha \in]0,1[$ (par exemple, 5%, soit $\alpha = 0.05$).

Le réel α est alors la probabilité de rejeter H_0 alors que H_0 est vraie.

Le rejet de H_0 est dit "significatif" si elle est rejetée au risque 5%.

Test statistique. Un test statistique est une procédure qui vise à apporter une réponse à la question :

Est-ce que les données nous permettent de rejeter H_0, donc d'accepter H_1, avec un faible risque de se tromper ?

Types de test statistique. En notant θ un paramètre inconnu, on dit que le test est

- bilatéral si H_1 est de la forme $H_1 : \theta \neq \dots$
- unilatéral à gauche (sens de $<$) si H_1 est de la forme $H_1 : \theta < \dots$
- unilatéral à droite (sens de $>$) si H_1 est de la forme $H_1 : \theta > \dots$

p-valeur. La p-valeur est le plus petit réel $\alpha \in]0,1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H_0 au risque $100\alpha\%$. Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H_0 alors que H_0 est vraie.

Les logiciels actuels travaillent principalement avec cette p-valeur.

Degré de significativité. La p-valeur nous donne un degré de significativité du rejet de H_0.

Le rejet de H_0 sera :

- significatif si p-valeur $\in]0.01,\ 0.05]$, symbolisé par $\star$,
- très significatif si p-valeur $\in]0.001,\ 0.01]$, symbolisé par $\star\star$,

- hautement significatif si p-valeur < 0.001, symbolisé par $\star\star\star$.

Il y a non rejet de H_0 si p-valeur > 0.05.

Non-rejet. S'il y a non-rejet de H_0, sauf convention, on ne peut rien conclure du tout (avec le risque considéré). En revanche, peut-être qu'un risque de départ plus élevé ou la disposition de plus de données peuvent conduire à un rejet de H_0.

3 Tests de conformité à une valeur de référence

Enjeu

L'enjeu d'un test de conformité est d'affirmer, avec un faible risque de se tromper, qu'une norme associée à un caractère X (sa moyenne, une proportion...) n'est plus conforme à la réalité.

Ainsi, en posant H_1 : "la norme n'est plus conforme", on se pose la question : Est-ce que les données $x_1, \dots, x_n$, observations de X, nous permettent de rejeter H_0, donc d'accepter H_1, avec un faible risque de se tromper ?

Formules : p-valeurs

Lois : $Z \sim \mathcal{N}(0,1)$, $T \sim \mathcal{T}(\nu)$ et $K \sim \chi^2(\nu)$, $\nu = n-1$. Outils : $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2}$.

$X \sim \mathcal{N}(\mu, \sigma^2)$	H_1	Stat. test obs.	p-valeurs
σ connu : Z-Test	$\mu \neq \mu_0$ $\mu > \mu_0$ $\mu < \mu_0$	$z_{obs} = \sqrt{n}\left(\frac{\overline{x} - \mu_0}{\sigma}\right)$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
σ inconnu : T-Test	$\mu \neq \mu_0$ $\mu > \mu_0$ $\mu < \mu_0$	$t_{obs} = \sqrt{n}\left(\frac{\overline{x} - \mu_0}{s}\right)$	$\mathbb{P}(\lvert T\rvert \geq \lvert t_{obs}\rvert)$ $\mathbb{P}(T \geq t_{obs})$ $\mathbb{P}(T \leq t_{obs})$
1-Chi2-Test	$\sigma^2 \neq \sigma_0^2$ $\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$	$\chi^2_{obs} = \frac{n-1}{\sigma_0^2}s^2$	$2\min\left(\mathbb{P}(K \geq \chi^2_{obs}),\ \mathbb{P}(K \leq \chi^2_{obs})\right)$ $\mathbb{P}(K \geq \chi^2_{obs})$ $\mathbb{P}(K \leq \chi^2_{obs})$
$X \sim \mathcal{B}(p)$	H_1	Stat. test obs. et *var*	p-valeurs
$n \geq 31$, $np_0 \geq 5$, $n(1-p_0) \geq 5$: 1-Prop-Z-Test cor	$p \neq p_0$ $p > p_0$ $p < p_0$	$z_{obs} = \sqrt{n}\left(\frac{\lvert\overline{x} - p_0\rvert - \frac{0.5}{n}}{\sqrt{p_0(1-p_0)}}\right)$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
$n \geq 31$, $np_0 \geq 5$, $n(1-p_0) \geq 5$: 1-Prop-Z-Test	$p \neq p_0$ $p > p_0$ $p < p_0$	$z_{obs} = \sqrt{n}\left(\frac{\overline{x} - p_0}{\sqrt{p_0(1-p_0)}}\right)$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$

Commandes

Pour les commandes ci-dessous et à venir, on considère les libraries `stats` et `OneTwoSamples` :

```
library(stats)

library(OneTwoSamples)
```

On propose les commandes R suivantes :

$X \sim \mathcal{N}(\mu, \sigma^2)$	H_1	Commandes
σ connu : Z-Test	$\mu \neq \mu_0$	`mean_test1(x, mu0, sigma)$p_value`
	$\mu > \mu_0$	`mean_test1(x, mu0, sigma, side = 1)$p_value`
	$\mu < \mu_0$	`mean_test1(x, mu0, sigma, side = -1)$p_value`
σ inconnu : T-Test	$\mu \neq \mu_0$	`t.test(x, mu = mu0)$p.value`
	$\mu > \mu_0$	`t.test(x, mu = mu0, alternative = "greater")$p.value`
	$\mu < \mu_0$	`t.test(x, mu = mu0, alternative = "less")$p.value`
1-Chi2-Test	$\sigma^2 \neq \sigma_0^2$	`var_test1(x, sigma20)$P_value`
	$\sigma^2 > \sigma_0^2$	`var_test1(x, sigma20, side = 1)$P_value`
	$\sigma^2 < \sigma_0^2$	`var_test1(x, sigma20, side = -1)$P_value`
$X \sim \mathcal{B}(p)$	H_1	Commandes
$n \geq 31,\ np_0 \geq 5,$ $n(1-p_0) \geq 5$: 1-Prop-Z-Test cor	$p \neq p_0$	`prop.test(x, n, p)$p.value`
	$p > p_0$	`prop.test(x, n, p, alt = "greater")$p.value`
	$p < p_0$	`prop.test(x, n, p, alternative = "less")$p.value`
$n \geq 31,\ np_0 \geq 5,$ $n(1-p_0) \geq 5$: 1-Prop-Z-Test	$p \neq p_0$	`prop.test(x, n, p, correct = F)$p.value`
	$p > p_0$	`prop.test(x, n, p, alternative = "greater", correct = F)$p.value`
	$p < p_0$	`prop.test(x, n, p, alternative = "less", correct = F)$p.value`

Remarque : En omettant les commandes `$p.value` (ou `$p_value`), les commandes renvoient plus d'éléments associés au test statistique considéré, dont la p-valeur (statistique de test observée, degré de liberté, intervalle de confiance...).

Exemples

Exemple 1. Une entreprise utilise une matière isolante pour fabriquer des appareils de contrôle industriel. Elle achète des composants isolants à un certain fournisseur qui certifie que l'épaisseur moyenne de ses composants est de 7.3 millimètres. Pour voir si le fournisseur respecte ses engagements, l'entreprise mesure l'épaisseur de 24 composants pris au hasard dans la livraison. Les résultats, en millimètres, sont :

6.47	7.02	7.15	7.22	7.44	6.99	7.47	7.61	7.32	7.22	7.52	6.92

7.28	6.69	7.24	7.19	6.97	7.52	6.22	7.13	7.32	7.67	7.24	6.21

On suppose que l'épaisseur en millimètres d'un de ces composants peut être modélisée par une *var* $X \sim \mathcal{N}(\mu, (0.38)^2)$, avec μ inconnu.

Peut-on affirmer, avec un faible risque de se tromper, que le fournisseur ne respecte pas ses engagements ?

Solution 1. Par l'énoncé, on observe la valeur de $X \sim \mathcal{N}(\mu, \sigma^2)$ pour chacun des n individus (composants) d'un échantillon avec $n = 24$, μ inconnu et $\sigma = 0.38$. On veut affirmer, avec un faible risque de se tromper, que le fournisseur ne respecte pas ses engagements. Cela est le cas si l'épaisseur moyenne de ses composants est différente de 7.3 millimètres, soit $\mu \neq 7.3$. Par conséquent, l'hypothèse H_1 est : $H_1 : \mu \neq 7.3$. On considère alors les hypothèses :

$$H_0 : \mu = 7.3 \qquad \text{contre} \qquad H_1 : \mu \neq 7.3.$$

Comme σ est connu, on utilise un Z-Test. Il est bilatéral.

On considère les commandes :

```
library(OneTwoSamples)
x = c(6.47, 7.02, 7.15, 7.22, 7.44, 6.99, 7.47, 7.61, 7.32, 7.22, 7.52,
6.92, 7.28, 6.69, 7.24, 7.19, 6.97, 7.52, 6.22, 7.13, 7.32, 7.67, 7.24,
6.21)
mean_test1(x, 7.3, 0.38)$p_value
```

Cela renvoie : `[1] 0.02509132`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, on peut affirmer que le fournisseur ne respecte pas ses engagements. En affirmant cela, il y a un peu moins de 2.6 chances sur 100 de se tromper.

Exemple 2. Une usine fabrique un certain type de récipient en plastique. On cherche à montrer, avec un faible risque de se tromper, que le contenu moyen d'un récipient est strictement supérieur à 10 litres. Le contenu de 12 récipients choisis au hasard dans la production est mesuré. Les résultats, en litres, sont :

10.1	9.8	10.2	10.3	10.4	9.8	9.9	10.4	10.2	9.5	10.4	9.6

On suppose que le contenu en litres d'un récipient de cet usine peut être modélisé par une *var* X suivant une loi normale.

Proposer un test statistique adapté et conclure.

Solution 2. Par l'énoncé, on observe la valeur de $X \sim \mathcal{N}(\mu, \sigma^2)$ pour chacun des n individus (récipients) d'un échantillon avec $n = 12$, et μ et σ inconnus. On veut montrer, avec un faible risque de se tromper, que le contenu moyen d'un récipient est strictement supérieur à 10 litres, soit $\mu > 10$. Par conséquent, l'hypothèse H_1 est : $H_1 : \mu > 10$.

On considère alors les hypothèses :

$$H_0 : \mu \leq 10 \qquad \text{contre} \qquad H_1 : \mu > 10.$$

Comme σ est inconnu, on utilise un T-Test. Il est unilatéral à droite.

On considère les commandes :

```
x = c(10.1, 9.8, 10.2, 10.3, 10.4, 9.8, 9.9, 10.4, 10.2, 9.5, 10.4, 9.6)
t.test(x, mu = 10, alternative = "greater")$p.value
```

Cela renvoie : `[1] 0.299845`

Comme p-valeur > 0.05, on ne rejette pas H_0. Les données ne nous permettent pas d'affirmer que le contenu moyen des récipients de cette usine est strictement supérieur à 10 litres.

Exemple 3. Dans une production, pour que le poids annoncé du contenu d'une boîte de conserve de tomates soit conforme, il faut régler la moyenne du conditionnement à 276 grammes.

Une panne est survenue dans la conditionneuse et le producteur craint que le réglage ne soit plus fiable. Il se pose la question : le réglage est-il encore à 276 grammes ? Il prélève 8 boîtes au hasard dans la production et les pèse une à une. Les résultats, en grammes, sont :

232	277	235	245	245	250	268	256

On suppose que le poids en grammes du contenu d'une boîte de conserve de tomates de cette production peut être modélisé par une *var* X suivant une loi normale.

Faire un test statistique pour répondre à la question du producteur.

Solution 3. Par l'énoncé, on observe la valeur de $X \sim \mathcal{N}(\mu, \sigma^2)$ pour chacun des n individus (boîtes de conserve de tomates) d'un échantillon avec $n = 8$, et μ et σ inconnus.

On considère les hypothèses :

$$H_0 : \mu = 276 \qquad \text{contre} \qquad H_1 : \mu \neq 276.$$

On utilise un T-Test. Il est bilatéral.

On fait :

```
x = c(232, 277, 235, 245, 245, 250, 268, 256)

t.test(x, mu = 276)$p.value
```

Cela renvoie : `[1] 0.00259146`

Comme p-valeur $\in]0.001, 0.01]$, le rejet de H_0 est très significatif $\star\star$.

Par conséquent, au risque au moins de 1%, on peut dire que le réglage de la conditionneuse n'est plus à 276 grammes.

Exemple 4. Un producteur affirme qu'exactement 25% des haricots verts de sa récolte sont extra-fins. Sur 400 haricots verts choisis au hasard dans la récolte, on en compte 118 extra-fins.

Est-ce que l'on peut affirmer, au risque 5%, que le producteur a tort ?

Solution 4. Soient p la proportion inconnue des haricots verts extra-fins dans la récolte et X la *var* qui vaut 1 si le haricot vert est extra-fin et 0 sinon ; $X \sim \mathcal{B}(p)$. Par l'énoncé, on observe la valeur de X pour chacun des n individus (haricots verts) d'un échantillon avec $n = 400$.

On considère les hypothèses :

$$H_0 : p = 0.25 \qquad \text{contre} \qquad H_1 : p \neq 0.25.$$

On utilise un 1-Prop-Z-Test cor. Il est bilatéral.

On considère les commandes :

```
prop.test(118, 400, 0.25)$p.value
```

Cela renvoie : `[1] 0.04330814`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur < 0.05, on peut affirmer, au risque 5%, que le producteur a tort.

On aurait aussi pu utiliser la version classique, sans correction de Yates :

```
prop.test(118, 400, 0.25, correct = F)$p.value
```

Cela renvoie : `[1] 0.03766692`

On aboutit à la même conclusion.

Remarque : Le 1-Prop-Z-Test avec la correction de Yates est plus fiable que sans la correction. Toutefois, il repose sur des résultats théoriques asymptotiques (convergence en loi). Pour mettre en œuvre un test utilisant la loi exacte (binomiale), on utilise les commandes :

```
binom.test(118, 400, 0.25)$p.value
```

(Cela renvoie : `[1] 0.04308655`)

Le résultat peut être différent. Par exemple, comparer les commandes :

```
prop.test(3, 5, 0.18)$p.value

binom.test(3, 5, 0.18)$p.value
```

Dans le premier, apparaît un "Warning message" signifiant que l'approximation normale n'est sans doute pas valide.

4 Tests d'homogénéité : échantillons indépendants

Contexte

On étudie un caractère dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On cherche à comparer $\mathcal{P}_1$ et $\mathcal{P}_2$ quant à ce caractère, et donc à analyser leur éventuelle homogénéité.

Pour ce faire, on considère

- un échantillon E_1 de n_1 individus de $\mathcal{P}_1$,
- un échantillon E_2 de n_2 individus de $\mathcal{P}_2$.

Échantillons indépendants

Si tous les individus sont différents, les échantillons E_1 et E_2 sont indépendants.

Données

On étudie un caractère représenté par une *var* X.

- La *var* X considérée dans $\mathcal{P}_1$ est une *var* X_1.
- La *var* X considérée dans $\mathcal{P}_2$ est une *var* X_2.

Les données sont constituées de

- la valeur de X_1 pour chacun des n_1 individus de E_1 : $x_{1,1}, \ldots, x_{1,n_1}$,
- la valeur de X_2 pour chacun des n_2 individus de E_2 : $x_{2,1}, \ldots, x_{2,n_2}$.

On suppose que les individus sont tous différents ; E_1 et E_2 sont indépendants.

On peut mettre les données sous la forme :

- pour E_1 :

$x_{1,1}$	$x_{1,2}$	$\ldots$	x_{1,n_1}

- pour E_2 :

$x_{2,1}$	$x_{2,2}$	$\ldots$	x_{2,n_2}

Formules : p-valeurs

Lois : $Z \sim \mathcal{N}(0,1)$, $F \sim \mathcal{F}(\nu_1,\nu_2)$, $(\nu_1,\nu_2) = \begin{cases} (n_1-1, n_2-1) & \text{si } s_1 > s_2, \\ (n_2-1, n_1-1) & \text{si } s_2 > s_1 \end{cases}$, $T_\nu \sim \mathcal{T}(\nu)$, $\nu = n_1+n_2-2$, $T_\gamma \sim \mathcal{T}(\gamma)$, $\gamma = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_1^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$.

Outils : $\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1,i}$, $\overline{x}_1 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2,i}$, $\overline{x}_p = \frac{n_1\overline{x}_1+n_2\overline{x}_2}{n_1+n_2}$, $s_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1,i}-\overline{x}_1)^2}$, $s_2 = \sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2,i}-\overline{x}_2)^2}$, $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

$X_1 \sim \mathcal{N}(\mu_1,\sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2,\sigma_2^2)$	H_1	Stat. test obs.	p-valeurs
σ_1, σ_2 connus : 2-Comp-Z-Test	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$z_{obs} = \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
σ_1, σ_2 inconnus : 2-Comp-F-Test	$\sigma_1^2 \neq \sigma_2^2$	$f_{obs} = \left(\dfrac{\max(s_1,s_2)}{\min(s_1,s_2)}\right)^2$	$2\mathbb{P}(F \geq f_{obs})$
σ_1, σ_2 inconnus, $\sigma_1^2 = \sigma_2^2$: 2-Comp-T-Test pooled yes	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \dfrac{\overline{x}_1-\overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$\mathbb{P}(\lvert T_\nu\rvert \geq \lvert t_{obs}\rvert)$ $\mathbb{P}(T_\nu \geq t_{obs})$ $\mathbb{P}(T_\nu \leq t_{obs})$
σ_1, σ_2 inconnus, $\sigma_1^2 \neq \sigma_2^2$: 2-Comp-T-Test pooled no	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$\mathbb{P}(\lvert T_\gamma\rvert \geq \lvert t_{obs}\rvert)$ $\mathbb{P}(T_\gamma \geq t_{obs})$ $\mathbb{P}(T_\gamma \leq t_{obs})$
$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Stat. test obs.	p-valeurs
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\overline{x}_1 \geq 5$, $n_1(1-\overline{x}_1) \geq 5$, $n_2\overline{x}_2 \geq 5$, $n_2(1-\overline{x}_2) \geq 5$: 2-Prop-Z-Test cor	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z_{obs} = \dfrac{\lvert\overline{x}_1-\overline{x}_2\rvert - \left(\frac{0.5}{n_1}+\frac{0.5}{n_2}\right)}{\sqrt{\overline{x}_p(1-\overline{x}_p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\overline{x}_1 \geq 5$, $n_1(1-\overline{x}_1) \geq 5$, $n_2\overline{x}_2 \geq 5$, $n_2(1-\overline{x}_2) \geq 5$: 2-Prop-Z-Test	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z_{obs} = \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\overline{x}_p(1-\overline{x}_p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$

Commandes

On considère les libraries `stats` et `OneTwoSamples` :

```
library(stats)

library(OneTwoSamples)
```

On propose les commandes R suivantes :

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2),\ X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$	H_1	Commandes
σ_1, σ_2 connus : 2-Comp-Z-Test	$\mu_1 \neq \mu_2$	`mean_test2(x1, x2, sigma = c(sigma1, sigma2))$p_value`
	$\mu_1 > \mu_2$	`mean_test2(x1, x2, sigma = c(sigma1, sigma2), side = -1)$p_value`
	$\mu_1 < \mu_2$	`mean_test2(x1, x2, sigma = c(sigma1, sigma2), side = 1)$p_value`
σ_1, σ_2 inconnus : 2-Comp-F-Test	$\sigma_1^2 \neq \sigma_2^2$	`var.test(x1, x2)$p.value`
σ_1, σ_2 inconnus, $\sigma_1^2 = \sigma_2^2$: 2-Comp-T-Test pooled yes	$\mu_1 \neq \mu_2$	`t.test(x1, x2, var.equal = T)$p.value`
	$\mu_1 > \mu_2$	`t.test(x1, x2, alternative = "greater", var.equal = T)$p.value`
	$\mu_1 < \mu_2$	`t.test(x1, x2, alternative = "less", var.equal = T)$p.value`
σ_1, σ_2 inconnus, $\sigma_1^2 \neq \sigma_2^2$: 2-Comp-T-Test pooled no	$\mu_1 \neq \mu_2$	`t.test(x1, x2)$p.value`
	$\mu_1 > \mu_2$	`t.test(x1, x2, alternative = "greater")$p.value`
	$\mu_1 < \mu_2$	`t.test(x1, x2, alternative = "less")$p.value`
$X_1 \sim \mathcal{B}(p_1),\ X_2 \sim \mathcal{B}(p_2)$	H_1	Commandes
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\overline{x}_1 \geq 5$, $n_1(1-\overline{x}_1) \geq 5$, $n_2\overline{x}_2 \geq 5$, $n_2(1-\overline{x}_2) \geq 5$: 2-Prop-Z-Test cor	$p_1 \neq p_2$	`prop.test(x = c(x1, x2), n = c(n1, n2))$p.value`
	$p_1 > p_2$	`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater")$p.value`
	$p_1 < p_2$	`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less")$p.value`
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\overline{x}_1 \geq 5$, $n_1(1-\overline{x}_1) \geq 5$, $n_2\overline{x}_2 \geq 5$, $n_2(1-\overline{x}_2) \geq 5$: 2-Prop-Z-Test	$p_1 \neq p_2$	`prop.test(x = c(x1, x2), n = c(n1, n2), correct = F)$p.value`
	$p_1 > p_2$	`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater", correct = F)$p.value`
	$p_1 < p_2$	`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less", correct = F)$p.value`

Exemples

Exemple 1. La société de Monsieur Labrador utilise deux machines, machine 1 et machine 2, pour remplir automatiquement des paquets de cacao en poudre.

- On prélève un échantillon de 10 paquets remplis par la machine 1 et on les pèse. Les résultats, en grammes, sont :

106.70	107.02	107.15	107.22	107.41	106.39	107.47	107.61	107.38	107.22

- On prélève un échantillon de 9 paquets remplis par la machine 2 et on les pèse. Les résultats, en grammes, sont :

107.68	106.69	107.24	107.69	106.97	107.52	106.22	107.23	107.32

On suppose que le poids en grammes d'un paquet rempli par la machine 1 peut être modélisé par une $X_1 \sim \mathcal{N}(\mu_1, 1.3^2)$ et celui avec la machine 2 peut être modélisé par une *var* $X_2 \sim \mathcal{N}(\mu_2, 0.9^2)$.

Peut-on affirmer, au risque 5%, que les machines sont réglées de manière différente ?

Solution 1. Par l'énoncé, on observe

- la valeur de $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ pour chacun des n_1 individus (paquets) d'un échantillon avec $n_1 = 10$, μ_1 inconnu et $\sigma_1 = 1.3$,
- la valeur de $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ pour chacun des n_2 individus (paquets) d'un échantillon avec $n_2 = 9$, μ_2 inconnu et $\sigma_2 = 0.9$.

Les échantillons sont indépendants car les individus considérés sont tous différents.

On veut affirmer, avec un faible risque de se tromper, que les machines sont réglées de manière différente. Cela est le cas si le poids moyen d'un paquet rempli par la machine 1 diffère de celui rempli par la machine 2, soit $\mu_1 \neq \mu_2$. Par conséquent, l'hypothèse H_1 est : $H_1 : \mu_1 \neq \mu_2$.

On considère alors les hypothèses :

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

Comme σ_1 et σ_2 sont connus, on utilise un 2-Comp-Z-Test. Il est bilatéral.

On considère les commandes :

```
library(OneTwoSamples)
x1 = c(106.70, 107.02, 107.15, 107.22, 107.41, 106.39, 107.47, 107.61,
107.38, 107.22)
x2 = c(107.68, 106.69, 107.24, 107.69, 106.97, 107.52, 106.22, 107.23,
107.32)
mean_test2(x1, x2, sigma = c(1.3, 0.9))$p_value
```

Cela renvoie : `[1] 0.974397`

Comme p-valeur > 0.05, on ne rejette pas H_0. Les données ne nous permettent pas d'affirmer que les machines 1 et 2 sont réglées de manière différente.

Exemple 2. On considère deux lots de tasses et on souhaite comparer la solidité de ceux-ci. Pour chacun des deux lots, on dispose d'un échantillon de 10 tasses et on mesure la résistance de chacune d'entre eux. Les résultats sont :

- pour le premier échantillon :

31.70	31.98	32.24	32.35	31.18	32.19	32.63	31.19	31.54	31.89

- pour le deuxième échantillon :

31.61	31.10	31.20	31.11	32.66	31.15	31.71	31.22	31.16	31.21

La solidité d'une tasse du premier lot peut être modélisée par une *var* X_1, et celle du tasse du second lot peut être modélisée par une *var* X_2. On suppose que X_1 et X_2 suivent des lois normales de variances égales.

Peut-on affirmer que ces deux échantillons ne proviennent pas de la même production ?

Solution 2. Par l'énoncé, on observe

- la valeur de $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ pour chacun des n_1 individus (tasses) d'un échantillon avec $n_1 = 10$, et μ_1 et σ_1 inconnus,
- la valeur de $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ pour chacun des n_2 individus (tasses) d'un échantillon avec $n_2 = 10$, et μ_2 et σ_2 inconnus.

On a $\sigma_1^2 = \sigma_2^2$. Les individus étant tous différents, les échantillons sont indépendants.

On considère les hypothèses :

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

On utilise un 2-Comp-T-Test avec pooled yes car on a l'égalité $\sigma_1^2 = \sigma_2^2$. Il est bilatéral.

On considère les commandes :

```
x1 = c(31.70, 31.98, 32.24, 32.35, 31.18, 32.19, 32.63, 31.19, 31.54, 31.89)
x2 = c(31.61, 31.10, 31.20, 31.11, 32.66, 31.15, 31.71, 31.22, 31.16, 31.21)
t.test(x1, x2, var.equal = T)$p.value
```

Cela renvoie : `[1] 0.04214053`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, on peut affirmer que les deux échantillons de tasses proviennent de deux productions différentes. En affirmant cela, il y a un peu moins de 5 chances sur 100 de se tromper.

Exercice 3. On dispose de deux lots de boîtes de sauce italienne conditionnées de la même manière mais provenant de producteurs différents. On s'intéresse à la teneur en grammes de viande dans celles-ci.

- On extrait 7 boîtes provenant du premier producteur et on mesure leur teneur de viande. Les résultats, en grammes, sont :

12.12	12.03	13.58	13.38	11.81	15.92	13.65

- On extrait 6 boîtes provenant du deuxième producteur et on mesure leur teneur de viande. Les résultats, en grammes, sont :

14.81	13.93	14.91	15.87	15.62	15.39

La teneur en grammes de viande dans une boîte provenant du premier producteur peut être modélisée par une *var* X_1, et celle dans une boîte provenant du deuxième producteur peut être modélisée par une *var* X_2. On suppose que X_1 et X_2 suivent des lois normales.

Peut-on affirmer qu'il y a une différence entre les producteurs quant à la teneur moyenne en viande dans les boîtes ?

Solution 3. Par l'énoncé, on observe

- la valeur de $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ pour chacun des n_1 individus (boîtes de sauce italienne) d'un échantillon avec $n_1 = 6$, et μ_1 et σ_1 inconnus,
- la valeur de $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ pour chacun des n_2 individus (boîtes de sauce italienne) d'un échantillon avec $n_2 = 5$, et μ_2 et σ_2 inconnus.

Les individus étant tous différents, les échantillons sont indépendants.

On considère les hypothèses :

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

On utilise un 2-Comp-T-Test sans connaissance de l'égalité : $\sigma_1^2 = \sigma_2^2$. Il est bilatéral.

On considère les commandes :

```
x1 = c(12.12, 12.03, 13.58, 13.38, 11.81, 15.92, 13.65)
x2 = c(14.81, 13.93, 14.91, 15.87, 15.62, 15.39)
t.test(x1, x2)$p.value
```

Cela renvoie : `[1] 0.01335816`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, on peut affirmer qu'il y a une différence "significative" entre les producteurs quant à la teneur moyenne en viande dans les boîtes.

Exercice 4. Un producteur de desserts lactés au caramel se trouve en concurrence avec d'autres marques. Au début de l'année 2010, il décide d'investir dans une nouvelle présentation de ses desserts. Avant d'avoir le bilan de l'année, il fait une rapide enquête auprès d'un certain nombre de magasins.

- Avant la nouvelle présentation, sur 230 desserts vendus, 54 étaient ceux du producteur.
- Après la nouvelle présentation, sur 340 desserts vendus, 110 étaient ceux du producteur.

Est-ce que le producteur peut affirmer que la nouvelle présentation a augmenté sa part de marché sur les desserts lactés au caramel ?

Solution 4. Soient

- p_1 la proportion inconnue de desserts vendus avec l'ancienne présentation et X_1 la *var* qui vaut 1 si le dessert avec l'ancienne présentation est vendu et 0 sinon ; $X_1 \sim \mathcal{B}(p_1)$,
- p_2 la proportion inconnue de dessert vendus avec la nouvelle présentation et X_2 la *var* qui vaut 1 si le dessert avec la nouvelle présentation est vendu et 0 sinon ; $X_2 \sim \mathcal{B}(p_2)$.

Par l'énoncé, on observe

- la valeur de X_1 pour chacun des n_1 individus (desserts) d'un échantillon avec $n_1 = 230$,
- la valeur de X_2 pour chacun des n_2 individus (desserts) d'un échantillon avec $n_2 = 340$.

Les individus étant tous différents, les échantillons sont indépendants.

On considère les hypothèses :

$$H_0 : p_1 \geq p_2 \qquad \text{contre} \qquad H_1 : p_1 < p_2.$$

On utilise un 2-Prop-Z-Test cor. Il est unilatéral à gauche.

On considère les commandes :

```
prop.test(x = c(54, 110), n = c(230, 340), alternative = "less")$p.value
```

Cela renvoie : `[1] 0.01383626`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, le producteur peut affirmer que la nouvelle présentation a significativement augmenté sa part de marché sur les desserts lactés au caramel.

Remarque : Le 2-Prop-Z-Test avec la correction de Yates est plus fiable que sans la correction. Toutefois, un test statistique plus puissant existe : le test exact de Fisher. On considère les commandes :

```
A = matrix(c(54, 110, 230 - 54, 340 - 110), 2)

fisher.test(A, alternative = "less")$p.value
```

(Cela renvoie : `[1] 0.01339192`)

5 Tests d'homogénéité : échantillons appariés

Contexte

On étudie un caractère dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On cherche à comparer $\mathcal{P}_1$ et $\mathcal{P}_2$ quant à ce caractère, et donc à analyser leur éventuelle homogénéité.

Pour ce faire, on considère

- un échantillon E_1 de n_1 individus de $\mathcal{P}_1$,
- un échantillon E_2 de n_2 individus de $\mathcal{P}_2$.

Échantillons appariés

Si les individus de $\mathcal{P}_1$ sont soumis à un certain traitement (ou aucun), et ceux de $\mathcal{P}_2$ sont les individus de $\mathcal{P}_1$ soumis à un autre traitement, les échantillons E_1 et E_2 sont appariés : ce sont les mêmes individus qui sont considérés dans les deux échantillons. On compare alors les effets des deux traitements en considérant un même échantillon de $n = n_1 = n_2$ individus.

Données

On étudie un caractère représenté par une *var* X.

- La *var* X considérée dans $\mathcal{P}_1$ est une *var* X_1,
- La *var* X considérée dans $\mathcal{P}_2$ est une *var* X_2.

Les données sont constituées de

- la valeur de X_1 pour chacun des $n_1 = n$ individus de E_1 : $x_{1,1}, \ldots, x_{1,n}$,
- la valeur de X_2 pour chacun des $n_2 = n$ individus de E_2 : $x_{2,1}, \ldots, x_{2,n}$.

Pour tout $i \in \{1, \ldots, n\}$, sur le i-ème individu, on observe donc une paire de valeurs : $(x_{1,i}, x_{2,i})$. Si on prend le schéma "Traitement 1" et "Traitement 2", on peut mettre les données sous la forme :

Individus	Traitement 1	Traitement 2
ω_1	$x_{1,1}$	$x_{2,1}$
ω_2	$x_{1,2}$	$x_{2,2}$
$\vdots$	$\vdots$	$\vdots$
ω_n	$x_{1,n}$	$x_{2,n}$

Formules : p-valeurs

Lois : $T \sim \mathcal{T}(\nu)$, $\nu = n-1$, $K \sim \chi^2(1)$.

Outils : $\overline{d} = \frac{1}{n}\sum_{i=1}^n d_i$, $d_i = x_{1,i} - x_{2,i}$, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (d_i - \overline{d})^2}$, pour tout $(i,j) \in \{0,1\}^2$, on pose $n_{i,j}$ le nombre d'individus dans l'échantillon vérifiant $X_1 = i$ et $X_2 = j$.

$X_1 - X_2 \sim \mathcal{N}$, $\mathbb{E}(X_1) = \mu_1$, $\mathbb{E}(X_2) = \mu_2$	H_1	Stat. test obs.	p-valeurs
Paired T-Test	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \sqrt{n}\left(\frac{\overline{d} - d_0}{s}\right)$	$\mathbb{P}(\lvert T\rvert \geq \lvert t_{obs}\rvert)$ $\mathbb{P}(T \geq t_{obs})$ $\mathbb{P}(T \leq t_{obs})$
$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Stat. test obs.	p-valeurs
MacNemarTest cor	$p_1 \neq p_2$	$\chi^2_{obs} = \frac{(\lvert n_{0,1} - n_{1,0}\rvert - 1)^2}{n_{0,1} + n_{1,0}}$	$\mathbb{P}(K \geq \chi^2_{obs})$
MacNemarTest	$p_1 \neq p_2$	$\chi^2_{obs} = \frac{(n_{0,1} - n_{1,0})^2}{n_{0,1} + n_{1,0}}$	$\mathbb{P}(K \geq \chi^2_{obs})$

Commandes

On considère la librairie `stats` :

```
library(stats)
```

On propose les commandes R suivantes :

$X_1 - X_2 \sim \mathcal{N}$, $\mathbb{E}(X_1) = \mu_1$, $\mathbb{E}(X_2) = \mu_2$	H_1	Commandes
Paired T-Test	$\mu_1 \neq \mu_2$	`t.test(x1, x2, paired = T)$p.value`
	$\mu_1 > \mu_2$	`t.test(x1, x2, paired = T, alternative = "greater")$p.value`
	$\mu_1 < \mu_2$	`t.test(x1, x2, paired = T, alternative = "less")$p.value`
$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Commandes
MacNemarTest cor	$p_1 \neq p_2$	`mcnemar(x1, x2)`
MacNemarTest	$p_1 \neq p_2$	`mcnemar(x1, x2, correct = F)`

Exemples

Exemple 1. Un médecin ne veut se tromper que 5 fois sur 100 en décidant que l'administration d'un traitement particulier à un malade provoque en moyenne un accroissement de poids au bout de 3 mois de traitement. Le médecin examine le poids avant traitement et le poids après traitement de 5 malades choisis au hasard. Les résultats, en kilogrammes, sont :

Sujet n^o	Poids avant traitement	Poids après traitement
1	80.82	83.76
2	60.12	64.13
3	102.52	101.81
4	51.65	56.63
5	65.96	68.21

Le poids en kilogrammes d'un malade avant traitement peut être modélisé par une *var* X_1, et le poids en kilogrammes d'un malade après 3 mois de traitement peut être modélisé par une *var* X_2. On suppose que $X_1 - X_2$ suit une loi normale.

Proposer une modélisation du problème via un test statistique adapté et énoncer clairement votre conclusion.

Solution 1. Par l'énoncé, on observe

- la valeur de X_1, *var* d'espérance μ_1, pour chacun des n_1 individus (malades) d'un échantillon avec $n_1 = 5$,
- la valeur de X_2, *var* d'espérance μ_2, pour chacun des n_2 individus (malades) d'un échantillon avec $n_2 = 5$.

On suppose que $X_1 - X_2$ suit une loi normale.

Les échantillons sont appariés car ce sont les mêmes individus qui reçoivent les deux traitements. On pose $n = n_1 = n_2 = 5$.

On considère les hypothèses :

$$H_0 : \mu_1 \geq \mu_2 \qquad \text{contre} \qquad H_1 : \mu_1 < \mu_2.$$

On utilise un Paired T-Test. Il est unilatéral à gauche.

On considère les commandes :

```
x1 = c(80.82, 60.12, 102.52, 51.65, 65.96)
x2 = c(83.76, 64.13, 101.81, 56.63, 68.21)
t.test(x1, x2, paired = T, alternative = "less")$p.value
```

Cela renvoie : `[1] 0.02494845`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

En particulier, avec un risque 5%, on peut affirmer que le traitement provoque un accroissement du poids moyen.

Exemple 2. La prise d'un médicament M_1 anti-inflammatoire provoque quelquefois des douleurs gastriques. Le médecin propose la prise d'un médicament supplémentaire M_2 pour tenter d'éviter cet inconvénient. Ainsi, 87 malades présentant une affection inflammatoire et prenant le remède M_1 sont testés. On leur demande d'observer l'apparition ou non de douleurs gastriques avant et après l'administration du médicament supplémentaire M_2. Les résultats sont :

- 61 malades n'ont eu de douleurs gastriques ni avant ni après M_2
- 2 malades qui n'avaient pas eu de douleurs avant M_2 ont en eu après
- 11 malades qui ont eu de douleurs avant M_2 n'en ont plus eu après
- 13 malades ont eu de douleurs aussi bien avant qu' après M_2.

Peut-on affirmer que l'administration de M_2 a modifié la probabilité d'avoir des douleurs gastriques ?

Solution 2. Soient

- p_1 la probabilité inconnue d'avoir des douleurs gastriques avant M_2 et X_1 la *var* qui vaut 1 si l'individu a des douleurs gastrique avant M_2 et 0 sinon ; $X_1 \sim \mathcal{B}(p_1)$,
- p_2 la probabilité inconnue d'avoir des douleurs gastriques après M_2 et X_2 la *var* qui vaut 1 si l'individu a des douleurs gastrique après M_2 et 0 sinon ; $X_2 \sim \mathcal{B}(p_2)$.

Par l'énoncé, on observe

- la valeur de X_1 pour chacun des n_1 individus d'un échantillon avec $n_1 = 87$,
- la valeur de X_2 pour chacun des n_2 individus d'un échantillon avec $n_2 = 87$.

Les échantillons sont appariés car ce sont les mêmes individus qui sont considérés.

On considère les hypothèses :

$$H_0 : p_1 = p_2 \qquad \text{contre} \qquad H_1 : p_1 \neq p_2.$$

On utilise un MacNemarTest cor.

Les données sont n couples de valeurs : $(x_{1,1}, x_{2,1}), (x_{1,2}, x_{2,2}), \ldots, (x_{1,n}, x_{2,n})$ où $x_{1,i} \in \{0, 1\}$ et $x_{2,i} \in \{0, 1\}$, $i \in \{1, \ldots, n\}$. Comme on n'y a pas directement accès, on pose, pour tout $(i, j) \in \{0, 1\}^2$, $n_{i,j}$ le nombre d'individus dans l'échantillon vérifiant $X_1 = i$ et $X_2 = j$, et on considère la matrice :

$$A = \begin{pmatrix} n_{0,0} & n_{0,1} \\ n_{1,1} & n_{1,0} \end{pmatrix} = \begin{pmatrix} 61 & 2 \\ 11 & 13 \end{pmatrix}.$$

On propose les commandes :

```
A = matrix(c(61, 11, 2, 13), ncol = 2)

mcnemar.test(A)$p.value
```

Cela renvoie : `[1] 0.02650028`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

On peut affirmer que l'administration de M_2 modifie significativement la probabilité d'avoir des douleurs.

En cas de non normalité

Pour la comparaison de 2 moyennes, si l'hypothèse que les *var* suivent des lois normales ne semblent pas vérifiée (en faisant une analyse graphique ou un test statistique adéquat, comme le "test de Shapiro-Wilk"), certains tests dit "non paramétriques" proposent des alternatives satisfaisantes. Notamment, on peut utiliser :

- le test de Mann et Witney si les échantillons sont indépendants,
- le test de Wilcoxon si les échantillons sont appariés.

Les commandes associées sont :

```
wilcox.test(x, y)$p.value

wilcox.test(x, y, paired = TRUE)$p.value
```

6 Tests d'indépendance entre deux caractères

6.1 Cas de deux caractères qualitatifs

Contexte

Soient X et Y deux caractères non chiffrés (qualitatifs). On suppose que

- le caractère X a k modalités notées $a_1, \ldots, a_k$,
- le caractère Y a h modalités notées $b_1, \ldots, b_h$.

Remarque : On peut aussi considérés des caractères chiffrés (quantitatifs) avec des valeurs réparties dans quelques intervalles disjoints appelés classes. Dans ce cas, on les traite comme des caractères qualitatifs et leurs classes joueront le rôle de modalités.

Données

On observe les valeurs de (X, Y) sur un échantillon de n individus.

Ainsi, les données sont n couples de modalités : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ où $(x_i, y_i) \in \{a_1, \ldots, a_k\} \times \{b_1, \ldots, b_h\}$. Pour tout $(i, j) \in \{1, \ldots, k\} \times \{1, \ldots, h\}$, on pose $n_{i,j}$ le nombre d'individus dans l'échantillon vérifiant $X = a_i$ et $Y = b_j$. On dispose du tableau :

X \ Y	b_1	$\ldots$	b_j	$\ldots$	b_h
a_1	$n_{1,1}$	$\ldots$	$n_{1,j}$	$\ldots$	$n_{1,h}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	$n_{i,1}$	$\ldots$	$n_{i,j}$	$\ldots$	$n_{i,h}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	$n_{k,1}$	$\ldots$	$n_{k,j}$	$\ldots$	$n_{k,h}$

Enjeu

À partir de ces données, on souhaite affirmer avec un faible risque de se tromper, que X et Y ne sont pas indépendants. Il y aurait alors une liaison entre elles.

Hypothèses

Étant donné la problématique, on considère les hypothèses :

$$H_0 : \text{"les caractères } X \text{ et } Y \text{ sont indépendants" contre}$$

$$H_1 : \text{"les caractères } X \text{ et } Y \text{ ne sont pas indépendants".}$$

En représentant les caractères X et Y par des *var*, on pose :

$$p_{i,j} = \mathbb{P}(\{X = a_i\} \cap \{Y = b_j\}), \qquad p_{i,.} = \mathbb{P}(X = a_i) = \sum_{j=1}^{h} p_{i,j}, \qquad p_{.,j} = \mathbb{P}(Y = b_j) = \sum_{i=1}^{k} p_{i,j}.$$

Par la définition d'indépendance de deux *var*, on peut alors reformuler les hypothèses comme :

$$H_0 : \text{"}p_{i,j} = p_{i,.}p_{.,j} \text{ pour tout } (i,j) \in \{1,\dots,h\} \times \{1,\dots,k\}\text{" contre}$$

$$H_1 : \text{"il existe } (i_0,j_0) \in \{1,\dots,h\} \times \{1,\dots,k\} \text{ tel que } p_{i_0,j_0} \neq p_{i_0,.}p_{.,j_0} \text{ ".}$$

Test d'indépendance du Chi-deux

Pour mettre en œuvre le test d'indépendance du Chi-deux, pour tout $(i,j) \in \{1,\dots,h\} \times \{1,\dots,k\}$, on considère les quantités :

$$n_{i,.} = \sum_{j=1}^{h} n_{i,j}, \qquad n_{.,j} = \sum_{i=1}^{k} n_{i,j}, \qquad n^*_{i,j} = \frac{n_{i,.}n_{.,j}}{n}.$$

On peut éventuellement les mettre sous la forme d'un tableau :

X \ Y	b_1	$\dots$	b_j	$\dots$	b_h	Total
a_1	$n_{1,1}$ $(n^*_{1,1})$	$\dots$	$n_{1,j}$ $(n^*_{1,j})$	$\dots$	$n_{1,h}$ $(n^*_{1,h})$	$n_{1,.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	$n_{i,1}$ $(n^*_{i,1})$	$\dots$	$n_{i,j}$ $(n^*_{i,j})$	$\dots$	$n_{i,h}$ $(n^*_{i,h})$	$n_{i,.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	$n_{k,1}$ $(n^*_{k,1})$	$\dots$	$n_{k,j}$ $(n^*_{k,j})$	$\dots$	$n_{k,h}$ $(n^*_{k,h})$	$n_{k,.}$
Total	$n_{.,1}$	$\dots$	$n_{.,j}$	$\dots$	$n_{.,h}$	n

On suppose que, pour tout $(i,j) \in \{1,\dots,k\} \times \{1,\dots,h\}$, $n^*_{i,j} \geq 5$, condition minimale pour valider test d'indépendance du Chi-deux.

On calcule

$$\chi^2_{obs} = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(n_{i,j} - n^*_{i,j})^2}{n^*_{i,j}} = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{n^2_{i,j}}{n^*_{i,j}} - n.$$

Soit $K \sim \chi^2(\nu)$, $\nu = (k-1)(h-1)$. Alors la p-valeur associée au test d'indépendance du Chi-deux est

$$\text{p-valeur} = \mathbb{P}(K \geq \chi^2_{obs}).$$

Commandes

Les commandes associées sont données par (avec $k = 2$ et $h = 3$ par exemple) :

```
A = matrix(c(n11, n12, n13, n21, n22, n23), nrow = 2, byrow = T)

chisq.test(A)$p.value
```

Remarque : Lorsque $h = 2$ et $k = 2$, la commande `chisq.test` utilise par défaut la correction de Yates dans la définition de χ^2_{obs} : $\chi^2_{obs} = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(|n_{i,j} - n^*_{i,j}| - 0.5)^2}{n^*_{i,j}}$.

Exemple

Le tableau ci-dessous donne le nombre d'étudiants qui ont été brillants et médiocres devant trois examinateurs : examinateur A, examinateur B et examinateur C.

Résultat \ Examinateur	A	B	C
brillants	50	47	56
médiocres	5	14	8

Peut-on affirmer, au risque 5%, que le résultat d'un individu dépend de l'examinateur ?

Solution

Soient X le caractère qualitatif "résultat" et Y le caractère qualitatif "examinateur". Les modalités de X sont "brillant" et "médiocre", et les modalités de Y sont "A", "B" et "C" (on a $k = 2$ et $h = 3$). Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (étudiants) d'un échantillon avec $n = 180$. On considère les hypothèses :

$$H_0 : \text{"les caractères } X \text{ et } Y \text{ sont indépendants" contre}$$

$$H_1 : \text{"les caractères } X \text{ et } Y \text{ ne sont pas indépendants".}$$

Vu le tableau de données, on pose la matrice :

$$A = \begin{pmatrix} 50 & 47 & 56 \\ 5 & 14 & 8 \end{pmatrix}.$$

On considère les commandes :

```
A = matrix(c(50, 47, 56, 5, 14, 8), nrow = 2, byrow = T)

chisq.test(A)$p.value
```

Cela renvoie : `[1] 0.08872648`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'application du test sont vérifiées.

Comme p-valeur > 0.05, on ne rejette pas H_0. Les données ne nous permettent pas de rejeter l'indépendance de X et Y.

Remarque : Si un "Warning message" apparaît, la p-valeur renvoyée n'est pas fiable. Une solution alternative est d'utiliser le test exact de Fisher (plus puissant). Un exemple de commande est :

```
fisher.test(A)$p.value
```

6.2 Cas de deux caractères quantitatifs

Contexte

Soient X et Y deux caractères chiffrés (quantitatifs). On observe les valeurs de (X, Y) sur un échantillon de n individus. Ainsi, les données sont n couples de valeurs : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. On dispose donc du tableau :

X	x_1	x_2	$\ldots$	x_n
Y	y_1	y_2	$\ldots$	y_n

Enjeu

À partir de ces données, on souhaite affirmer avec un faible risque de se tromper, que X et Y ne sont pas indépendants. Il y aurait alors une liaison entre elles.

Hypothèses

Étant donné la problématique, on considère les hypothèses :

H_0 : "les caractères X et Y sont indépendants" contre

H_1 : "les caractères X et Y ne sont pas indépendants".

En représentant les caractères X et Y par des *var*, on définit le coefficient de corrélation ρ par

$$\rho = \frac{\mathbb{C}(X,Y)}{\sigma(X)\sigma(Y)}.$$

De plus, on suppose que (X,Y) est un vecteur de *var* suivant une loi normale bidimensionnelle. Grâce à cette hypothèse, on a l'équivalence : X et Y indépendantes $\Leftrightarrow \rho = 0$.

On peut alors reformuler les hypothèses comme :

$$H_0 : \rho = 0 \qquad \text{contre} \qquad H_1 : \rho \neq 0.$$

Remarque : En pratique, on peut représenter les données sur le repère orthonormé (O,I,J) par les points de coordonnées : $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. L'ensemble de ces points est appelé nuage de points. Si la silhouette de ce nuage de points est de forme ellipsoïdale, on peut admettre l'hypothèse de normalité sur (X,Y).

Test de nullité du coefficient de corrélation (de Pearson)

Pour mettre en œuvre le test de nullité du coefficient de corrélation, on considère les quantités :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2 \sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

On calcule

$$t_{obs} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}.$$

Soit $T \sim T(\nu)$, $\nu = n-2$. Alors la p-valeur associée au test de nullité du coefficient de corrélation est

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

Commandes

Les commandes associées sont données par :

```
cor.test(x,y)$p.value
```

Exemple

Sur 14 familles composées d'un père et d'un fils, on examine le QI du père et le QI du fils. Les résultats sont :

Père	121	142	108	111	97	139	131	90	115	107	124	103	115	151
Fils	102	138	126	133	95	146	115	100	142	105	130	120	109	123

Peut-on affirmer qu'il y a une liaison significative entre le QI du père et le QI du fils ?

Solution

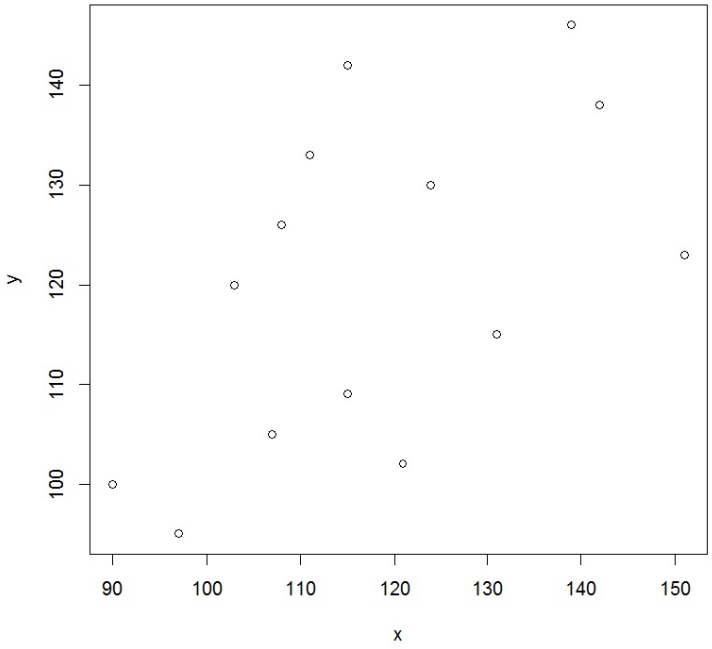
Soient X le caractère quantitatif "QI du père" et Y le caractère quantitatif "QI du fils". Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (familles) d'un échantillon avec $n = 14$. On considère les hypothèses :

$$H_0 : \text{"les caractères } X \text{ et } Y \text{ sont indépendants" contre}$$
$$H_1 : \text{"les caractères } X \text{ et } Y \text{ ne sont pas indépendants".}$$

Dans un premier temps, on considère les commandes :

```
x = c(121, 142, 108, 111, 97, 139, 131, 90, 115, 107, 124, 103, 115, 151)
y = c(102, 138, 126, 133, 95, 146, 115, 100, 142, 105, 130, 120, 109, 123)
plot(x, y)
```

Cela renvoie :



On constate que ce nuage de points est de forme ellipsoïdale ; en représentant les caractères comme des *var*, on peut admettre l'hypothèse de normalité sur la loi de (X, Y).

On fait :

```
cor.test(x, y)$p.value
```

Cela renvoie : `[1] 0.04090612`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, on peut affirmer qu'il y a une liaison significative entre le QI du père et le QI du fils.

En cas de non normalité

Si l'hypothèse que (X, Y) suit une loi normale bidimensionnelle ne semble pas vérifiée, on peut étudier l'indépendance de X et Y en faisant le test de nullité du coefficient de corrélation de Spearman. Les commandes associées sont :

```
cor.test(x, y, method = "spearman")$p.value
```

7 Exercices

Exercice 1. Sur un paquet de céréale "Croqus", une étiquette assure que le taux moyen de magnésium dans un paquet est de 94 milligrammes. On extrait au hasard 8 paquets "Croqus" dans la production et on mesure leur quantité de magnésium. Les résultats, en milligrammes, sont :

81.23	95.12	85.67	81.35	81.77	85.21	80.34	82.34

On suppose que la quantité en milligrammes de magnésium que contient un paquet de céréales de la production peut être modélisée par une *var* X suivant une loi normale.

Peut-on conclure, au risque 0.1%, que le taux moyen de magnésium d'un paquet n'est pas conforme à la valeur de référence ?

Exercice 2. Un charcutier normand produit des pâtés de campagne. Il affirme que ses pâtés pèsent en moyenne 223 grammes. On extrait de la production un échantillon de 25 pâtés et on les pèse. Les résultats, en grammes, sont :

225	224	225	221	230	229	219	224	226	222	220	221	229

226	221	231	219	222	223	224	220	223	223	224	222

On suppose que le poids en grammes d'un pâté de porc du charcutier peut être modélisé par une *var* X suivant une loi normale.

Peut-on affirmer, au risque 5%, que le charcutier a tort ?

Exercice 3. On effectue un sondage auprès de 620 personnes : 232 disent être en faveur d'une réforme fiscale. Peut-on affirmer que la proportion de personnes favorables à cette réforme est strictement supérieure à 22% ? Si oui, préciser le degré de significativité.

Exercice 4. On souhaite étudier l'homogénéité de deux champs de fraises, notés champ A et champ B, quant aux poids des fraises.

- Les pesées en grammes de 15 fraises choisies au hasard dans le champ A sont :

48.73	43.44	46.71	51.62	47.24	54.64	47.00	48.40
45.86	47.70	46.14	47.68	44.73	51.69	50.54	

- Les pesées en grammes de 15 fraises choisies au hasard dans le champ B sont :

44.89	34.31	42.74	53.36	41.98	41.64	47.24	37.86
45.89	40.88	40.85	38.60	44.38	44.52	38.26	

Le poids d'une fraise dans le champ A peut être modélisé par une *var* X_1 et le poids d'une fraise dans le champ B peut être modélisé par une *var* X_2. On suppose que X_1 et X_2 suivent des lois normales de variances égales.

Peut-on affirmer, au risque 2%, que le poids moyen d'une fraise diffère selon les champs ?

Exercice 5. Deux hypermarchés H1 et H2 appartenant à un même groupe mais situés dans des villes différentes proposent au rayon "pâtes alimentaires", à la fois des produits de la marque du groupe, notée MG, et des produits d'autres marques. Soient p_1 la proportion de produits MG vendus par H1 et p_2 la proportion de produits MG vendus par H2. Avant le bilan de l'année, le groupe veut savoir s'il y a une différence entre ces proportions. Pour cela il fait faire une enquête rapide :

- Sur 532 produits vendus par H1, 231 étaient de la marque MG.
- Sur 758 produits vendus par H2, 272 étaient de la marque MG.

A la suite de l'enquête, le groupe conclut qu'il a moins d'une chance sur 100 de se tromper en affirmant qu'il y a une différence entre les deux hypermarchés quant à la proportion réelle de produits vendus de la marque MG.

A-t-il raison ? Justifier votre réponse.

Exercice 6. Un médecin mesure la tension de 9 patients volontaires le matin et le soir.

Les résultats, en centimètres de mercure, sont :

Matin	Soir
13.12	13.92
13.54	13.89
15.12	14.51
14.51	14.78
12.12	10.97
13.10	13.58
13.98	14.52
11.21	11.54
14.44	13.54

La tension en centimètres de mercure d'un patient le matin peut être modélisée par une *var* X_1, et celle du soir peut être modélisée par une *var* X_2. On suppose que $X_1 - X_2$ suit une loi normale.

Peut-on affirmer, au risque 5%, qu'en moyenne la tension du soir est différente de celle du matin ?

Exercice 7. Un commercial fournissant les stations-service souhaite savoir s'il y a un lien entre l'achat de bières bouteilles et l'achat de paquets de chips. Pour le tester, il tire au hasard parmi les tickets de caisse d'une année.

- 92 clients ont acheté à la fois des bières et des chips
- 32 clients ont acheté des bières mais pas de chips
- 10 clients ont acheté des chips mais pas de bières
- 12 clients n'ont acheté ni bières ni chips

Il ne veut se tromper qu'une fois sur 100 en disant qu'il y a un lien entre ces deux types d'achat. Proposer un test statistique adapté au problème.

Exercice 8. On a interrogé 200 élèves d'un lycée sur le type d'études supérieures qu'ils désiraient entreprendre. Les résultats de l'enquête figurent dans le tableau ci-dessous :

Type d'étude \ Sexe	garçon	fille
littéraire	60	60
scientifique	42	18
technique	18	2

Semble-t-il exister une relation entre le choix des études et le sexe ?

Exercice 9. Dans une grande entreprise, on a évalué le niveau de stress au travail et mesuré le temps en minutes mis pour se rendre au travail de 550 salariés. Les résultats sont :

Niveau de stress \ Temps	<15	$[15, 45]$	>45
faible	91	136	48
modéré	39	37	38
élevé	38	69	54

Est-ce que le temps mis pour se rendre au travail a une influence sur le niveau de stress ?

Exercice 10. On s'intéresse à la dépendance possible entre l'âge d'un client d'une banque et le fait qu'il soit interdit de chéquier ou pas. Pour 810 clients, on dispose :

- de leur classe d'âge (caractère Y),
- du fait qu'il soit interdit de chéquier ou pas (caractère X, avec $X = 1$ si interdiction, et $X = 0$ sinon).

Le jeu de données "chequiers" est disponible ici :

`http://www.math.unicaen.fr/~chesneau/chequiers.txt`

1. Mettre le jeu de données sous la forme d'une data frame `w`, puis attacher les noms des colonnes.
2. Donner une brève description de `w`.

3. Compléter le tableau des effectifs suivant :

Âge / Interdit de chéquier	ai25	ai35	ai45	ai55	ai75
0					
1					

4. Peut-on affirmer, au risque 5%, qu'il ya une laison entre X et Y ?

Exercice 11. Le pouls est la traduction des battements du cœur au niveau des artères. Sa fréquence est une indication précieuse dans nombre de situations aigües. Celle-ci se mesure comme suit : une fois le pouls bien repéré, on compte les battements pendant 15 secondes et on multiplie par 4 ce nombre. Soient Y la fréquence maximale du pouls d'une personne et X son âge. Sur $n = 15$ personnes, on observe les valeurs de (X, Y) suivantes :

X	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Y	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

Évaluer le degré de significativité du lien existant entre X et Y.

Exercice 12. Reproduire et comprendre l'enjeu des commandes suivantes :

```
x = c(16.2, 24.8, 4.2, 10, 20, 15, 11.8, 1, 21, 14, 3, 13, 6.2, 12.2, 15.4)
y = c(6.1, 23.7, 3.4, 21.1, 18.2, 9.3, 8, 1.1, 14.7, 15.8, 11.1, 13.4, 22.5,
12.2, 22.8)
plot(x, y)
cor.test(x, y, method = "spearman")$p.value
```

8 Solutions

Solution 1. Par l'énoncé, on observe la valeur de $X \sim \mathcal{N}(\mu, \sigma^2)$ pour chacun des n individus (paquets) d'un échantillon avec $n = 8$, et μ et σ inconnus.

On considère les hypothèses :

$$H_0 : \mu = 94 \qquad \text{contre} \qquad H_1 : \mu \neq 94.$$

Comme σ est inconnu, on utilise un T-Test. Il est bilatéral.

On considère les commandes :

```
x = c(81.23, 95.12, 85.67, 81.35, 81.77, 85.21, 80.34, 82.34)
t.test(x, mu = 94)$p.value
```

Cela renvoie : `[1] 0.000680449`

Comme p-valeur < 0.001, le rejet de H_0 est hautement significatif $\star\star\star$.

On peut conclure, avec un risque 0.1%, que le taux moyen de magnésium d'un paquet n'est pas conforme à la valeur de référence.

Solution 2. Par l'énoncé, on observe la valeur de $X \sim \mathcal{N}(\mu, \sigma^2)$ pour chacun des n individus (pâtés) d'un échantillon avec $n = 25$, et μ et σ inconnus.

On considère les hypothèses :

$$H_0 : \mu = 223 \qquad \text{contre} \qquad H_1 : \mu \neq 223.$$

Comme σ est inconnu, on utilise un T-Test. Il est bilatéral.

On considère les commandes :

```
x = c(225, 224, 225, 221, 230, 229, 219, 224, 226, 222, 220, 221, 229, 226,
221, 231, 219, 222, 223, 224, 220, 223, 223, 224, 222)
t.test(x, mu = 223)$p.value
```

Cela renvoie : `[1] 0.291206`

Comme p-valeur > 0.05, on ne rejette pas H_0. Les données ne nous permettent pas d'affirmer que le charcutier à tort.

Solution 3. Soient p la proportion inconnue des personnes favorables à la réforme fiscale et X la *var* qui vaut 1 si l'individu y est favorable et 0 sinon ; $X \sim \mathcal{B}(p)$. Par l'énoncé, on observe la valeur de X pour chacun des n individus (personnes) d'un échantillon avec $n = 620$.

On considère les hypothèses :

$$H_0 : p \leq 0.22 \qquad \text{contre} \qquad H_1 : p > 0.22.$$

On utilise un 1-Prop-Z-Test cor. Il est unilatéral à droite.

On considère les commandes :

```
prop.test(232, 620, 0.22, alternative = "greater")$p.value
```

Cela renvoie : `[1] 1.48694e-20`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur < 0.001, le rejet de H_0 est hautement significatif $\star\star\star$.

On peut affirmer que la proportion de personnes favorables à cette réforme est "hautement significativement" strictement supérieure à 22%.

Solution 4. Par l'énoncé, on observe

- la valeur de $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ pour chacun des n_1 individus (fraises) d'un échantillon avec $n_1 = 10$, et μ_1 et σ_1 inconnus,
- la valeur de $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ pour chacun des n_2 individus (fraises) d'un échantillon avec $n_2 = 10$, et μ_2 et σ_2 inconnus.

On a $\sigma_1^2 = \sigma_2^2$. Les individus étant tous différents, les échantillons sont indépendants.

On considère les hypothèses :

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

On utilise un 2-Comp-T-Test avec pooled yes car on a l'égalité $\sigma_1^2 = \sigma_2^2$. Il est bilatéral.

On considère les commandes :

```
x1 = c(48.73, 43.44, 46.71, 51.62, 47.24, 54.64, 47.00, 48.40, 45.86, 47.70,
46.14, 47.68, 44.73, 51.69, 50.54))
x2 = c(44.89, 34.31, 42.74, 53.36, 41.98, 41.64, 47.24, 37.86, 45.89, 40.88,
40.85, 38.60, 44.38, 44.52, 38.26)
t.test(x1, x2, var.equal = T)$p.value
```

Cela renvoie : `[1] 0.0003957631`

Comme p-valeur < 0.001, le rejet de H_0 est hautement significatif $\star\star\star$.

On peut affirmer que le poids moyen d'une fraise diffère de manière "hautement significative" selon les champs.

Solution 5. Soient

- X_1 la *var* qui vaut 1 si le produit de H1 est vendu et 0 sinon ; $X_1 \sim \mathcal{B}(p_1)$,
- X_2 la *var* qui vaut 1 si le produit de H2 est vendu et 0 sinon ; $X_2 \sim \mathcal{B}(p_2)$.

Par l'énoncé, on observe

- la valeur de X_1 pour chacun des n_1 individus (produits vendus) d'un échantillon avec $n_1 = 532$,
- la valeur de X_2 pour chacun des n_2 individus (produits vendus) d'un échantillon avec $n_2 = 758$.

Les individus étant tous différents, les échantillons sont indépendants.

On considère les hypothèses :

$$H_0 : p_1 = p_2 \qquad \text{contre} \qquad H_1 : p_1 \neq p_2.$$

On utilise un 2-Prop-Z-Test. Il est bilatéral.

On considère les commandes :

```
prop.test(x = c(231, 272), n = c(532, 758))$p.value
```

Cela renvoie : `[1] 0.007489186`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur $\in]0.001, 0.01]$, le rejet de H_0 est très significatif $\star\star$. Ainsi, le groupe a raison.

Solution 6. Par l'énoncé, on observe

- la valeur de X_1, *var* d'espérance μ_1, pour chacun des n_1 individus (patients) d'un échantillon avec $n_1 = 9$,
- la valeur de X_2, *var* d'espérance μ_2, pour chacun des n_2 individus (patients) d'un échantillon avec $n_2 = 9$.

On suppose que $X_1 - X_2$ suit une loi normale.

Les échantillons sont appariés car ce sont les mêmes individus qui sont considérés dans les deux échantillons.

On considère les hypothèses :

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

On utilise un Paired T-Test. Il est bilatéral.

On considère les commandes :

```
x1 = c(13.12, 13.54, 15.12, 14.51, 12.12, 13.10, 13.98, 11.21, 14.44
x2 = c(13.92, 13.89, 14.51, 14.78, 10.97, 13.58, 14.52, 11.54, 13.54)
t.test(x1, x2, paired = T, alternative = "less")$p.value
```

Cela renvoie : `[1] 0.4798814`

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Comme p-valeur > 0.05, on ne rejette pas H_0. Les données ne nous permettent pas d'affirmer qu'il y a une différence de tension moyenne chez les patients entre le matin et le soir.

Solution 7. Soient X le caractère qualitatif "achat de chips" et Y le caractère qualitatif "achat de bières". Les modalités de X sont "oui" et "non", et les modalités de Y sont "oui" et "non". Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (étudiants) d'un échantillon avec $n = 146$. On considère les hypothèses :

H_0 : "les caractères X et Y sont indépendants" contre

H_1 : "les caractères X et Y ne sont pas indépendants".

Vu le tableau de données, on pose la matrice :

$$A = \begin{pmatrix} 92 & 32 \\ 10 & 12 \end{pmatrix}.$$

On considère les commandes :

```
A = matrix(c(92, 32, 10, 12), nrow = 2, byrow = T)

chisq.test(A)$p.value
```

Cela renvoie : `[1] 0.01407829`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est (seulement) significatif $\star$.

Ainsi, en considérant un risque 1%, on ne rejette pas H_0. Les données ne permettent pas d'affirmer qu'il y a un lien entre l'achat de bières et l'achat de chips.

Solution 8. Soient X le caractère qualitatif "sexe" et Y le caractère qualitatif "type d'étude". Les modalités de X sont "garçon" et "fille, et les modalités de Y sont "littéraire", "scientifique" et "technique". Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (élèves) d'un échantillon avec $n = 146$. On considère les hypothèses :

$$H_0 : \text{"les caractères } X \text{ et } Y \text{ sont indépendants" contre}$$
$$H_1 : \text{"les caractères } X \text{ et } Y \text{ ne sont pas indépendants".}$$

Vu le tableau de données, on pose la matrice :

$$A = \begin{pmatrix} 60 & 60 \\ 42 & 18 \\ 18 & 2 \end{pmatrix}.$$

On considère les commandes :

```
A = matrix(c(60, 60, 42, 18, 18, 2), nrow = 2, byrow = T)

chisq.test(A)$p.value
```

Cela renvoie : `[1] 0.02172885`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est (seulement) significatif $\star$.

Ainsi, on peut affirmer qu'il y a un lien "significatif" entre le sexe de l'élève et le type d'études supérieures qu'il désire entreprendre.

Solution 9. Soient X le caractère qualitatif "Niveau de stress" et Y le caractère (considéré comme) qualitatif "Temps". Les modalités de X sont "faible", "modéré" et "élevé", et les modalités de Y sont "<15", "[15, 45]" et ">45". Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (salariés) d'un échantillon avec $n = 550$. On considère les hypothèses :

$$H_0 : \text{"les caractères } X \text{ et } Y \text{ sont indépendants" contre}$$
$$H_1 : \text{"les caractères } X \text{ et } Y \text{ ne sont pas indépendants".}$$

Vu le tableau de données, on pose la matrice :

$$A = \begin{pmatrix} 91 & 136 & 48 \\ 39 & 37 & 38 \\ 38 & 69 & 54 \end{pmatrix}.$$

On considère les commandes :

```
A = matrix(c(91, 136, 48, 39, 37, 38, 38, 69, 54), nrow = 3, byrow = T)
chisq.test(A)$p.value
```

Cela renvoie : `[1] 0.0001378628`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur < 0.001, le rejet de H_0 est hautement significatif $\star\star\star$.

Ainsi, on peut affirmer que le temps mis pour se rendre au travail a une influence "hautement significative" sur le niveau de stress.

Solution 10.

1. On fait :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/chequiers.txt", header = T)
attach(w)
```

2. On fait :

```
str(w)
```

Cela renvoie :

```
'data.frame':   810 obs. of  2 variables:
 $ X: int  0 0 0 1 0 0 0 0 0 0 ...
 $ Y: Factor w/ 5 levels "ai25","ai35",..: 5 2 5 3 5 3 5 5 2 4 ...
```

3. On fait :

```
table(X, Y)
```

Cela renvoie :

```
   Y
X   ai25 ai35 ai45 ai55 ai75
  0   84  136  196  165  171
  1    6   20   16    9    7
```

D'où le tableau :

Interdit de chéquier \ Âge	ai25	ai35	ai45	ai55	ai75
0	84	136	196	165	171
1	6	20	16	9	7

4. Les caractères X et Y sont (considérés comme) qualitatif. On considère les hypothèses :

H_0 : "les caractères X et Y sont indépendants" contre

H_1 : "les caractères X et Y ne sont pas indépendants".

Comme on dispose des données brutes, on considère les commandes :

```
chisq.test(X, Y)$p.value
```

Cela renvoie : `[1] 0.02220152`

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'applications du test sont vérifiées.

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, on peut affirmer que l'âge du clien a une influence "significative" sur le fait qu'il soit interdit de chéquier.

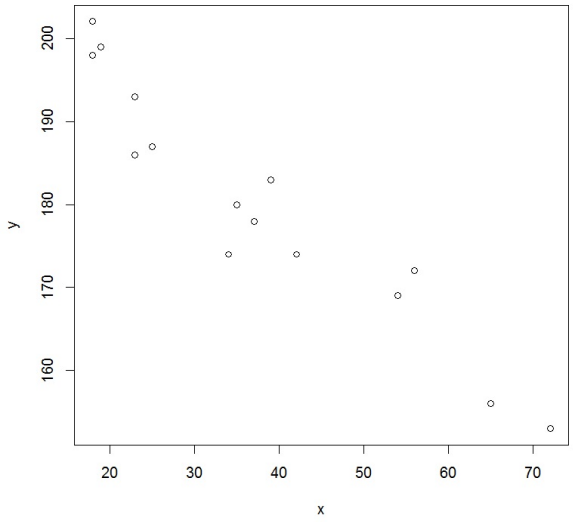
Solution 11. Soient X le caractère quantitatif "âge" et Y le caractère quantitatif "fréquence maximale du pouls". Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (personnes) d'un échantillon avec $n = 15$. On considère les hypothèses :

$$H_0 : \text{"les caractères } X \text{ et } Y \text{ sont indépendants" contre}$$
$$H_1 : \text{"les caractères } X \text{ et } Y \text{ ne sont pas indépendants".}$$

Dans un premier temps, on considère les commandes :

```
x = c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
y = c(202, 186, 187, 180, 156, 169, 174, 172, 153, 199, 193, 174, 198, 183,
178)
plot(x, y)
```

Cela renvoie :



On constate que ce nuage de points est de forme ellipsoïdale ; en représentant les caractères comme des *var*, on peut admettre l'hypothèse de normalité sur la loi de (X, Y).

On fait :

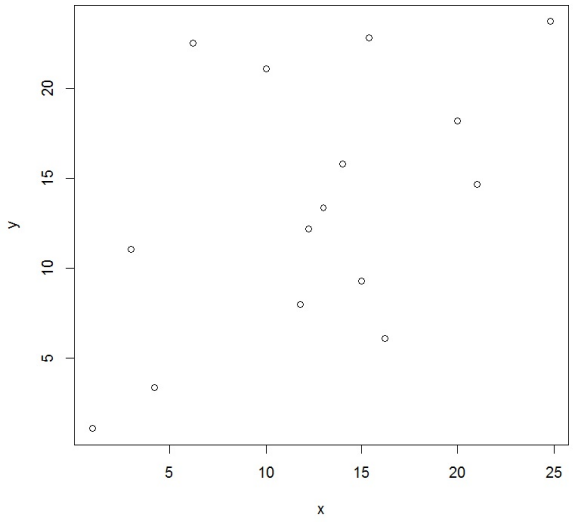
```
cor.test(x,y)$p.value
```

Cela renvoie : `[1] 3.847987e-08`

Comme p-valeur < 0.001, le rejet de H_0 est hautement significatif $\star\star\star$.

Ainsi, on peut affirmer qu'il y a une liaison "hautement significative" entre X et Y.

Solution 12. Dans un premier temps, on obtient le nuage de points :



On constate que la forme ellipsoïdale de celui-ci est discutable. Ainsi, pour étudier l'indépendance des caractères X et Y d'où émanent les données, on utilise le test de nullité du coefficient de corrélation de Spearman. La p-valeur de celui-ci est donnée par :

`[1] 0.0783316`

Comme p-valeur> 0.05, les données ne nous permettent pas de conclure ; on ne rejette pas l'hypothèse d'indépendance entre X et Y.