Version du 7 avril 2015

A. Samson

Rappels sur quelques tests statistiques (et les fonctions associées sous **R**)

Table des matières

1 Tests concernant des variables de Bernoulli

1.1 Test d'une probabilité

On s'intéresse au caractère A dans une population. La probabilité qu'un individu ait le caractère A est égale à p. Au vu d'un échantillon de taille n, on désire prendre une décision quant à la valeur de cette probabilité, au niveau α. On cherche à tester si la probabilité est égale à p_0. Les hypothèses du test sont

$$\begin{cases} H_0 & : p = p_0 \\ H_1 & : p \neq p_0 \end{cases}$$

A partir de l'échantillon, l'estimateur de la probabilité théorique sera la fréquence empirique $p_n = \frac{x}{n}$ où x est le nombre d'individus possédant le caractère A dans l'échantillon.

Méthode exacte

Sous H_0, $p_n \sim \mathcal{B}(n, p_0)$ et sous H_1, $p_n \sim \mathcal{B}(n, p)$ avec $p \neq p_0$.

La règle de décision est

– si $t_1 \leq p_n \leq t_2$, alors on rejette H_0

– si $t_1 \geq p_n$ ou $p_n \geq t_2$, alors on ne rejette pas H_0

où t_1 et t_2 sont définis par

$$P(t_1 \leq \mathcal{B}(n, p_0) \leq t_2) = 1 - \alpha$$

La fonction R permettant de réaliser ce test est `binom.test(x,n,p=p0)`.

Méthode approchée

Si la taille de l'échantillon est suffisamment grande (en pratique, $np_0 > 5$ et $n(1-p_0) > 5$), on considère la statistique de test

$$S_n = \frac{p_n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

On peut montrer que

$$S_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est

– si $|S_n| \geq s_\alpha$, alors on rejette H_0

– si $|S_n| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est `prop.test(x,n,p=p0)`.

1.2 Comparaison de deux probabilités : échantillons appariés

On s'intéresse au caractère A d'une population dans deux conditions différentes appelées expérience 1 et expérience 2. On suppose que sous l'expérience 1, la probabilité qu'un individu possède le caractère A est égale à p_1 et vaut p_2 sous l'expérience 2. On considère un échantillon de n individus qu'on soumet aux deux conditions d'expériences. On note $(X_1, \ldots, X_n)$ les observations obtenues lors de l'expérience 1 et $(Y_1, \ldots, Y_n)$ celles obtenues lors de l'expérience 2. Par nature, les échantillons X et Y ne sont pas indépendants puisqu'ils sont mesurés sur les mÍmes individus. Les hypothèses du test d'égalité des probabilités sont

$$\begin{cases} H_0 & : p_1 = p_2 \\ H_1 & : p_1 \neq p_2 \end{cases}$$

A partir des deux expériences, on mesure le nombre $x_{1,1}$ d'individus ayant le caractère A dans les deux expériences, $x_{1,0}$ le nombre d'individus n'ayant pas le caractère A dans l'expérience 1 mais l'ayant dans l'expérience 2, $x_{0,1}$ le nombre d'individus ayant le caractère A dans l'expérience 1 mais pas dans l'expérience 2 et enfin $x_{0,0}$ le nombre d'individus n'ayant pas le caractère A dans les deux expériences. On peut résumer les données dans le tableau de contingence suivant Comme les échantillons sont appariés, les effectifs $n_{1,1}$ et $n_{0,0}$ n'apportent pas d'information sur l'écart entre les deux expériences. En revanche, les individus associés aux deux autres couplages sont ceux qui contribuent à la différence entre les deux expériences. On reformule les hypothèses du test. On considère la population formée des individus qui ont changé de caractère entre les deux expériences. L'hypothèse nulle est l'égale répartition de ces individus entre le couplage "caractère

	expérience 1	
expérience 2	A	non A
A	$x_{1,1}$	$x_{0,1}$
non A	$x_{1,0}$	$x_{0,0}$

A à l'expérience 1 et non A à l'expérience 2" et le couplage "caractère non A à l'expérience 1 et A à l'expérience 2". On note $p_{1,0}$ et $p_{0,1}$ les probabilités de ces deux couplages respectivement. Les hypothèses du test sont

$$\begin{cases} H_0 & : p_{1,0} = p_{0,1} \\ H_1 & : p_{1,0} \neq p_{0,1} \end{cases}$$

On construit alors un tableau des effectifs observés et des effectifs théoriques. On note $n_{1/0} = x_{1,0} + x_{0,1}$ la taille de la population considérée.

	$(1,0)$	$(0,1)$	total
Observé	$x_{1,0}$	$x_{0,1}$	$n_{1/0}$
Théorique	$n_{1/0}/2$	$n_{1/0}/2$	$n_{1/0}$

On peut alors construire un test du χ^2 de symétrie, aussi appelé test de Mc Nemar. La statistique du test est

$$mcN_n = \frac{(x_{1,0} - n_{1/0}/2)^2}{x_{1/0}/2} = \frac{(x_{1,0} - x_{0,1})^2}{x_{1,0} + x_{0,1}}$$

On peut montrer que

$$mcN_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(1) \text{ sous } H_0$$

La règle de décision est
- si $mcN_n \geq s_\alpha$, alors on rejette H_0
- si $mcN_n < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(\chi^2(1) \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est `mcnemar.test(table)` *où* `table=matrix(c(`$x_{1,1}$`,`$x_{0,1}$`,`$x_{1,0}$`,`$x_{0,0}$`),2,2)` *est le tableau de contingence.*

1.3 Comparaison de deux probabilités : échantillons indépendants

Soient p_1 et p_2 les probabilités qu'un individu ait une certaine modalité A dans les populationsM_1 et M_2 respectivement. On extrait un échantillon de taille n_1 et n_2 dans les populations M_1 et M_2 respectivement. On note $(X_1, \ldots, X_n)$ les observations obtenues dans la population M_1 et $(Y_1, \ldots, Y_n)$ celles obtenues dans la population M_2. On teste à partir de ces échantillons si les probabilités sont les mÍmes dans les deux populations. Les hypothèses du test d'égalité des probabilités sont

$$\begin{cases} H_0 & : p_1 = p_2 \\ H_1 & : p_1 \neq p_2 \end{cases}$$

Les variables aléatoires X_i, $i = 1, \ldots, n$ sont i.i.d. de loi $\mathcal{B}(p_1)$ et les Y_i, $i = 1, \ldots, n$ sont i.i.d. de loi $\mathcal{B}(p_2)$. On note x et y le nombre d'individus possédant le caractère A dans l'échantillon de taille n_1 de la population M_1 et dans l'échantillon de taille n_2 de la population M_2 respectivement. On dispose d'une estimation $f_1 = \frac{x}{n_1}$ et $f_2 = \frac{y}{n_2}$ de p_1 et p_2 respectivement. On se ramène au test d'une probabilité en reformulant les hypothèses du test

$$\begin{cases} H_0 & : p_1 - p_2 = 0 \\ H_1 & : p_1 - p_2 \neq 0 \end{cases}$$

Il s'agit donc de comparer à 0 la différence $p_1 - p_2$.

Sous H_0, x et y ont mÍme loi. On estime alors la variance de la variable $f_1 - f_2$ par $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$. La statistique de test est

$$S_{n_1,n_2} = \frac{f_1 - f_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Méthode approchée Si la taille de l'échantillon est suffisamment grande (en pratique, $n_1 p_1 > 5$, $n_2 p_2 > 5$, $n_1(1-p_1) > 5$ et $n_2(1-p_2) > 5$), on peut montrer que

$$S_{n_1,n_2} \xrightarrow[n_1,n_2 \to \infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|S_{n_1,n_2}| \geq s_\alpha$, alors on rejette H_0
- si $|S_{n_1,n_2}| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est
`prop.test(matrix(x1,x2,n1,n2),2,2).`

Ce test est équivalent au test du chi-deux d'indépendance pour deux variables X et Y binaires, qu'on peut réaliser à l'aide de la fonction `chisq.test`.

1.4 Comparaison de plus de deux probabilités : échantillons appariés

On s'intéresse au caractère A d'une population dans K conditions différentes. On suppose que sous les conditions k, $k = 1, \ldots, K$, la probabilité qu'un individu possède le caractère A est égale à p_k. On considère un échantillon de n individus qu'on soumet aux K conditions d'expériences. On désire prendre une décision quant à l'égalité des K probabilités, au niveau α. Par nature, les proportions $p_1, \ldots, p_K$ sont liées puisqu'elles sont mesurées sur les mÍmes individus. Les hypothèses du test d'égalité des probabilités sont

$$\begin{cases} H_0 & : p_1 = \ldots = p_K \\ H_1 & : \exists i \neq j, p_i \neq p_j \end{cases}$$

au risque α.

La méthode statistique pour répondre à cette question généralise le test de Mac Nemar d'égalité de deux probabilités pour échantillons appariés présenté dans le paragraphe 1.2. Ce test du Chi-deux est appelé test de Cochran-Mantel-Haenszel.

La fonction R permettant de réaliser ce test est `mantelhaen.test(table)` *où* `table` *est le tableau de contingence des K populations.*

1.5 Comparaison de plus de deux probabilités : échantillons indépendants (test du Chi-deux d'independance de Pearson

Soit p_{ij} la probabilité qu'un individu ait une certaine modalité A_i $(i = 1, \ldots, I)$ dans la population M_j, $j = 1, \ldots, J$. On extrait J échantillons de taille $n_1, \ldots, n_J$ dans les populations $M_1, \ldots, M_J$ respectivement. On teste à partir de ces échantillons si les probabilités sont les mÍmes dans les J populations. Les hypothèses du test d'égalité des probabilités sont

$$\begin{cases} H_0 & : p_{ij} = p_{ij'} \forall i, j, j' \\ H_1 & : \exists i, \exists j \neq j', p_{ij} \neq p_{ij'} \end{cases}$$

On utilise alors un test du Chi-deux d'indépendance de Pearson qui généralise le test du Chi-deux d'égalité de deux probabilités présenté dans le paragraphe 1.3.

Le tableau de fréquences observées se présente de la faÁon suivante.

La proportion théorique attendue est $E_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$. La statistique du test de Chi2 d'indépendance de Pearson est

$$X_N^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

modalité	population				Total
A_1	$O_{1,1}$	$O_{1,2}$	$\dots$	$O_{1,J}$	$n_{1\cdot} = \sum_{j=1}^{J} O_{1j}$
A_2	$O_{2,1}$	$O_{2,2}$	$\dots$	$O_{2,J}$	$n_{2\cdot} = \sum_{j=1}^{J} O_{2j}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
A_I	$O_{I,1}$	$O_{I,2}$	$\dots$	$O_{I,J}$	$n_{I\cdot} = \sum_{j=1}^{J} O_{Ij}$
total	$n_{\cdot 1} = \sum_{i=1}^{I} O_{i1}$	$n_{\cdot 2} = \sum_{i=1}^{I} O_{i2}$	$\dots$	$n_{\cdot J} = \sum_{i=1}^{I} O_{iJ}$	$N = \sum_{i=1}^{I} O_{i1}$

On peut montrer que sous H_0,

$$X_N^2 \xrightarrow[N\to\infty]{\mathcal{L}} \chi^2((I-1)(J-1)) \text{ sous } H_0$$

La fonction R permettant de réaliser ce test est `chisq.test(table)` *où* `table` *est le tableau de contingence des I populations.*

2 Tests concernant des variables quantitatives

2.1 Test d'une espérance

On dispose de n observations $(x_1, \ldots, x_n)$. On suppose que les x_i sont des réalisations de variables aléatoires $(X_i)_{1 \leq i \leq n}$, qui sont indépendantes, identiquement distribuées d'espérance μ et de variance σ^2. Les paramètres μ et σ^2 sont supposés inconnus.

On cherche à savoir si l'espérance μ est égale à une valeur donnée μ_0. Les hypothèses du test sont donc

$$\begin{cases} H_0 & : \mu = \mu_0 \\ H_1 & : \mu \neq \mu_0 \end{cases}$$

La moyenne empirique $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ est un estimateur de la moyenne et la variance empirique $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ un estimateur de la variance. $\bar{x}$ et s^2 sont des réalisations de variables aléatoires $\bar{X}$ et S^2. On considère alors la statistique de test

$$T_n = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Cas gaussien On suppose que les variables aléatoires $(X_i)_{1 \leq i \leq n}$ sont de loi normale $\mathcal{N}(\mu, \sigma^2)$. Par construction, sous H_0, $\overline{X}$ suit une loi normale d'espérance μ_0 et de variance σ^2/n, et S^2 suit une loi du Chi-deux. On peut montrer que

$$T_n \sim \mathcal{T}(n-1) \text{ sous } H_0$$

où $\mathcal{T}(n-1)$ est une loi de Student à $n-1$ degrés de liberté. La règle de décision est
- si $|T_n| \geq s_\alpha$, alors on rejette H_0
- si $|T_n| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{T}(n-1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est `t.test(x,mu=mu0)`.

Il faut alors vérifier les conditions d'application du test de Student, en particulier l'hypothèse de normalité de l'échantillon. Pour cela, on réalise un test de normalité dont les hypothèses sont

$$\begin{cases} H_0 & : (X_i) \text{ suit une loi normale} \\ H_1 & : (X_i) \text{ ne suit pas une loi normale} \end{cases}$$

Les fonctions R permettant de réaliser ce test sont `shapiro.test` *ou* `lillie.test` *du package* `nortest`.

Cas non gaussien On suppose que les variables aléatoires $(X_i)_{1 \leq i \leq n}$ sont de loi quelconque mais que n est grand. Par le théorème central limite, on peut montrer que

$$T_n \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|T_n| \geq s_\alpha$, alors on rejette H_0
- si $|T_n| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

Il n'existe pas de fonction R qui réalise ce test. Lorsque n est grand, les quantiles de la loi normale centrée réduite sont très proches des quantiles d'une loi de Student à $n-1$ degrés de liberté. On pourra donc utiliser la fonction `t.test(x,mu=mu0)` *comme* ***approximation*** *du test d'une espérance dans le cas non gaussien.*

2.2 Comparaison de deux échantillons : échantillons appariés

On dispose de n couples d'observations (x_i, y_i), mesures effectuées sur un mÍme individu. Plus précisément, on considère qu'une mÍme variable a été mesurée sur un mÍme individu i, dans des conditions différentes ou à deux instants différents. On considère que ces couples (x_i, y_i) sont les réalisations de variables aléatoires (X_i, Y_i) qui sont indépendantes, identiquement distribuées. Par nature les variables X_i et Y_i sont liées puisqu'elles sont des mesures effectuées sur un mÍme individu. Dans ce contexte on se demande si la loi de X est différente de celle de Y.

2.2.1 Test paramétrique d'égalité de deux espérances

On suppose que X a pour espérance μ_X et variance σ^2 et Y a pour espérance μ_Y et variance σ^2. Dans le cadre d'un test paramétrique, on cherche à comparer les espérances des deux échantillons. Les hypothèses du test sont

$$\begin{cases} H_0 & : \mu_X = \mu_Y \\ H_1 & : \mu_X \neq \mu_Y \end{cases}$$

On introduit la variable différence $D_i = X_i - Y_i$. Le test revient donc à tester la nullité de l'espérance de la variable D_i, en se ramenant au test d'une espérance. On considère donc la statistique de test :

$$T_n = \sqrt{n}\frac{\overline{D}}{S}$$

où $\overline{D} = \frac{1}{n}\sum_{i=1}^n D_i$ est la moyenne empirique des D_i et $S^2 = \frac{1}{n-1}\sum_{i=1}^n (D_i - \overline{D})^2$ est l'estimateur de la variance.

Cas gaussien On suppose que les variables D_i sont de lois gaussiennes. Sous H_0, l'espérance de D_i est nulle. On peut montrer que

$$T_n \sim \mathcal{T}(n-1) \text{ sous } H_0$$

où $\mathcal{T}(n-1)$ est une loi de Student à $n-1$ degrés de liberté.

La règle de décision est
- si $|T_n| \geq s_\alpha$, alors on rejette H_0
- si $|T_n| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{T}(n-1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est `t.test(x,y,paired=T)`.

Il faut alors vérifier l'hypothèse de normalité de l'échantillon $D_1, \ldots, D_n$ à l'aide d'un test de normalité. Les fonctions R permettant de réaliser ce test sont `shapiro.test` ou `lillie.test` du package `nortest`.

Cas non gaussien On suppose que les variables aléatoires $(D_i)_{1 \leq i \leq n}$ sont de loi quelconque mais que n est grand. Par le théorème central limite, on peut montrer que

$$T_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|T_n| \geq s_\alpha$, alors on rejette H_0
- si $|T_n| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

Il n'existe pas de fonction R qui réalise ce test. Lorsque n est grand, les quantiles de la loi normale centrée réduite sont très proches des quantiles d'une loi de Student à $n-1$ degrés de liberté. On pourra donc utiliser la fonction `t.test(x,y,paired=T)` *comme* ***approximation*** *du test d'une espérance dans le cas non gaussien.*

2.2.2 Test non paramétrique de symétrie de la distribution des différences

Les tests non paramétriques de comparaison de deux échantillons appariés sont le test du signe ou le test du signe et rangs de Wilcoxon. On introduit la variable différence $D_i = X_i - Y_i$. Ces deux tests sont des tests sur la médiane de D. On teste

$$\begin{cases} H_0 & : med_D = 0 \\ H_1 & : med_D > 0 \end{cases}$$

Test du signe Le test du signe ne s'intéresse qu'au signe de $X - Y$. C'est un test de symétrie de $X - Y$. On teste si $P(X - Y > 0) = P(X - Y \leq 0)$. On note $Z_i = I_{D_i > 0}$. Les variables Z_i sont des variables i.i.d., de loi $\mathcal{B}(p)$ avec $p = P(D > 0)$. Sous H_0, $p = 1/2$. On est dans le cadre d'un test sur le paramètre d'une variable de Bernoulli à la valeur $1/2$. On construit la statistique de test

$$S_n = \frac{\sqrt{n}(\overline{Z} - 1/2)}{\sqrt{1/2(1-1/2)}}$$

avec $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} I_{D_i>0}$. La loi de la statistique de test sous H_0 peut se calculer de manière exacte (loi binomiale) ou approchée (loi normale) (voir le paragraphe 1.1).

Test du signe et rangs de Wilcoxon Le test du signe et rangs de Wilcoxon teste si (X, Y) a mÍme loi que (Y, X). On range par ordre croissant les $|D_i|$ et on attribue un rang R_i à $|D_i|$ (rang moyen en cas d'ex-aequo). La statistique de test est

$$W_n = \sum_{i=1}^{n} R_i I_{D_i>0}$$

On peut montrer que

$$\frac{W_n - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par

- si $\left|\frac{W_n - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right| \geq s_\alpha$, alors on rejette H_0
- si $\left|\frac{W_n - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right| < s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser le test de signe et rangs de Wilcoxon est `wilcox.test(x,y, paired=T)`. *L'option* `exact=T` *permet de calculer des p-values exactes lorsqu'il y a moins de* $n = 50$ *observations. Attention, lorsqu'il y a des ex-aequos, la p-value calculée par la fonction* `wilcox.test` *est fausse. Il faut utiliser la fonction* `wilcox.exact` *du package* `exactRankTests` *qui réalise le test exact de signe et rang de Wilcoxon.*

2.3 Comparaison de deux échantillons : échantillons indépendants

On considère $X_1, \ldots, X_{n_1}$, n_1 variables aléatoires i.i.d. et $Y_1, \ldots, Y_{n_2}$, n_2 variables aléatoires i.i.d., avec $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$ indépendants. On cherche à savoir si les lois des deux variables aléatoires sont les mÍmes.

2.3.1 Test paramétrique d'égalité de deux espérances

On suppose que X et Y ont pour espérances μ_X et μ_Y et pour écart types σ_X et σ_Y. On teste

$$\begin{cases} H_0 & : \mu_X = \mu_Y \\ H_1 & : \mu_X \neq \mu_Y \end{cases}$$

On estime les moyennes empiriques $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ et $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ et les variance $S_X^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$ et $S_Y^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$ à partir des deux échantillons. La statistique de test dépend de l'égalité des variances.

Si $\sigma_X = \sigma_Y$. La statistique de test est

$$Z_n = \frac{m_X - m_Y}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ avec } S = \sqrt{\frac{n_1 s_X^2 + n_2 s_Y^2}{n_1 + n_2 - 2}}$$

Cas gaussien On suppose que les variables X et Y sont distribuées selon des lois gaussiennes. On peut montrer que

$$Z_n \sim \mathcal{T}(n_1 + n_2 - 2) \text{ sous } H_0$$

où $\mathcal{T}(n_1 + n_2 - 2)$ est une loi Student à $n_1 + n_2 - 2$ degrés de liberté. C'est un test de Student.

La règle de décision est donnée par

- si $|Z_n| > t_\alpha$, alors on rejette H_0
- si $|Z_n| \leq t_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{T}(n_1+n_2-2)| \geq t_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est `t.test(x,y,var.equal=T)`.

Cas non gaussien On suppose que les variables X et Y sont de lois quelconques mais que n_1 et n_2 sont suffisamment grands. On peut montrer que

$$Z_n \xrightarrow[n_1,n_2\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|Z_n| > s_\alpha$, alors on rejette H_0
- si $|Z_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

Il n'existe pas de fonction R qui réalise ce test. Lorsque n est grand, les quantiles de la loi normale centrée réduite sont très proches des quantiles d'une loi de Student à $n-1$ degrés de liberté. On pourra donc utiliser la fonction `t.test(x,y,var.equal=T)` *comme* ***approximation*** *du test de deux espérances dans le cas non gaussien.*

Si $\sigma_X \neq \sigma_Y$

La statistique de test est

$$Z_n = \frac{m_X - m_Y}{\sqrt{\frac{s_X^2}{n_1-1} + \frac{s_Y^2}{n_2-1}}}$$

Cas gaussien On suppose que les variables X et Y sont distribuées selon des lois gaussiennes. On peut montrer que

$$Z_n \sim \mathcal{T}(\nu) \text{ sous } H_0$$

où $\mathcal{T}(\nu)$ est une loi Student à ν degrés de liberté avec ν est l'entier le plus proche de

$$\frac{\left[\frac{s_X^2}{n_1-1} + \frac{s_Y^2}{n_2-1}\right]^2}{\frac{s_X^4}{(n_1-1)n_1^2} + \frac{s_Y^4}{(n_2-1)n_2^2}}$$

C'est le test de Welch. La règle de décision est donnée par
- si $|Z_n| > t_\alpha$, alors on rejette H_0
- si $|Z_n| \leq t_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{T}(n_1+n_2-2)| \geq t_\alpha) = \alpha$$

La fonction R permettant de réaliser ce test est `t.test(x,y,var.equal=F)`.

L'égalité des variances des deux échantillons peut se vérifier par un test d'homogénéité des variances. On peut par exemple utiliser un test de Fisher-Snedecor (fonction R : `var.test()`*).*

L'hypothèse de normalité de l'échantillon se vérifie à l'aide d'un test de normalité. Les fonctions R permettant de réaliser ce test sont `shapiro.test` *ou* `lillie.test` *du package* `nortest`.

Cas non gaussien On suppose que les variables X et Y sont de lois quelconques mais que n_1 et n_2 sont suffisamment grands. On peut montrer que

$$Z_n \xrightarrow[n_1,n_2\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|Z_n| > s_\alpha$, alors on rejette H_0
- si $|Z_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

Il n'existe pas de fonction R qui réalise ce test. Lorsque n est grand, les quantiles de la loi normale centrée réduite sont très proches des quantiles d'une loi de Student à $n-1$ degrés de liberté. On pourra donc utiliser la fonction `t.test(x,y,var.equal=F)` *comme* ***approximation*** *du test de deux espérances dans le cas non gaussien.*

2.3.2 Test non paramétrique d'égalité de deux distributions

On souhaite savoir si les lois des deux variables aléatoires X et Y sont les mÍmes, autrement dit on va tester

$$\begin{cases} H_0 & : X \text{ et } Y \text{ ont mÍme loi} \\ H_1 & : X \text{ et } Y \text{ n'ont pas mÍme loi} \end{cases}$$

Il existe deux tests non paramétriques dans ce cadre : le test de Wilcoxon de la somme des rangs et le test de Mann-Whitney.

Test de la somme des rangs de Wilcoxon On rassemble les 2 échantillons en un seul. On ordonne l'échantillon global : on interclasse les X_i et les Y_j pour obtenir une suite mélangée et ordonnée de X_i et Y_j. Sous H_0, l'alternance des X_i et Y_j doit Ítre à peu près régulière. Cette régularité dans l'alternance est mesurée par les rangs des X_i et Y_j dans l'échantillon ordonné. A chaque X_i, on associe son rang R_i dans l'échantillon global ordonné. On note $W_n = \sum_{i=0}^{n} R_i$. La statistique de test est

$$U_n = \frac{W_n - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}$$

On peut montrer que

$$U_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|U_n| > s_\alpha$, alors on rejette H_0
- si $|U_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

Test de Mann-Whitney On compte le nombre de couples (X_i, Y_j) pour lesquels $X_i > Y_j$. On introduit

$$MW_n = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} I_{X_i > Y_j} = W_n - n_1(n_1/1)/2$$

On peut montrer que

$$Z_n = \frac{MW_n - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|U_n| > s_\alpha$, alors on rejette H_0
- si $|U_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser le test de Mann-Whitney est
`wilcox.test(x,y)` *(contrairement à son nom, cette fonction ne réalise pas le test de la somme des rangs de Wilcoxon). Attention, lorsqu'il y a des ex-aequos, la p-value calculée par la fonction* `wilcox.test` *est fausse. Il faut utiliser la fonction* `wilcox.exact` *du package* `exactRankTests` *qui réalise le test exact de Mann-Whitney.*

2.4 Comparaison de plus de deux échantillons, échantillons indépendants

On dispose de K groupes. Pour chaque groupe, on mesure n_k observations indépendantes $x_{k,1}, \ldots, x_{k,n_k}$ d'une v.a. X_k ($k = 1, \ldots, K$). On veut savoir si les lois dans les K groupes sont les mÍmes. On note $n = \sum_{k=1}^{K} n_k$ l'effectif de l'échantillon global.

On note `x` *le vecteur de toutes les observations,* `groupe` *la variable associant le numéro de groupe à chaque observations. Attention dans R : la variable groupe doit Ítre de type "facteur". Si ce n'est pas le cas, on peut la transformer à l'aide de la commande* `groupe = factor(groupe)`.

2.4.1 Test paramétrique d'égalité d'espérances

On suppose que les v.a. X_k sont indépendantes, de lois normales de mÍme variance. On est dans le cadre de l'analyse de la variance (ANOVA) à 1 facteur. On note $\mu_1, \ldots, \mu_K$ les espérances de ces K v.a. On va donc tester les hypothèses

$$\begin{cases} H_0 & : \mu_1 = \ldots = \mu_K \\ H_1 & : \exists k \neq k', \mu_k \neq \mu_{k'} \end{cases}$$

L'analyse de la variance est basée sur une décomposition de la variance du modèle. On définit la somme des carrés totale $SCT = \sum_{k=1}^{K}(x_{ki} - \bar{x}_{..})^2$, la somme des carrés du modèle $SCM = \sum_{k=1}^{K}(\bar{x}_{k.} - \bar{x}_{..})^2$ et la somme des carrés résiduels $SCR = \sum_{k=1}^{K}(x_{ki} - \bar{x}_{k.})^2$ où $\bar{x}_{k.} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}$ est la moyenne du groupe k et $\bar{x}_{..} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} x_{ki}$ est la moyenne totale. On a alors la formule de décomposition de la variance :

$$SCT = SCM + SCR$$

La statistique de test de l'analyse de la variance est

$$F_n = \frac{SCT/(K-1)}{SCR/(n-K)}$$

On peut montrer que

$$F_n \sim \mathcal{F}(K-1, n-K) \text{ sous } H_0$$

où $\mathcal{F}(K-1, n-K)$ est une loi de Fisher à $K-1$ et $n-K$ degrés de liberté.

La règle de décision est donnée par

- si $F_n > f_\alpha$, alors on rejette H_0
- si $F_n \leq f_\alpha$, alors on ne rejette pas H_0

où f_α est défini par

$$P(\mathcal{F}(K-1, n-K) \geq f_\alpha) = \alpha$$

On peut ensuite réaliser une ANOVA à l'aide de la commande `summary(aov(x~groupe))`.

2.4.2 Test non paramétrique de comparaison des distributions

Dans le cadre non paramétrique, on ne fait aucune hypothèse sur les distributions des v.a. X_k. On note $P_1, \ldots, P_K$ les distributions des v.a. On va donc tester les hypothèses

$$\begin{cases} H_0 & : P_1 = \ldots = P_K \\ H_1 & : \exists k \neq k', P_k \neq P_{k'} \end{cases}$$

Le test de Kruskal-Wallis est une généralisation du test de Mann-Whitney au cas de K échantillons. On ordonne l'échantillon global. Sous H_0, l'alternance entre les K groupes est a peu près régulière. On affecte un rang R_{ki} à l'observation X_{ki} dans l'échantillon global ordonné. On calcule $\bar{R}_{k.} = \frac{1}{n_k}\sum_{i=1}^{n_k} R_{ki}$ la moyenne des rangs du groupe k, $\bar{R}_{..} = \frac{n+1}{2}$ la moyenne de tous les rangs. On introduit la somme des carrés des rangs du modele $SCM = \sum_{k=1}^{K} n_k(\bar{R}_{k.} - \bar{R}_{..})^2$ et la somme des carrés des rangs résiduels $SCR = \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(R_{ki} - \bar{R})^2$. La statistique du test de Kruskal Wallis est

$$\begin{aligned} KW_n &= (n-1)\frac{\sum_{k=1}^{K} n_k(\bar{R}_{k\cdot} - \bar{R}_{\cdot\cdot})^2}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(R_{ki} - \bar{R}_{\cdot\cdot})^2} \\ &= \frac{12}{n(n+1)}\sum_{k=1}^{K}\frac{(\sum_{i=1}^{n_k} R_{ki})^2}{n_k} - 3(n+1) \\ &= \frac{SCM}{SCR} \end{aligned}$$

On peut montrer que

$$KW_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(K-1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $KW_n > t_\alpha$, alors on rejette H_0
- si $KW_n \leq t_\alpha$, alors on ne rejette pas H_0

où t_α est défini par

$$P(\chi^2(K-1) \geq t_\alpha) = \alpha$$

On peut remarquer que la statistique KW_n est similaire à l'analyse de la variance paramétrique. *La fonction R permettant de réaliser le test de Kruskal-Wallis est* `kruskal.test(x~groupe)`.

2.5 Comparaison de plus de deux échantillons, échantillons appariés

On mesure chez n individus le caractère A dans K conditions différentes ou à K temps différents. Pour chaque condition d'expérience, on mesure donc n observations indépendantes $x_{k,1}, \ldots, x_{k,n}$ d'une v.a. X_k $(k = 1, \ldots, K)$. On veut savoir si les lois dans les K conditions sont les mÍmes.

On note `x` *le vecteur de toutes les observations,* `groupe` *la variable associant le numéro de groupe à chaque observation. On note* `id` *la variable identité associant le numéro de l'individu à chaque observation. Attention dans R : les variable groupe et identité doivent Ítre de type "facteur". Si ce n'est pas le cas, on peut les transformer à l'aide de* la commande `groupe = factor(groupe)` et `id = factor(id)`.

2.5.1 Test paramétrique d'égalité des espérances

On suppose que les v.a. X_k sont indépendantes et de lois normales de mÍme variance. On est dans le cadre de l'analyse de la variance (ANOVA) à 2 facteurs : le facteur condition d'expérience et le facteur individu. C'est un plan d'expérience sans répétition, on ne dispose que d'une seule observation pour chaque traitement. On note $\mu_1, \ldots, \mu_K$ les espérances de ces K v.a. On va donc tester les hypothèses

$$\begin{cases} H_0 & : \mu_1 = \ldots = \mu_K \\ H_1 & : \exists k \neq k', \mu_k \neq \mu_{k'} \end{cases}$$

Ce test revient à tester l'effet du facteur expérience dans un modèle d'ANOVA 2 sans interaction. On définit la somme des carrés du au facteur expérience $SC_E = n\sum_{k=1}^{K}(\bar{x}_{k\cdot} - \bar{x}_{\cdot\cdot})^2$ et la somme des carrés résiduels du modèle d'ANOVA 2 sans interaction $SCR = \sum_{k=1}^{K}\sum_{i=1}^{n}(x_{ki} - \bar{x}_{k\cdot} - \bar{x}_{\cdot i} + \bar{x}_{\cdot\cdot})^2$. La statistique du test d'ANOVA 2 est

$$F_n = \frac{SC_E/(K-1)}{SCR/((n-1)(K-1))}$$

On peut montrer que

$$F_n \sim \mathcal{F}(K-1, (n-1)(K-1)) \text{ sous } H_0$$

où $\mathcal{F}(K-1, (n-1)(K-1))$ est une loi de Fisher à $K-1$ et $(n-1)(K-1)$ degrés de liberté.

La règle de décision est donnée par
- si $F_n > f_\alpha$, alors on rejette H_0
- si $F_n \leq f_\alpha$, alors on ne rejette pas H_0

où f_α est défini par

$$P(\mathcal{F}(K-1,(n-1)(K-1)) \geq f_\alpha) = \alpha$$

Il faut vérifier l'hypothèse d'homogénéité des variances, par exemple en utilisant le test de Bartlett `bartlett.test(x~groupe)` *et* `bartlett.test(x~ id)`.

On peut ensuite réaliser une ANOVA à 2 facteurs à l'aide de la commande `summary(aov(x~groupe+id))`.

2.5.2 Test non paramétrique de comparaison des distributions

Dans le cadre non paramétrique, on ne fait aucune hypothèse sur les lois des v.a. X_k. On note (P_k) les distributions des v.a. On va donc tester les hypothèses

$$\begin{cases} H_0 & : P_1 = \ldots = P_K \\ H_1 & : \exists k \neq k', P_k \neq P_{k'} \end{cases}$$

Le test de Friedman compare les sommes des rangs des I échantillons appariés. On ordonne l'échantillon global composé des $(x_{ki})_{K\times n}$ valeurs observées. On affecte à chaque observation x_{ki} son rang R_{ki} dans l'échantillon global. On calcule la moyenne des rangs du groupe k $\bar{R}_{k\cdot} = \frac{1}{n}\sum_{i=1}^{n} R_{ki}$, la moyenne de tous les rangs $\bar{R}_{\cdot\cdot} = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n} R_{ki} = \frac{nK+1}{2}$. On introduit la somme des carrés des rangs du modele $SCM = n\sum_{k=1}^{K}(\bar{R}_{k\cdot} - \bar{R}_{\cdot\cdot})^2$ et la somme des carrés des rangs résiduels $SCR = \frac{1}{n(K-1)}\sum_{k=1}^{K}\sum_{i=1}^{n}(R_{ki} - \bar{R})^2$. La statistique du test de Friedman est

$$F_{K,n} = n^2(K-1)\frac{\sum_{k=1}^{K}(\bar{R}_{k\cdot} - \bar{R}_{\cdot\cdot})^2}{\sum_{k=1}^{K}\sum_{i=1}^{n}(R_{ki} - \bar{R}_{\cdot\cdot})^2} = \frac{SCM}{SCR}$$

On peut montrer que

$$F_{K,n} \xrightarrow[K\to\infty]{\mathcal{L}} \chi^2(n-1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $F_{K,n} > s_\alpha$, alors on rejette H_0
- si $F_{K,n} \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(\chi^2(n-1) \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser le test de Friedman est `friedman.test(x,groupe,id)`.

2.6 Test d'égalité de deux variances

On considère deux échantillons indépendants $(X_1,\ldots,X_{n_x})$ et $(Y_1,\ldots,Y_{n_y})$ de variances respectives σ_X^2 et σ_Y^2. On cherche à tester

$$\begin{cases} H_0 & : \sigma_X^2 = \sigma_Y^2 \\ H_1 & : \sigma_X^2 \neq \sigma_Y^2 \end{cases}$$

2.6.1 Test paramétrique

Cas gaussien

On considère que les deux échantillons sont gaussiens, de lois respectives $\mathcal{N}(\mu_X,\sigma_X^2)$ et $\mathcal{N}(\mu_Y,\sigma_Y^2)$.

On introduit les estimateurs des variances : $S_x^2 = \dfrac{1}{n_x-1}\displaystyle\sum_{i=1}^{n_x}(X_i - \overline{X})^2$ et $S_y^2 = \dfrac{1}{n_y-1}\displaystyle\sum_{j=1}^{n_y}(Y_j - \overline{Y})^2$. La statistique du test de Fisher est :

$$F_n = \frac{S_x^2}{S_y^2}$$

On peut montrer que

$$F_n \sim \mathcal{F}(n_x-1, n_y-1) \text{ sous } H_0$$

où $\mathcal{F}(n_x-1,n_y-1)$ est une loi de Fisher à n_x-1 et n_y-1 degrés de liberté.

La règle de décision est donnée par

– si $F_n > f_\alpha$, alors on rejette H_0
– si $F_n \leq f_\alpha$, alors on ne rejette pas H_0

où f_α est défini par

$$P(\mathcal{F}(n_x-1, n_y-1) \geq f_\alpha) = \alpha$$

La fonction R permettant de réaliser le test d'égalité de variance de Fisher est `var.test(x,y)`.

2.6.2 Test non paramétrique

On ordonne l'échantillon global des X_i et Y_j de taille $n = n_x + n_y$. On attribue à chaque valeur un rang de symétrie : le rang 1 est à attribué à la plus petite et à la plus grande des n valeurs, le rang 2 est attribué à la plus petite et à la plus grande des $n-2$ valeurs restantes, etc. On note R_i^s le rang de symétrie des X_i dans l'échantillon global. La statistique du test d'Ansari-Bradley est

$$AB_{n_x,n_y} = \sum_{i=1}^{n_x} R_i^s$$

La loi de AB_{n_x,n_y} sous H_0 est tabulée pour de petites valeurs de n_x et n_y. On peut montrer que

$$S_n = \frac{AB_{n_x,n_y} - E(AB_{n_x,n_y})}{Var(AB_{n_x,n_y})} \xrightarrow[n_y\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
– si $|S_n| > s_\alpha$, alors on rejette H_0
– si $|S_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser le test d'égalité de variances d'Ansari-Bradley est `ansari.test(x,y)`.

2.7 Test d'égalité de plusieurs variances

2.7.1 Test paramétrique

Cas gaussien

On considère K échantillons indépendants $(X_{11}, \ldots, X_{1n_1}), \ldots, (X_{K1}, \ldots, X_{1n_K})$, de lois respectives $\mathcal{N}(\mu_1, \sigma_1^2), \ldots, \mathcal{N}(\mu_K, \sigma_K^2)$. On cherche à tester

$$\begin{cases} H_0 & : \sigma_1^2 = \ldots = \sigma_K^2 \\ H_1 & : \exists k, k', \sigma_k^2 \neq \sigma_{k'}^2 \end{cases}$$

On introduit les estimateurs de la variance de chaque échantillon $S_k^2 = \dfrac{1}{n_k - 1} \displaystyle\sum_{i=1}^{n_k} (X_{ki} - \overline{X_k})^2$. La statistique du test de Barlett est

$$B_n = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^k (n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^k (\frac{1}{n_i - 1}) - \frac{1}{N-k}\right)}$$

où $n = \sum_{k=1}^K n_k$ et $S^2 = \frac{1}{n-K}\sum_{k=1}^K (n_k - 1) S_k^2$ est l'estimateur global de la variance. On peut montrer que

$$B_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(K-1) \text{ sous } H_0$$

La règle de décision est donnée par
– si $B_n > s_\alpha$, alors on rejette H_0
– si $B_n \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(\chi^2(K-1) \geq s_\alpha) = \alpha$$

On note `data` *l'échantillon global, et* `groupe` *la variable prenant des valeurs entre* 1 *et* K *indiquant le groupe pour chaque observation (* `groupe` *doit Ítre un facteur* `groupe=factor(groupe)`*). La fonction R permettant de réaliser le test d'égalité de variance de Fisher est* `bartlett.test(data,groupe)`.

3 Test de corrélation

On mesure deux caractères continus sur n individus. On dispose ainsi d'un ensemble de couples de valeurs (x_i, y_i) chez l'ensemble des n individus. On suppose que les observations proviennent d'un échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ i.i.d. On cherche à savoir si les deux variables X et Y sont corrélées.

3.1 Test paramétrique

On suppose que l'échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ provient d'une loi normale bidimensionnelle, d'espérance (μ_x, μ_y) et de matrice de covariance :

$$\begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}.$$

C'est la loi d'un couple de variables, dont les espérances respectives sont μ_x et μ_y et les variances σ_x^2 et σ_y^2, le coefficient de corrélation étant ρ. L'estimateur naturel de ρ est le coefficient de corrélation empirique, à savoir la variable aléatoire R suivante :

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

où $\overline{X}$ et $\overline{Y}$ désignent les moyennes empiriques des X_i et des Y_i respectivement. L'hypothèse nulle que l'on souhaite tester est

$$\begin{cases} H_0 & : \rho = 0 \\ H_1 & : \rho \neq 0 \end{cases}$$

La statistique de test est

$$T_n = \sqrt{n-2}\frac{R}{\sqrt{1-R^2}}$$

On peut montrer que

$$T_n \sim \mathcal{T}(n-2) \text{ sous } H_0$$

où $\mathcal{T}(n-2)$ est la loi de Student à $n-2$ degrés de liberté. Ce test est appelé test de corrélation de Pearson.

La règle de décision est donnée par

- si $T_n > t_\alpha$, alors on rejette H_0
- si $T_n \leq t_\alpha$, alors on ne rejette pas H_0

où t_α est défini par

$$P(\mathcal{T}(n-2) \geq t_\alpha) = \alpha$$

La fonction R permettant de réaliser le test de corrélation de Pearson est
`cor.test(x,y,method="pearson").`

3.2 Test non paramétrique

On ne fait aucune hypothèse sur la loi de l'échantillon. On cherche à tester

$$\begin{cases} H_0 & : X \text{ et } Y \text{ sont indépendantes} \\ H_1 & : X \text{ et } Y \text{ ne sont pas indépendantes} \end{cases}$$

Il existe deux tests de corrélation, le test de corrélation de Spearman et celui de Kendall.

Test de corrélation de Spearman On ordonne chaque échantillon séparément. On calcule pour chaque individu i la différence d_i entre le classement de l'observation X_i dans l'échantillon des X ordonné et le classement de l'observation Y_i dans l'échantillon des Y ordonné. Le coefficient de corrélation de Spearman est donné par :

$$\rho_n = 1 - 6\frac{\sum_{i=1}^{i=n} d_i^2}{n(n^2-1)}$$

On peut montrer que

$$\sqrt{n-1}\rho_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

La règle de décision est donnée par
- si $|\rho_n| > s_\alpha$, alors on rejette H_0
- si $|\rho_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser le test de corrélation de Spearman est `cor.test(x,y,method="spearman")`.

Test de corrélation de Kendall On note c_k le nombre de paires de couples (x_i, y_i), (x_j, y_j) qui sont concordantes, c'est à dire telles que $(x_j - x_i)(y_j - y_i) > 0$ avec $1 \leq i < j \leq n$. On note d_k le nombre de paires de couples (x_i, y_i), (x_j, y_j) qui sont discordantes, c'est à dire telles que $(x_j - x_i)(y_j - y_i) < 0$ avec $1 \leq i < j \leq n$. On note $s_k = c_k - d_k$. Le tau empirique de Kendall est défini par

$$\tau_n = \frac{2s_k}{n(n-1)}$$

On peut montrer que

$$\frac{\tau_n}{var(\tau_n)} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ sous } H_0$$

On en déduit alors la région de rejet du test. La règle de décision est donnée par
- si $|\tau_n| > s_\alpha$, alors on rejette H_0
- si $|\tau_n| \leq s_\alpha$, alors on ne rejette pas H_0

où s_α est défini par

$$P(|\mathcal{N}(0,1)| \geq s_\alpha) = \alpha$$

La fonction R permettant de réaliser le test de corrélation de Kendall est `cor.test(x,y,method="kendall")`.

4 Test d'ajustement à la famille gaussienne

On considère une variable aléatoire réelle continue X de loi inconnue P_X, de fonction de répartition continue F. A partir d'un n-échantillon $(X_1, \ldots, X_n)$ de X de loi P, on cherche à tester si la loi P appartient à la famille gaussien de fonction de répartition F_θ avec $\theta = (\mu, \sigma^2)$ inconnu. On va tester

$$\begin{cases} H_0 \ : \text{ la loi } F = F_\theta \\ H_1 \ : \text{ la loi } F \neq F_\theta \end{cases}$$

4.1 Test de Kolmogorov-Smirnov

La première étape consiste à estimer θ sous H_0 par $\widehat{\theta} = (\overline{X}, S^2)$ avec $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. La deuxième étape consiste à tester

$$\begin{cases} H_0 \ : \text{ la loi } F = F_{\widehat{\theta}} \\ H_1 \ : \text{ la loi } F \neq F_{\widehat{\theta}} \end{cases}$$

Pour construire le test, il faut estimer la fonction de répartition de X. La fonction de répartition est définie par $F(x) = P(X \leq x) = E(I_{X \leq x})$. Pour tout x dans $\mathbb{R}$, on peut estimer $F(x)$ par

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{X_i \leq x}$$

Cet estimateur $\widehat{F}_n(x)$ est la fonction de répartition empirique de $(X_1, \ldots, X_n)$.

La statistique de test est

$$T_n = \left\| \widehat{F}_n - \Phi_{\widehat{\mu},\widehat{\sigma}^2} \right\|_\infty \text{ avec } \Phi_{\mu,\sigma^2} = P(\mathcal{N}(\mu, \sigma^2) \leq x)$$

On peut montrer que

$$P(\sqrt{n}(1 + 0.85 - 0.01/n)T_n > 0.895) = 0.05 \text{ sous } H_0$$

La fonction R permettant de réaliser le test de Kolmogorov-Smirnov d'ajustement à la famille gaussienne est `lillie.test(x)`*, du package* `nortest`*. ATTENTION, la fonction* `ks.test` *ne réalise pas le test d'ajustement à une famille de loi, mais le test d'ajustement à une loi connue. On ne peut pas l'utiliser dans ce contexte.*

4.2 Test de Shapiro-Wilks

On estime la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. On ordonne l'échantillon $X_1, \ldots, X_n$. On note $X_{(1)} \leq \ldots \leq X_{(n)}$ l'échantillon ordonné.

La statistique de test est

$$W_n = \frac{(\sum_{i=1}^n a_i X(i))^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

où les coefficients a_i sont données par le vecteur $a = (a_1, \ldots, a_n)$ défini par

$$a' = MV^{-1}(M'V^{-1}V^{-1}M)^{-1/2}$$

où M sont les valeurs attendues pour une statistique d'ordre d'un échantillon de taille n de loi normale centrée réduite, et V est la matrice de variance-covariance associée. Le numérateur est la fonction des étendues partielles alors que le dénominateur est la fonction des carrés des écarts à la moyenne. On peut tabuler la loi de la statistique sous H_0.

La règle de décision est donnée par

- si $W_n > s_\alpha$, alors on rejette H_0
- si $W_n \leq s_\alpha$, alors on ne rejette pas H_0

où s_α se lit dans la table du test de Shapiro-Wilks.

La fonction R permettant de réaliser le test de normalité de Shapiro-Wilks est `shapiro.test(x)`*.*