Statistique pour petits échantillons

Ecole Doctorale des Sciences de la Vie et de la Santé
Université Louis Pasteur
Module de formation

Frédéric Bertrand et Myriam Maumy[1]

19 et 20 septembre 2007

[1]Institut de Recherche Mathématique Avancée, Université Louis Pasteur
7, rue René Descartes, 67084 Strasbourg Cedex
fbertran@math.u-strasbg.fr
mmaumy@math.u-strasbg.fr

Première partie

Notes de cours

Chapitre 1

Quelques tests non paramétriques.[1]

1.1. Les tests non paramétriques sur un échantillon

Dans cette section nous nous intéressons à deux tests non paramétriques :
– le test du signe et
– le test des rangs signés.
Nous utiliserons de préférence le test des rangs signés dès que les conditions de son utilisation sont remplies, sa puissance étant alors supérieure à celle du test du signe.

1.1.1. Test du signe

Soit un échantillon indépendant et identiquement distribué $X_1, \ldots, X_n$ d'une loi continue F dont la valeur médiane est notée m_e et la moyenne μ. Le test du signe permet de tester les hypothèses suivantes :

Hypothèses :

$$\boxed{\mathcal{H}_0 : m_e = 0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] = \frac{1}{2}}$$

contre

$$\boxed{\mathcal{H}_1 : m_e \neq 0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] \neq \frac{1}{2}.}$$

Remarque 1.1.1. La formulation de ce test est bien sûr la formulation d'un test bilatéral. Nous pouvons envisager les deux tests unilatéraux correspondants. À ce moment là, la formulation de l'hypothèse alternative $\mathcal{H}_1$ est différente et s'écrit soit :

[1] Les références [10], [13] et [8] ayant servi à l'élaboration de ce document sont mentionnées dans la bibliographie.

$$\boxed{\mathcal{H}_1' : \mathbb{P}\left[X_i > 0\right] < \frac{1}{2}}$$

soit

$$\boxed{\mathcal{H}_1'' : \mathbb{P}\left[X_i > 0\right] > \frac{1}{2}.}$$

Remarque 1.1.2. Plus généralement ce test permet de tester l'hypothèse nulle

$$\boxed{\mathcal{H}_0 : m_e = m_0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] = p}$$

contre

$$\boxed{\mathcal{H}_1 : m_e \neq m_0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] \neq p.}$$

où m_0 est un nombre réel et p est une constante comprise entre 0 et 1,
ou encore, dans la version unilatérale, contre l'hypothèse alternative

$$\boxed{\mathcal{H}_1' : m_e < m_0}$$

ou encore, dans la version unilatérale, contre l'hypothèse alternative

$$\boxed{\mathcal{H}_1'' : m_e > m_0}$$

Pour cela il suffit de considérer l'échantillon $Z_1, \ldots, Z_n$ avec $Z_i = X_i - m_0$ et de lui appliquer le test décrit ci-dessous.

Statistique : S_n désigne le nombre de variables X_i, $1 \leqslant i \leqslant n$, qui prennent une valeur positive.

Propriétés 1.1.1. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire S_n suit exactement une loi binomiale $\mathcal{B}(n, p)$ de paramètres n et p.*

Concrètement cette hypothèse nulle $\mathcal{H}_0$ signifie que l'effectif de l'échantillon considéré est faible devant celui de la population dont il est issu.

Remarque 1.1.3. Nous pourrons prendre comme taille limite des échantillons dont les effectifs sont inférieurs à une fraction de 1/10 de la population. Dans ce cas nous pouvons assimiler les tirages réalisés ici à des tirages avec remise.

Cas le plus souvent utilisé : p =1/2. Nous nous proposons de tester :

Hypothèses :

$$\boxed{\mathcal{H}_0 : \mathbb{P}\left[X_i > 0\right] = \frac{1}{2}}$$

contre

$$\boxed{\mathcal{H}_1 : \mathbb{P}\left[X_i > 0\right] \neq \frac{1}{2}.}$$

Statistique : S_n désigne le nombre de variables X_i, $1 \leqslant i \leqslant n$, qui prennent une valeur positive.

Propriétés 1.1.2. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire S_n a les trois propriétés suivantes :*

1. *La variable aléatoire S_n suit une loi binomiale $\mathcal{B}(n, 1/2)$ de paramètres n et $1/2$. De ce fait, découle les deux propriétés suivantes :*
2. $\mathbb{E}[S_n] = n/2$.
3. $\mathrm{Var}[S_n] = n/4$.

Cette distribution binomiale est symétrique. Pour n grand ($n \geqslant 40$), nous pouvons utiliser l'approximation normale avec correction de continuité :

$$\mathbb{P}_{\mathcal{H}_0}[S_n \leqslant h] = \mathbb{P}_{\mathcal{H}_0}[S_n \geqslant n - h] = \frac{\Phi(2h + 1 - n)}{\sqrt{n}}$$

où Φ est la fonction de répartition d'une loi normale centrée réduite.

Décision 1.1.1. *Pour un seuil α donné ($= \alpha = 5\% = 0,05$ en général), nous cherchons le plus grand entier $s_\alpha^\star$ tel que $\mathbb{P}[Y \leqslant s_\alpha^\star] \leqslant \alpha/2$ où Y suit une loi binomiale $\mathcal{B}(n, 1/2)$ de paramètres n et $1/2$. Alors nous décidons :*

$$\begin{cases} \mathcal{H}_1 \textit{ est vraie si} & S_{n,obs} \notin]s_\alpha^\star, n - s_\alpha^\star[\\ \mathcal{H}_0 \textit{ est vraie si} & S_{n,obs} \in]s_\alpha^\star, n - s_\alpha^\star[. \end{cases}$$

Remarque 1.1.4. Le niveau de signification réel du test est alors égal à $2\mathbb{P}[Y \leqslant s_\alpha^\star]$ qui est généralement différent de α.

Remarque 1.1.5. Pour voir un exemple, nous renvoyons à la feuille de travaux dirigés qui sera traitée lors de la première séance de travaux dirigés.

1.1.2. Test des rangs signés de Wilcoxon

Soit un échantillon indépendant et identiquement distribué $X_1, \ldots, X_n$ d'une loi continue F dont la valeur médiane est notée m_e et la moyenne μ.

Hypothèses : Le test des rangs signés permet de tester l'hypothèse nulle

$$\boxed{\mathcal{H}_0 : \text{La loi continue } F \text{ est symétrique en } 0}$$

contre

$$\boxed{\mathcal{H}_1 : \text{La loi continue } F \text{ n'est pas symétrique en } 0.}$$

De plus, si nous savons que la loi continue F est symétrique, alors le test des rangs signés de Wilcoxon devient

$$\boxed{\mathcal{H}_0 : \mu = \mu_0}$$

contre

$$\boxed{\mathcal{H}_1 : \mu \neq \mu_0.}$$

Ici μ_0 est un nombre réel et ce jeu d'hypothèses permet alors de s'intéresser à la moyenne de la loi continue F.

1.1.2.1. Cas où il n'y a pas d'ex æquo.

Soit $x_1, \ldots, x_n$ n réalisations de l'échantillon précédent. À chaque x_i nous attribuons le rang r_i^a qui correspond au rang de $|x_i|$ lorsque que les n réalisations sont classées par ordre croissant de leurs valeurs absolues.

Statistique : Nous déterminons alors la somme w des rangs r_i^a des seules observations positives. La statistique W_n^+ des rangs signés de Wilcoxon est la variable aléatoire qui prend pour valeur la somme w. Par conséquent, la statistique W_n^+ des rangs signés de Wilcoxon s'écrit

$$W_n^+ = \sum_{\substack{1 \leqslant i \leqslant n \\ X_i > 0}} R_i^a.$$

Propriétés 1.1.3. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire W_n^+ a les trois propriétés suivantes :*

1. *W_n^+ est symétrique autour de sa valeur moyenne $\mathbb{E}[W_n^+] = n(n+1)/4$.*
2. $\mathrm{Var}[W_n^+] = n(n+1)(2n+1)/24$.
3. *Elle est tabulée pour de faibles valeurs de n. Pour $n \geqslant 15$, nous avons l'approximation normale avec correction de continuité :*

$$\mathbb{P}\left[W_n^+ \leqslant w\right] = \Phi\left(\frac{w + 0,5 - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Décision 1.1.2.

– **Premier cas :** *Pour tester l'hypothèse nulle « $\mathcal{H}_0$: La loi continue F est symétrique en 0 » contre l'hypothèse alternative « $\mathcal{H}_1$: La loi continue F n'est pas symétrique en 0 » pour un seuil donné α, nous cherchons l'entier w_α tel que $\mathbb{P}[W_n^+ \leqslant w_\alpha] \approx \alpha/2$. Alors nous décidons :*

$$\begin{cases} \mathcal{H}_1 \text{ est vraie si} & W_{n,obs}^+ \notin]w_\alpha + 1, n(n+1)/2 - w_\alpha - 1[, \\ \mathcal{H}_0 \text{ est vraie si} & W_{n,obs}^+ \in]w_\alpha + 1, n(n+1)/2 - w_\alpha - 1[. \end{cases}$$

– **Second cas :** *Pour tester l'hypothèse nulle « $\mathcal{H}_0 : \mu = \mu_0$ », nous introduisons l'échantillon $Z_1, \ldots, Z_n$ avec $Z_i = X_i - \mu$, $1 \leqslant i \leqslant n$.*

1.1.2.2. Cas où il y a des ex æquo.

Les observations $x_1, \ldots, x_n$ peuvent présenter des ex æquo et *a fortiori* leurs valeurs absolues. Il s'agit en particulier du cas où la loi F est discrète. Deux procédures sont alors employées.

- *Méthode de départition des ex æquo*

  Nous départageons les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales nous associons un entier au hasard puis nous affectons, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et nous pouvons directement appliquer les résultats du paragraphe précédent.

- *Méthode des rangs moyens*

  En associant à l'observation X_i son rang moyen $R_i^{a\star}$ dans le classement des valeurs absolues et en sommant tous les rangs pour lesquels $X_i > 0$ nous obtenons la statistique :

$$W_n^{+\star} = \sum_{\substack{1 \leqslant i \leqslant n \\ X_i > 0}} R_i^{a\star}.$$

  Les valeurs absolues observées $|x_1|, \ldots, |x_n|$ étant ordonnées puis regroupées en classes d'ex æquo, C_0 pour la première classe qui est constituée des nombres $|x_i|$ nuls, s'il en existe, et C_j, $1 \leqslant j \leqslant h$ pour les autres nombres, certaines classes C_j pouvant comporter un seul élément, si cet élément n'a pas d'ex æquo, notons d_j le nombre d'ex æquo de la classe C_j. Nous avons

$$d_0 + \sum_{j=1}^{h} d_j = n.$$

  Sous l'hypothèse nulle $\mathcal{H}_0$ et si $n > 15$, il est d'usage d'utiliser l'approximation normale

$$\frac{W_n^{+\star} - m^\star}{\sigma^\star} \approx \mathcal{N}(0,1)$$

  où

$$m^\star = \frac{1}{4}\left(n(n+1) - d_0(d_0+1)\right)$$

  et

$$(\sigma^\star)^2 = \frac{1}{24}\left(n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)\right) - \frac{1}{48}\sum_{j=1}^{h}\left(d_j^3 - d_j\right).$$

  Dans le cas où nous utilisons cette méthode des rangs moyens, nous ne pouvons pas utiliser les tables statistiques usuelles qui concernent la distribution de la variable aléatoire W_n^+.

1.2. Les tests non paramétriques sur deux échantillons

1.2.1. Les échantillons sont indépendants : Test de Mann-Whitney

Nous observons, de manière indépendante, une variable Y, continue, sur deux populations, ou sur une population divisée en deux sous-populations. Nous notons $\mathcal{L}_i$ la loi de Y sur la (sous-)population d'ordre i. Nous allons présenter le test :

Hypothèses :

$$\boxed{\mathcal{H}_0 : \text{Les deux lois } \mathcal{L}_i \text{ sont égales ou encore de façon équivalente : } \mathcal{L}_1 = \mathcal{L}_2}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les deux lois } \mathcal{L}_i \text{ ne sont pas égales ou encore de façon équivalente : } \mathcal{L}_1 \neq \mathcal{L}_2.}$$

1.2.1.1. Cas où il n'y a pas d'ex aequo.

Statistique : Pour obtenir la statistique du test notée U en général, nous devons procéder à des étapes successives :

1. En se plaçant sous l'hypothèse nulle $\mathcal{H}_0$, nous classons par ordre croissant l'ensemble des observations des deux échantillons $(x_1, \ldots, x_{n_1})$ et $(y_1, \ldots, y_{n_2})$ de taille respective n_1 et n_2.
2. Nous affectons le rang correspondant.
3. Nous effectuons la somme des rangs pour chacun des deux échantillons, notés R_1 et R_2.
4. Nous en déduisons les quantités U_1 et U_2 qui se calculent ainsi :

$$U_{n_1} = n_1 \times n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{1.2.1}$$

et

$$U_{n_2} = n_1 \times n_2 + \frac{n_2(n_2+1)}{2} - R_2 = n_1 \times n_2 - U_1. \tag{1.2.2}$$

La plus petite des deux valeurs U_{n_1} et U_{n_2}, notée U_{n_1,n_2}, est utilisée pour tester l'hypothèse nulle $\mathcal{H}_0$.

Propriétés 1.2.1. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire U_{n_1,n_2} a les trois propriétés suivantes :*

1. $\mathbb{E}\left[U_{n_1,n_2}\right] = (n_1 \times n_2)/2$.
2. $\mathrm{Var}\left[U_{n_1,n_2}\right] = (n_1 \times n_2)(n_1 + n_2 + 1)/12$.

3. *La variable aléatoire U_{n_1,n_2} est tabulée pour de faibles valeurs de n. Pour $n \geqslant 20$, nous avons l'approximation normale :*

$$\mathbb{P}\left[U_{n_1,n_2} \leqslant u\right] = \Phi\left(\frac{u - (n_1 \times n_2)/2}{\sqrt{(n_1 \times n_2)(n_1 + n_2 + 1)/12}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Décision 1.2.1.

– **Premier cas :** *Si les tailles n_1 ou n_2 sont inférieures à 20. Pour un seuil donné α ($= 5\% = 0,05$ en général), la table de Mann-Whitney nous fournit une valeur critique c. Alors nous décidons :*

$$\begin{cases} \mathcal{H}_1 \text{ est vraie si} & U_{n_1,n_2,obs} \leqslant c, \\ \mathcal{H}_0 \text{ est vraie si} & U_{n_1,n_2,obs} > c. \end{cases}$$

– **Second cas :** *Si les tailles n_1 et n_2 sont supérieures à 20, alors la quantité est décrite approximativement par une loi normale et nous utilisons alors le test de l'écart réduit :*

$$Z_{n_1,n_2} = \frac{U_{n_1,n_2} - (n_1 \times n_2)/2}{\sqrt{(n_1 \times n_2)(n_1 + n_2 + 1)/12}}.$$

Pour un seuil donné α ($= 5\% = 0,05$ en général), la table de la loi normale centrée réduite nous fournit une valeur critique c. Alors nous décidons :

$$\begin{cases} \mathcal{H}_1 \text{ est vraie si} & Z_{n_1,n_2,obs} \geqslant c, \\ \mathcal{H}_0 \text{ est vraie si} & Z_{n_1,n_2,obs} < c. \end{cases}$$

1.2.1.2. Cas où il y a des ex æquo.

Les observations $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ peuvent présenter des ex æquo. Il s'agit en particulier du cas où les lois F et G dont sont issus les deux échantillons sont discrètes. Deux procédures sont alors employées.

- *Méthode de départition des ex æquo*

  Nous départageons les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales nous associons un entier au hasard puis nous affectons, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et nous pouvons directement appliquer les résultats du paragraphe précédent.

- *Méthode des rangs moyens*

  Les valeurs absolues observées $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ étant ordonnées puis regroupées en h classes d'ex æquo C_j, $1 \leqslant j \leqslant h$, certaines classes C_j pouvant comporter un seul élément, si cet élément n'a pas d'ex æquo, notons d_j le nombre d'ex æquo de la classe

C_j. Nous avons $\sum_{j=1}^{h} d_j = n_1 + n_2$.
En associant à l'observation X_i son rang moyen $R_i^\star$ dans ce classement et en sommant tous les rangs de tous les X_i, on obtient la statistique :

$$U_{n_1,n_2}^\star = \sum_{i=1}^{n_2} R_i^\star.$$

Sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y ont la même distribution » et pour $n_1 > 15$ et $n_2 > 15$, il est d'usage d'utiliser l'approximation normale

$$\frac{U_{n_1,n_2}^\star - m^\star}{\sigma^\star} \approx \mathcal{N}(0,1)$$

où

$$m^\star = \frac{1}{2}\left(n_1(n_1+n_2+1)\right)$$

et

$$(\sigma^\star)^2 = \frac{1}{12}\left(n_1 n_2(n_1+n_2+1)\right) - \frac{1}{12}\frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)}\sum_{j=1}^{h}\left(d_j^3 - d_j\right).$$

Dans le cas où nous utilisons cette méthode des rangs moyens nous ne pouvons pas utiliser les tables statistiques usuelles qui concernent la distribution de la variable aléatoire U_{n_1,n_2}.

1.2.2. Les échantillons sont indépendants : Test de la médiane de Mood

On considère deux échantillons indépendants $(X_1,\ldots,X_{n_1})$ et $(Y_1,\ldots,Y_{n_2})$.
$(X_1,\ldots,X_{n_1})$ est un échantillon indépendant et identiquement distribué d'une loi continue F et $(Y_1,\ldots,Y_{n_2})$ est un échantillon indépendant et identiquement distribué d'une loi continue G.

Après regroupement des n_1+n_2 valeurs des deux échantillons, $n_1 \times M_N$ est le nombre d'observations X_i qui sont supérieures à la médiane des $N = n_1+n_2$ observations.

Sous l'hypothèse nulle $\mathcal{H}_0$: « Les variables X et Y suivent la même loi continue c'est-à-dire $G = F$ », la variable $n_1 \times M_N$ peut prendre les valeurs $0,1,\ldots,n_1$ selon la distribution hypergéométrique suivante :

$$\mathbb{P}\left[n_1 \times M_N = k\right] = \frac{C_{n_1}^k C_{n_2}^{N/2-k}}{C_N^{N/2}}.$$

Ainsi on a :

$$\mathbb{E}\left[n_1 \times M_N\right] = \frac{n_1(n_1+n_2-\epsilon_N)}{2N}$$
$$\mathrm{Var}\left[n_1 \times M_N\right] = \frac{n_1 n_2(n_1+n_2+1)}{4(n_1+n_2-1+\epsilon_N)(n_1+n_2+1-\epsilon_N)},$$

où $\epsilon_N = 0$ si N est pair et $\epsilon_N = 1$ si N est impair.

Lorsque n_1 et n_2 sont grands, c'est-à-dire $n_1 \geqslant 25$ et $n_2 \geqslant 25$, on utilise l'approximation normale :

$$\frac{n_1 \times M_N - \mathbb{E}\left[n_1 \times M_N\right]}{\sqrt{\mathrm{Var}\left[n_1 \times M_N\right]}} \approx \mathcal{N}(0,1)$$

avec correction de continuité.

La distribution est symétrique lorsque N est pair.

Pour tester l'hypothèse nulle $\mathcal{H}_0$: « $G = F$ » contre $\mathcal{H}_1$: « $G \neq F$ » avec un niveau de signification égal à α, on cherche les entiers k_α et $k_\alpha^{'}$ tels que $\mathbb{P}\left[n_1 \times M_N \leqslant k_\alpha\right] \approx \alpha/2$ et $\mathbb{P}\left[n_1 \times M_N \geqslant n_1 - k_\alpha^{'}\right] \approx \alpha/2$, puis on rejette l'hypothèse nulle $\mathcal{H}_0$ si la réalisation de la statistique du test calculée à l'aide de l'échantillon n'est pas dans l'intervalle $[k_\alpha, k_\alpha^{'}]$. Cette statistique permet également de réaliser des tests unilatéraux.

1.2.3. Les échantillons sont dépendants : Test de Wilcoxon

Nous considérons deux variables aléatoires X et Y, de même nature, observées toutes les deux sur les mêmes unités d'un n-échantillon. Les observations se présentent alors sous la forme d'une suite de couples $(x_1, y_1), \ldots, (x_n, y_n)$. Ce test concerne les lois des deux variables. Pour ce faire nous testons :

Hypothèses :

$$\boxed{\mathcal{H}_0 : \text{Les deux lois sont égales ou encore de façon équivalente } \mathcal{L}(X) = \mathcal{L}(Y)}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les deux lois ne sont pas égales ou encore de façon équivalente } \mathcal{L}(X) \neq \mathcal{L}(Y).}$$

1.2.3.1. Cas où il n'y a pas d'ex aequo.

Statistique : Pour obtenir la statistique du test notée S^+ en général, nous devons procéder à des étapes successives :

1. Ce test suppose que la loi de la différence entre les deux variables étudiées est symétrique par rapport à 0.

2. Après avoir calculé les différences d_i, nous classons par ordre croissant les $|d_i|$ non nulles, c'est-à-dire les d_i sans tenir compte des signes.
3. Nous attribuons à chaque $|d_i|$ le rang correspondant.
4. Nous restituons ensuite à chaque rang le signe de la différence correspondante.
5. Enfin, nous calculons la somme S^+ des rangs positifs (P) et la somme S^- des rangs négatifs (M).

La somme S^+ des rangs positifs (P) permet de tester l'hypothèse nulle $\mathcal{H}_0$.

Décision 1.2.2.
- **Premier cas :** *Si $n < 15$, nous utilisons une table et nous comparons la valeur de (S^+) à la valeur critique c associée au seuil α du test.*
- **Second cas :** *Si $n \geqslant 15$, nous utilisons l'approximation normale avec correction de continuité :*

$$\mathbb{P}_{\mathcal{H}_0}\left[S^+ \leqslant h\right] \approx \Phi\left(\frac{h + 0,5 - n(n+1)/4}{\sqrt{(n(n+1)(2n+1))/24}}\right)$$

où Φ est la fonction de répartition d'un loi normale centrée réduite.

1.2.3.2. Cas où il y a des ex æquo.

Il se traite de la même manière que pour la statistique de Wilcoxon pour un échantillon, voir le paragraphe 1.1.2.

1.3. Les tests non paramétriques sur k échantillons : 1 facteur

1.3.1. Les échantillons sont indépendants : Test de Kruskal-Wallis

On suppose que l'on dispose de k échantillons **indépendants** et identiquement distribués $(X_{1,1}, \ldots, X_{1,n_1}), \ldots, (X_{k,1}, \ldots, X_{k,n_k})$. La distribution du $i-$ème échantillon est notée F_i. On admet **a priori** que $F_i(x) = G(x - \alpha_i)$ où G est une fonction de répartition inconnue mais continue de moyenne μ et les α_i sont des nombres réels. Ainsi on suppose que le seul paramètre qui diffère d'une distribution F_i à l'autre est un paramètre de position α_i. C'est pourquoi même lorsque vous effectuez un test de Kruskal-Wallis vous devez vous assurer que vous pouvez au moins supposer que les variances des variables sont égales d'un échantillon à l'autre à l'aide d'un test non paramétrique de Levene d'égalité des variances. Les hypothèses ci-dessus impliquent que l'on peut écrire, pour tout $1 \leqslant i \leqslant k$ la décomposition suivante :

$$X_{i,j} = \mu + \alpha_i + \epsilon_{i,j}, \quad 1 \leqslant j \leqslant n_i,$$

les $N = \sum_{i=1}^{k} n_i$ variables aléatoires $\epsilon_{i,j}$ étant indépendantes et ayant une même distribution inconnue et de moyenne nulle.

1.3.1.1. Cas où il n'y a pas d'ex æquo.

La variable KW_N de Kruskal-Wallis est utilisée pour tester l'hypothèse

$$\boxed{\mathcal{H}_0 : \alpha_1 = \ldots = \alpha_k = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Il existe au moins un } i_0 \text{ tel que } \alpha_{i_0} \neq 0.}$$

On commence par calculer le rang $R_{i,j}$ de $X_{i,j}$ parmi les N valeurs, puis la somme des rangs associée à chaque échantillon : $R_{i,\bullet} = \sum_{j=1}^{n_i} R_{i,j}$ et enfin la moyenne des rangs de chaque échantillon : $\overline{R_{i,\bullet}} = \dfrac{R_{i,\bullet}}{n_i}$.
La statistique de Krukal-Wallis KW_N prend en compte l'écart entre la moyenne des rangs de chaque échantillon et la moyenne de tous les rangs, qui vaut $(N+1)/2$:

$$KW_N = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\overline{R_{i,\bullet}} - \frac{N+1}{2} \right).$$

Sous l'hypothèse nulle $\mathcal{H}_0$: « $X_1, \ldots, X_k$ ont la même distribution continue », qui dans notre cas est équivalente à $\mathcal{H}_0$: « $\alpha_1 = \ldots = \alpha_k = 0$ », il est possible de déterminer la distribution de KW_N bien que le calcul soit complexe.
– Si l'un des effectifs n_i, $1 \leqslant i \leqslant k$, est inférieur ou égal à 4, on utilise une table spécifique.
– Si $n_i \geqslant 5$, pout tout $1 \leqslant i \leqslant k$ on utilise l'approximation $KW_N \approx \chi^2_{k-1}$.
Pour un seuil de signification de α, on détermine c_α tel que $\mathbb{P}\left[KW_N \geqslant c_\alpha\right] \cong \alpha$ et l'on rejette l'hypothèse nulle $\mathcal{H}_0$ lorsque la valeur prise par KW_N est supérieure à c_α.

1.3.1.2. Cas où il y a des ex æquo.

• *Méthode de départition des ex æquo*

On départage les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales on associe un entier au hasard puis on affecte, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et on peut directement appliquer les résultats du paragraphe précédent.

• *Méthode des rangs moyens*

À chaque nombre appartenant à un groupe d'ex æquo on attribue le rang moyen du groupe auquel il appartient puis on détermine la somme $T = \sum_{l=1}^{h}(t_l^3 - t_l)$ où t_l désigne le nombre d'éléments du $l-$ème groupe d'ex æquo. Il est d'usage de substituer à KW_N la variable $KW_N^\star$ définie par :

$$KW_N^\star = \frac{KW_N}{1 - \dfrac{T}{N^3 - N}}.$$

Comparaisons multiples

Si l'on rejette l'hypothèse nulle $\mathcal{H}_0$: « $\alpha_1 = \ldots = \alpha_k = 0$ » d'absence de différence entre les distributions F_i des k échantillons, on peut être amené à se demander quelles sont les distributions qui sont différentes.

On décide que **deux distributions** F_i et $F_{i'}$ sont significativement différentes au seuil $\boldsymbol{\alpha}$ si :

$$|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}| \geqslant \sqrt{\chi^2(k-1, 1-\alpha)}\sqrt{\frac{N(N+1)}{12}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}},$$

où $\chi^2(k-1, 1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi du χ^2 à $k-1$ degrés de liberté.

On décide qu'**au seuil global** $\boldsymbol{\alpha}$ deux distributions F_i et $F_{i'}$, parmi les $\boldsymbol{k(k-1)}$ **comparaisons** que l'on va faire, sont significativement différentes si :

$$|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}| \geqslant u\left(1 - \frac{\alpha}{k(k-1)}\right)\sqrt{\frac{N(N+1)}{12}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}},$$

où $u\left(1 - \dfrac{\alpha}{k(k-1)}\right)$ est le $100\left(1 - \dfrac{\alpha}{k(k-1)}\right)$ quantile de la loi normale centrée réduite. Il s'agit d'une application des inégalités de Bonferroni[2]. Cette procédure est plus puissante que la précédente.

On décide qu'**au seuil global** $\boldsymbol{\alpha}$ deux distributions F_i et $F_{i'}$, parmi les $\boldsymbol{k(k-1)}$ **comparaisons** que l'on va faire, sont significativement différentes si :

[2]On pourra consulter le cours sur les modèles d'analyse de la variance pour plus de détails sur les procédures de comparaisons multiples.

$$\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \geqslant q(k, +\infty, 1-\alpha)\sqrt{\frac{N(N+1)}{12}}\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)},$$

où $q(k, +\infty, 1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi de l'étendue studentisée pour k moyennes et $+\infty$ degrés de liberté. Il s'agit d'une procédure analogue à celle de Tukey-Kramer[2] dans le cas paramétrique et valide asymptotiquement. Elle est généralement plus puissante que les deux approches précédentes.

1.3.2. Les échantillons sont indépendants : Test de Jonckheere-Terpstra

La statistique J_N de Jonckheere-Terpstra permet de raffiner l'approche de la statistique KW_N de Kruskal-Wallis : supposons que les k modalités du facteur pour lequel on a réalisé les expériences soient naturellement ordonnées. C'est par exemple le cas dans la situation suivante : vous souhaitez trouver la dose optimale d'engrais à utiliser pour améliorer un rendement. Vous allez donc réaliser des expériences avec des doses de plus en plus importantes d'engrais et les modalités de votre facteur explicatif seront donc naturellement ordonnées par la quantité croissante d'engrais utilisé.

La statistique J_N de Jonckheere-Terpstra permet de tester l'hypothèse :

$$\boxed{\mathcal{H}_0 : \alpha_1 = \ldots = \alpha_k = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \alpha_1 \leqslant \ldots \leqslant \alpha_k = 0 \text{ et il existe au moins un } i_0 \text{ tel que } \alpha_{i_0} < \alpha_{i_0+1}.}$$

1.3.2.1. Cas où il n'y a pas d'ex æquo.

La statistique J_N est construite à l'aide de toutes les variables de Mann-Whitney $U_{i,j}$, associées à l'échantillon i et l'échantillon j, lorsque $1 \leqslant i < j \leqslant k$:

$$J_N = \sum_{1 \leqslant i < j \leqslant k} U_{i,j}.$$

Sous l'hypothèse nulle $\mathcal{H}_0$: « $\alpha_1 = \ldots = \alpha_k = 0$ » :
– L'espérance et la variance de la statistique J_N sont :

$$\begin{aligned}
\mathbb{E}[J_N] &= \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4}, \\
\mathrm{Var}[J_N] &= \frac{1}{72}\left(N^2(3+2N) - \sum_{i=1}^{k} n_i^2(3+2n_i)\right).
\end{aligned}$$

- Les valeurs critiques de la statistique J_N sont tabulées pour de faibles valeurs de k et des n_i.
- Lorsque $n_i \geqslant 5$, pour tout $1 \leqslant i \leqslant k$, on a l'approximation normale avec correction de continuité :

$$\frac{J_N - \mathbb{E}\left[J_n\right]}{\sqrt{\mathrm{Var}\left[J_N\right]}} \approx \mathcal{N}\left(0,1\right).$$

On cherche l'entier ϕ_α tel que $\mathbb{P}\left[J_N \geqslant \phi_\alpha\right] \approx \alpha$ puis on rejette l'hypothèse nulle $\mathcal{H}_0$ au seuil α si la valeur prise par la statistique J_N est supérieure ou égale à ϕ_α.

1.3.2.2. Cas où il y a des ex æquo.

On peut utiliser une méthode de départition des ex æquo ou des tests de Mann-Whitney basés sur des rangs moyens, l'inconvient de la seconde méthode étant qu'on ne peut utiliser les mêmes tables qu'en absence d'ex æquo.

1.3.3. Les échantillons ne sont pas indépendants : Test de Friedman

On se place ici dans le cas où les échantillons utilisés pour tester l'influence d'un facteur **ne sont pas indépendants**.

Individu	Facteur A		
	1	$\cdots$	n
1	$x_{1,1}$	$\cdots$	$x_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1,k}$	$\cdots$	$x_{n,k}$

On construit alors le tableau des rangs :

Individu	Facteur A			Total
	1	$\cdots$	n	
1	$r_{1,1}$	$\cdots$	$r_{n,1}$	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$n(n+1)/2$
k	$r_{1,k}$	$\cdots$	$r_{n,k}$	$n(n+1)/2$
Total	$r_{1,\bullet}$	$\cdots$	$r_{n,\bullet}$	$kn(n+1)/2$

Si on est en présence de répétitions $x_{i,j,k}$ on remplace $x_{i,j}$ par la moyenne $\overline{x_{i,j}}$ des valeurs pour chaque cas où il y a des répétitions.

On cherche alors à tester l'hypothèse :

$\boxed{\mathcal{H}_0 : \text{Les niveaux du facteur ont tous la même influence}}$

contre

$\boxed{\mathcal{H}_1 : \text{Les niveaux du facteur n'ont pas tous la même influence.}}$

1.3.3.1. Cas où il n'y a pas d'ex æquo.

La statistique de Friedman $F_{k,n}$ est définie par :

$$F_{k,n} = \frac{12k}{n(n+1)} \sum_{i=1}^{n} \left(\frac{R_{i,\bullet}}{k} - \frac{n+1}{2} \right)^2 = \frac{12}{kn(n+1)} \sum_{i=1}^{n} R_{i,\bullet}^2 - 3k(n+1).$$

On admet que sous l'hypothèse nulle $\mathcal{H}_0$: « Les niveaux du facteur ont tous la même influence » les distributions pour chaque individu ne diffèrent que par un paramètre de position, ce que l'on peut vérifier par un test non paramétrique de Levene par exemple.

- Pour de petites valeurs de k on utilise une table spécifique. Il se peut que l'on vous fournisse une table du coefficient de concordance $W_{k,n}$ de Kendall car la statistique de Friedman $F_{k,n} = k(n-1)W_{k,n}$.
- Pour des valeurs de k assez grandes on utilise l'approximation asymptotique suivante :

$$F_{k,n} \approx \chi^2_{n-1}.$$

On rejettera l'hypothèse nulle $\mathcal{H}_0$ si la valeur prise par $F_{k,n}$ est trop grande.

1.3.3.2. Cas où il y a des ex æquo.

• *Méthode de départition des ex æquo*

On départage les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales on associe un entier au hasard puis on affecte, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et on peut directement appliquer les résultats du paragraphe précédent.

• *Méthode des rangs moyens*

Dans chaque classement présentant des ex æquo on attribue à chacun de ceux-ci le rang moyen du groupe d'ex æquo auquel il appartient et qui n'est pas nécessairement un entier. Lorsque le classement numéro m a h_m groupes d'ex æquo, on lui attribue la somme $T_m = \sum_{l=1}^{h_m} \left(t_{l,m}^3 - t_{l,m} \right)$ où $t_{l,m}$ désigne le nombre d'éléments du $l-$ème de ces h_m groupes. S'il n'y a pas d'ex æquo on a évidemment $T_m = 0$ puisque la répartition des n entiers du

classement en classes de nombres égaux donne $h_m = n$ et $t_{l,m} = 1$ pour tout l. Alors la statistique de Friedman corrigée est définie par :

$$\begin{aligned} F^{\star}_{k,n} &= \frac{12k(n-1)}{(n^3-n) - \frac{1}{k}\sum_{m=1}^{k} T_m} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2} \right)^2 \\ &= \frac{1}{1 - \frac{1}{(n^3-n)} \frac{1}{k} \sum_{m=1}^{k} T_m} \frac{12k}{n(n+1)} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2} \right)^2 . \end{aligned}$$

On en déduit que :

$$F^{\star}_{k,n} = \frac{F_{k,n}}{1 - \frac{1}{(n^3-n)} \frac{1}{k} \sum_{m=1}^{k} T_m} .$$

1.4. Les tests non paramétriques sur nk échantillons : 2 facteurs

1.4.1. Les échantillons sont indépendants : Test de Friedman

On se place ici dans le cas où les échantillons utilisés pour tester l'influence d'un facteur **sont indépendants**.

Facteur B	Facteur A		
	1	$\cdots$	n
1	$x_{1,1}$	$\cdots$	$x_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1,k}$	$\cdots$	$x_{n,k}$

On construit alors le tableau des rangs :

Facteur B	Facteur A			Total
	1	$\cdots$	n	
1	$r_{1,1}$	$\cdots$	$r_{n,1}$	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$n(n+1)/2$
k	$r_{1,k}$	$\cdots$	$r_{n,k}$	$n(n+1)/2$
Total	$r_{1,\bullet}$	$\cdots$	$r_{n,\bullet}$	$kn(n+1)/2$

Si on est en présence de répétitions $x_{i,j,k}$ on remplace $x_{i,j}$ par la moyenne $\overline{x_{i,j}}$ des valeurs pour chaque cas où il y a des répétitions.

On admet **a priori** que l'influence des couples de niveaux (A_i, B_j) des facteurs A et B, pour $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant k$, se traduit par une décomposition de la forme :

$$X_{i,j} = \mu + \alpha_i + \beta_j + \epsilon_{i,j}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant k$$

avec $\sum_{i=1}^{n} \alpha_i = 0$ et $\sum_{j=1}^{k} \beta_j = 0$. Les $N = \sum_{i=1}^{n} k = n \times k$ variables aléatoires $\epsilon_{i,j}$ étant indépendantes et ayant une même distribution inconnue et de moyenne nulle.

1.4.1.1. Cas où il n'y a pas d'ex æquo.

La variable $F_{k,n}$ de Friedman est utilisée pour tester l'hypothèse

$$\boxed{\mathcal{H}_0 : \alpha_1 = \ldots = \alpha_n = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Il existe au moins un } i_0 \text{ tel que } \alpha_{i_0} \neq 0.}$$

On commence par calculer le rang $R_{i,j}$ de $X_{i,j}$ parmi les n valeurs de la colonne i, puis la somme des rangs associée à chaque colonne : $R_{i,\bullet} = \sum_{j=1}^{k} R_{i,j}$ et enfin la moyenne des rangs de chaque colonne : $\overline{R_{i,\bullet}} = \dfrac{R_{i,\bullet}}{k}$.

La statistique de Friedman $F_{k,n}$ est définie par :

$$F_{k,n} = \frac{12k}{n(n+1)} \sum_{i=1}^{n} \left(\frac{R_{i,\bullet}}{k} - \frac{n+1}{2} \right)^2 = \frac{12}{kn(n+1)} \sum_{i=1}^{n} R_{i,\bullet}^2 - 3k(n+1).$$

- Pour de petites valeurs de k on utilise une table spécifique. Il se peut que l'on vous fournisse une table du coefficient de concordance $W_{k,n}$ de Kendall car $F_{k,n} = k(n-1)W_{k,n}$.
- Pour des valeurs de k assez grandes on utilise l'approximation asymptotique suivante :

$$F_{k,n} \approx \chi^2_{n-1}.$$

On rejettera l'hypothèse nulle $\mathcal{H}_0$ si la valeur prise par la statistique de Friedman $F_{k,n}$ est trop grande.
Si l'on voulait également tester l'influence du facteur B on aurait analysé les tableaux ci-dessous avec la même méthode.

Facteur A	Facteur B		
	1	$\cdots$	n
1	$x_{1,1}$	$\cdots$	$x_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1,k}$	$\cdots$	$x_{n,k}$

Facteur A	Facteur B			Total
	1	$\cdots$	n	
1	$r_{1,1}$	$\cdots$	$r_{n,1}$	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$n(n+1)/2$
k	$r_{1,k}$	$\cdots$	$r_{n,k}$	$n(n+1)/2$
Total	$r_{1,\bullet}$	$\cdots$	$r_{n,\bullet}$	$kn(n+1)/2$

Comme on a échangé le rôle du facteur A et du facteur B on teste maintenant :

$$\boxed{\mathcal{H}_0 : \text{Les niveaux du facteur } B \text{ ont tous la même influence}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les niveaux du facteur } B \text{ n'ont pas tous la même influence.}}$$

On ne peut pas tester l'existence d'une interaction par cette méthode puisque le modèle utilisé ne comporte pas de terme d'interaction. Il existe d'autres tests pour étudier l'existence d'une interaction.

1.4.1.2. Cas où il y a des ex æquo.

• *Méthode de départition des ex æquo*

On départage les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales on associe un entier au hasard puis on affecte, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et on peut directement appliquer les résultats du paragraphe précédent.

• *Méthode des rangs moyens*
Dans chaque classement présentant des ex æquo on attribue à chacun de ceux-ci le rang moyen du groupe d'ex æquo auquel il appartient et qui n'est pas nécessairement un entier. Lorsque le classement numéro m a h_m groupes d'ex æquo, on lui attribue la somme $T_m = \sum_{l=1}^{h_m} \left(t_{l,m}^3 - t_{l,m}\right)$ où $t_{l,m}$ désigne le nombre d'éléments du $l-$ème de ces h_m groupes. S'il n'y a pas d'ex æquo on a évidemment $T_m = 0$ puisque la répartition des n entiers du

classement en classes de nombres égaux donne $h_m = n$ et $t_{l,m} = 1$ pour tout l. Alors la statistique de Friedman corrigée est définie par :

$$\begin{aligned} F^{\star}_{k,n} &= \frac{12k(n-1)}{\left(n^3-n\right) - \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} T_m} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2} \right)^2 \\ &= \frac{1}{1 - \dfrac{1}{(n^3-n)} \dfrac{1}{k} \displaystyle\sum_{m=1}^{k} T_m} \frac{12k}{n(n+1)} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2} \right)^2 . \end{aligned}$$

On en déduit que :

$$F^{\star}_{k,n} = \frac{F_{k,n}}{1 - \dfrac{1}{(n^3-n)} \dfrac{1}{k} \displaystyle\sum_{m=1}^{k} T_m} .$$

Chapitre 2

Valeurs non représentatives.[1]

2.1. Valeurs extrêmes, valeurs non représentatives

La plupart des distributions statistiques ont la majeure partie de leurs réalisations comprises dans un intervalle d'une largeur de 6 écarts types autour de la moyenne μ. Par exemple dans le cas de la loi normale de paramètres μ et σ, la probabilité d'obtenir une valeur dans un intervalle $[\mu - 3\sigma, \mu + 3\sigma]$ est de 99,8 %. Il ne faut pas déduire de cette propriété que l'on n'observera jamais de réalisations se trouvant en dehors de cet intervalle dans la pratique. En effet, bien que la soit probabilité de se trouver face à une telle situation soit faible, elle n'est pas nulle. De surcroît, plus l'effectif de l'échantillon est important, plus ce cas de figure est susceptible de se produire.

En effet, comme vous le savez, en notant X une variable aléatoire d'espérance μ et d'écart type σ et n le nombre de réalisations de X que l'on considère, la fréquence théorique de l'évènement considéré est

$$n \times (1 - \mathbb{P}[-3\sigma \leqslant X - \mu \leqslant +3\sigma]).$$

En considérant le cas d'un loi normale de moyenne $\mu = 0$ et d'écart type $\sigma = 1$, on obtient :

$$n \times (1 - \mathbb{P}[-3 \leqslant X \leqslant +3]) = n \times 0,00270.$$

Ainsi pour un effectif de 400 la fréquence attendue est de $1,080$.

Plus généralement, l'inégalité de Bienaymé-Tchebychev, valable pour toute variable aléatoire Y admettant une variance σ_Y^2, et par conséquent une moyenne μ_Y, permet d'obtenir

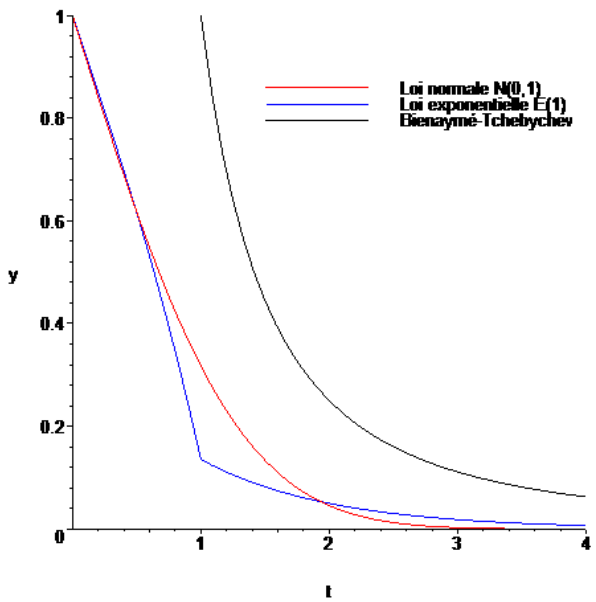
[1] Les références [12], [14] ayant servi à l'élaboration de ce document sont mentionnées dans la bibliographie.

la relation non paramétrique suivante :

$$\Phi_Y(\lambda) \;=\; 1 - \mathbb{P}\left[\mu - \lambda\sigma < Y < \mu + \lambda\sigma\right] \;\leqslant\; \frac{1}{\lambda^2}.$$

Ainsi pour toute variable aléatoire Y dont la loi est une distribution de probabilité admettant une variance l'évènement $\mathbb{P}\left[|Y-\mu| \geqslant 2\sigma\right]$ a au plus une probabilité de $1/4 = 0,25$ et l'évènement $\mathbb{P}\left[|Y-\mu| \geqslant 3\sigma\right]$ a au plus une probabilité de $1/9 \approx 0,11$. On note que cette estimation est trop pessimiste pour un emploi pratique et est donc exclusivement réservé au cas on l'on ne connaît rien sur la loi de Y.

Exemple 2.1.1.



Dans le graphique ci-contre, la valeur de la probabilité Φ_Y est réprésentée sur l'axe des ordonnées en fonction de la valeur de λ qui varie selon l'axe des abscisses.

Trois courbes ont été tracées, l'une associée à l'inégalité de Bienaymé-Tchebychev, les deux autres étant les valeurs exactes des probabilités Φ_Y pour une loi normale centrée-réduite $\mathcal{N}(0,1)$ et pour une loi exponentielle de moyenne 1, $\mathcal{E}(1)$.

2.2. Que vérifier ?

Lorsque que vous rencontrez ce type de valeurs dans un échantillon vous devez avoir la réaction suivante :

- L'hypothèse qui est faite sur la loi de ma variable aléatoire est-elle fondée ? Une alternative non paramétrique ou robuste n'est-elle pas préférable dans ma situation ? Certains des phénomènes que vous étudiez sont connus pour avoir des modélisations qui ont déjà été étudiées. Vous devez alors vous référer à la bibliographie pour déterminer si la modélisation que l'on fait est en accord avec les connaissances a priori que l'on a du phénomène.
- Si l'on utilise un modèle statistique, les autres hypothèses sous-jacentes au modèle sont-elles toutes vérifiées ?
- Les données sont-elles fiables ? Si vous les avez récoltées, les conditions expérimentales étaient-elles semblables à celles que vous avez fixées ou observées lors des autres essais ?
- Ne s'agit-il pas d'une simple erreur de copie et donc alors d'une valeur aberrante ?

2.3. Que faire avec une valeur potentiellement non-représentative ?

Si vous soupçonnez une valeur d'être une valeur non représentative, il existe plusieurs procédures de tests possibles. Attention, il ne faut pas abuser de ces pratiques et ne les utiliser que pour des cas légitimes. De plus si vous soupçonnez plusieurs valeurs d'être des valeurs non représentatives, on verra par la suite comment il faut procéder : il ne s'agit pas simplement de répéter le traitement d'une seule valeur non reprséentative plusieurs fois de suite. Outre les problèmes d'évaluation du risque de première espèce, la présence d'autre valeurs non représentatives peut fausser les résultats des tests.

2.3.1. Notations

Soit $\boldsymbol{x} = (x_1, \ldots, x_n)$ un $n-$échantillon d'une variable aléatoire X.
On note $x_{(1)}, \ldots, x_{(n)}$ les valeurs $x_1, \ldots, x_n$ **ordonnées** dans l'ordre croissant ($x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$). On note $\boldsymbol{x}_{(\bullet)}$ ce nouvel échantillon.
Le classement des valeurs de $\boldsymbol{x}$ ne change ni la valeur de la moyenne de l'échantillon, notée $\bar{\boldsymbol{x}}$, ni celle de l'écart type de l'échantillon, noté $s(\boldsymbol{x})$. En effet on montre que :

$$\begin{aligned}
\bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_{(i)} = \bar{\boldsymbol{x}}_{(\bullet)}, \\
s(\boldsymbol{x}) &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{(i)} - \bar{\boldsymbol{x}}_{(\bullet)}\right)^2 = s(\boldsymbol{x}_{(\bullet)}).
\end{aligned}$$

On appelle **variation** d'un échantillon $\boldsymbol{x}$ la quantité $\mathrm{VA}(\boldsymbol{x})$:

$$\begin{aligned}
\mathrm{VA}(\boldsymbol{x}) &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= (n-1)\, s^2(\boldsymbol{x}) = (n-1)\, s^2(\boldsymbol{x}_{(\bullet)}) = \mathrm{VA}(\boldsymbol{x}_{(\bullet)}).
\end{aligned}$$

La variation n'est pas affectée par le classement des valeurs dans l'ordre croissant.

On appelle **somme des carrés** d'un échantillon $\boldsymbol{x}$ la quantité $\mathrm{SC}(\boldsymbol{x})$:

$$\mathrm{SC}(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_{(i)}^2 = \mathrm{SC}(\boldsymbol{x}_{(\bullet)}).$$

La somme des carrés n'est pas affectée par le classement des valeurs dans l'ordre croissant.

On considèrera **systématiquement** dans ce cours que X suit une **loi normale**. Il est possible de réaliser les mêmes types de tests pour des variables aléatoires qui suivent des

lois autres que la loi normale ; il faut alors utiliser d'autres valeurs critiques que celles qui vous ont été fournies.

2.3.2. Test de Grubbs pour une seule valeur non représentative

On construit les trois statistiques suivantes :

$$T = \max\left(T_1, T_n\right),$$

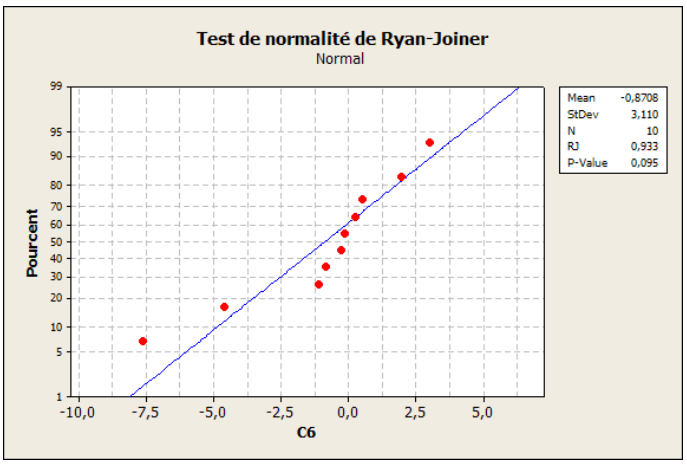
avec

$$T_1 = \frac{(\bar{x} - x_{(1)})}{s},$$
$$T_n = \frac{(x_{(n)} - \bar{x})}{s},$$

où $\bar{x}$ est la moyenne de l'échantillon $\boldsymbol{x} = (x_1, \ldots, x_n)$ et s est son écart type, donc corrigé. On utilise T_1 ou T_n si l'on suspecte la présence d'une valeur non représentative dans une direction donnée et T si l'on n'a pas d'idée a priori sur la direction dans laquelle il pourrait y avoir une valeur non représentative. En effet T_1 mesure la déviation de la plus petite valeur de l'échantillon par rapport à la moyenne standardisée par l'écart type corrigé de l'échantillon. De même T_n mesure la déviation de la plus grande valeur de l'échantillon par rapport à la moyenne standardisée par l'écart type corrigé de l'échantillon.

On compare alors la valeur de T_1, T_n ou T avec la valeur de référence correspondante de la table adéquate (Table de Grubbs). Ainsi pour T_1 et T_n on prendra la valeur du $(1-\alpha)$ % quantile et pour T celle du $(1-\alpha/2)$ % quantile.

Exemple 2.3.1.



x_i	$x_{(i)}$
0,26787	-7,61567
3,01367	-4,60385
-0,27047	-1,11060
-7,61567	-0,82072
-4,60385	-0,27047
0,54445	-0,10821
-0,10821	0,26787
1,99539	0,54445
-1,11060	1,99539
-0,82072	3,01367

$T_1 = 2,85558$ et $T_{10} = 0,79458$ donc $T = 2,85558$. La valeur de T est supérieure à celle de la table pour $\alpha = 5$ % (2,290). On a donc détecté une valeur non représentative. On en déduit que $x_{(1)} = -7,61567$ ou que $x_{(10)} = 3,01367$ est une valeur non représentative avec un risque de première espèce de 5 %. On aurait pu procéder uniquement au test de $x_{(1)}$ ou respectivement de $x_{(10)}$ à l'aide de la statistique T_1 ou respectivement de T_{10}. Pour un risque de première espèce de 5 %, la valeur critique à utiliser avec les statistiques T_1 et T_{10} est de 2,176. On en déduit que $x_{(1)} = -7,61567$ est une valeur non représentative avec un risque de première espèce de 5 %.

.....

2.3.3. Test de Dixon pour une seule valeur non représentative

La statistique du test de Dixon pour détecter une valeur non représentative parmi les grandes valeurs de l'échantillon, $r_{1,0}$, est le rapport entre l'écart entre les deux observations les plus grandes et l'étendue de l'échantillon.

$$r_{1,0} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}.$$

La statistique du test de Dixon pour détecter une valeur non représentative parmi les petites valeurs de l'échantillon, $r'_{1,0}$, est le rapport entre l'écart entre les deux observations les plus petites et l'étendue de l'échantillon.

$$r'_{1,0} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

En posant $r = \max(r_{1,0}, r'_{1,0})$, on obtient une statistique de test, r, permettant de tester la présence d'une valeur non repésentative dans l'une des deux directions. On compare alors la valeur de $r_{1,0}, r'_{1,0}$ ou r avec la valeur de référence, au niveau α, de la table.

On note qu'il s'agit du test le plus simple à réaliser en pratique puisqu'il ne nécessite que peu de calculs.

Exemple 2.3.2.
On reprend les données de l'exemple ci-dessus. On a alors :

$$\begin{aligned} x_{(1)} &= -7,61567, \\ x_{(2)} &= -4,60385, \\ x_{(9)} &= 1,99539, \\ x_{(10)} &= 3,01367, \\ x_{(10)} - x_{(1)} &= 10,62934. \end{aligned}$$

On calcule :

$$\begin{aligned} r_{1,0} &= 0,09580, \\ r'_{1,0} &= 0,28335, \\ r &= 0,28335. \end{aligned}$$

La valeur critique, pour $\alpha = 5$ % et $n = 10$, est de 0,412 pour $r_{1,0}$ et $r'_{1,0}$. Celle pour r est égale à 0,469.
On constate qu'aucune des réalisations des statistiques de test n'est supérieure à la valeur critique au seuil $\alpha = 5$ % qui lui est associée. Les tests ne sont donc pas significatifs. Ainsi pour ces procédures de tests il n'y a pas de valeurs non représentatives.
Cette conclusion n'est pas la même que celle du test de Grubbs. Afin de savoir quel crédit on peut lui accorder on devrait calculer le risque de deuxième espèce associé à une telle décision. Malheusement sa détermination n'est pas aisée et on ne pourra le faire ici.

.....

On note que le test de Dixon proposé ci-avant ne parvient pas à détecter ni la valeur non-représentative $x_{(10)}$ de la partie supérieure de l'échantillon, ni la valeur non-représentative $x_{(1)}$ de la partie inférieure de l'échantillon. Ceci peut être dû à un effet masquant lié aux valeurs de $x_{(1)}$, $x_{(2)}$ et $x_{(9)}$.

On introduit alors les statistiques $r_{j,k}$ suivantes qui permettent de ne pas tenir compte des $j-1$ valeurs les plus fortes, $x_{(n-1)}, \ldots, x_{(n-j+1)}$, et des k valeurs les plus faibles de l'échantillon, $x_{(1)}, \ldots, x_{(k+1)}$, lors du test de la valeur $x_{(n)}$. En effet si celles-ci sont elles aussi non représentatives, cela peut fausser le résultat du test.

$$r_{j,k} = \frac{x_{(n)} - x_{(n-j)}}{x_{(n)} - x_{(k+1)}}.$$

Symétriquement, on introduit les statistiques $r'_{j,k}$ pour la détermination de la représentativité de $x_{(1)}$ sans tenir compte des valeurs les plus fortes ou les plus faibles.

$$r'_{j,k} = \frac{x_{(j+1)} - x_{(1)}}{x_{(n-k)} - x_{(1)}}.$$

Quelles valeurs choisir pour les valeurs de j et de k ?
Dixon a formulé les recommandations d'utilisation suivantes. Si l'effectif n de l'échantillon est inférieur ou égal à 7, il faut utiliser $j = 1$ et $k = 0$. Si $8 \leqslant n \leqslant 14$, il faut utiliser $j = 2$ et $k = 1$. Enfin si $n \geqslant 15$, il faut utiliser $j = 2$ et $k = 2$.

Exemple 2.3.3.
L'effectif de l'échantillon utilisé dans les exemples ci-dessus est de 10. On calcule donc $r_{2,1}$ et $r'_{2,1}$:

$$\begin{aligned} r_{2,1} &= \frac{x_{(10)} - x_{(8)}}{x_{(10)} - x_{(2)}} \\ &= \frac{3,01367 - 0,54445}{3,01367 - (-4,60385)} = 0,32415, \\ r'_{2,1} &= \frac{x_{(3)} - x_{(1)}}{x_{(9)} - x_{(1)}} \\ &= \frac{-1,11060 - (-7,61567)}{1,99539 - (-7,61567)} = 0,67683. \end{aligned}$$

Les valeurs critiques du test sont : pour $\alpha = 10$ %, 0,551, pour $\alpha = 5$ %, 0,612, et pour $\alpha = 1$ %, 0,726. Ainsi, même au seuil $\alpha = 10$ %, le test basé sur $r_{2,1}$ n'est pas significatif : $x_{(10)}$ n'est pas une valeur non-représentative. Par contre, aux seuils $\alpha = 10$ % et $\alpha = 5$ %, le test basé sur $r'_{2,1}$ est significatif. Ainsi après avoir éliminé l'effet masquant de $x_{(2)}$, on peut à nouveau mettre en évidence la non représentativité de $x_{(1)}$.

.....

L'exemple que l'on vient de traiter pose le problème pratique suivant : en utilisant r ou $r'_{1,0}$ les tests ne sont pas significatifs. Tandis qu'en utilisant $r'_{2,1}$, le test est significatif. Cette situation ne serait-elle pas due au fait que la valeur $x_{(2)}$ n'est pas, elle non plus, représentative. Or si, comme dans cet exemple, l'on soupçonne non pas la présence d'une valeur non représentative mais de plusieurs, il existe d'autres procédures de tests à appliquer qui seront détaillées à la section 2.4 et à la section 2.6.

2.3.4. Test basé sur l'étendue

La statistique du test basé sur l'étendue de l'échantillon est :

$$u = \frac{x_{(n)} - x_{(1)}}{s}.$$

où l'on rappelle que l'étendue e d'un échantillon est la longueur de l'intervalle séparant la plus petite valeur de la plus grande valeur de l'échantillon, entre d'autres termes, $e = x_{(n)} - x_{(1)}$. Ce test permet de détecter une valeur non représentative dans l'une quelconque des directions, c'est-à-dire à la fois une valeur trop faible ou trop forte. Il est néanmoins plus naturel de s'en sevir pour tester la paire de valeurs potentiellement non représentative $(x_{(1)}, x_{(n)})$ puisque l'on compare l'étendue e de l'échantillon à son écart type s.

Exemple 2.3.4.
En reprenant les données ci-dessus, on calcule u pour cet exemple :

$$u = \frac{3,01367 - (-7,61567)}{3,10974} = 3,41808.$$

Les valeurs critiques pour $\alpha = 5$ % et $\alpha = 10$ % sont respectivement 3,68 et 3,57. Ainsi à aucun de ces niveaux le test n'est significatif. La paire $(x_{(1)}, x_{(n)})$ n'est pas constituée de valeurs toutes les deux non représentatives, ce qui est bien cohérent avec les résultats précédents.

.....

2.4. Test simultané de k valeurs non représentatives

Le nombre k est ici fixé avant la procédure de test. Le cas où k serait également à déterminer sera abordé aux sections 2.5 et 2.6.

2.4.1. Test de Grubbs pour k valeurs non représentatives dans une direction donnée.

La statistique exposée ci-après a été proposée par Grubbs en 1950 pour $k = 2$ puis étendue par Tietjen and Moore en 1972 au cas $1 < k < n$. À la fois la queue de la distribution (valeurs fortes ou valeurs faibles) dans laquelle on procède au test et le nombre de valeurs testées k doivent être fixés à l'avance. On note L_k les statistiques associées à la détection de valeurs non représentatives parmi les fortes valeurs de l'échantillon et $L_k^\star$ celles associées à la recherche parmi les faibles valeurs de l'échantillon.

$$L_k = \frac{\displaystyle\sum_{i=1}^{n-k} \left(x_{(i)} - \bar{x}_{n-k}\right)^2}{(n-1)s^2}$$

où $\bar{x}_{n-k}$ est la moyenne des $n-k$ plus petites valeurs de l'échantillon, c'est-à-dire $\bar{x}_{n-k} = \dfrac{1}{n-k}\displaystyle\sum_{i=1}^{n-k} x_{(i)}$ et s^2 est toujours la variance corrigée de l'échantillon de taille n.

$$L_k^\star = \frac{\displaystyle\sum_{i=k+1}^{n} \left(x_{(i)} - \bar{x}_{n-k}^\star\right)^2}{(n-1)s^2}$$

où $\bar{x}_{n-k}^\star$ est la moyenne des $n-k$ plus grandes valeurs de l'échantillon, c'est-à-dire $\bar{x}_{n-k}^\star = \dfrac{1}{n-k}\displaystyle\sum_{i=k+1}^{n} x_{(i)}$.

Ainsi dans chacun des deux cas on fait le rapport de la somme des carrés des déviations à la moyenne de l'échantillon privé des k valeurs que l'on suspecte d'être non représentatives par la somme des carrés des déviations à la moyenne de l'échantillon tout entier. Ces

rapports sont toujours inférieurs ou égaux à 1. Une valeur proche de 0 indiquera que la présence des valeurs testées augmente de manière conséquente la variation de l'échantillon. On comparera la réalisation du test au quantile à α %.

Exemple 2.4.1.
On reprend toujours les mêmes données. Les résultats des tests de Dixon ont confirmé le fait que $x_{(1)}$ est une valeur non représentative et ont amené à envisager que $x_{(2)}$ en était aussi une. On cherche à tester deux valeurs, donc $k = 2$, dans la partie inférieure de l'échantillon et on calcule ainsi $L_2^\star$.

$$\begin{aligned}
\bar{x}_8^\star &= \frac{1}{8}\sum_{i=3}^{10} x_{(i)} = 0,43892 \\
L_2^\star &= \frac{\displaystyle\sum_{i=3}^{10}\left(x_{(i)} - \bar{x}_8^\star\right)^2}{9s^2} \\
&= \frac{13,8826}{87,0345} = 0,15951.
\end{aligned}$$

On compare ce résultat avec les valeurs critiques (associées aux quantiles inférieurs de la distribution cette fois-ci) pour un niveau de $\alpha = 1\%$ et de $\alpha = 5\%$. Après lecture dans la table on trouve respectivement $0,142$ et $0,233$. Puisque $L_2^\star = 0,160 < 0,233$ le test est significatif au niveau $\alpha = 5\%$. Il ne l'est pas au niveau $\alpha = 1\%$ car $L_2^\star = 0,160 > 0,142$. Ainsi avec un risque de première espèce de 5%, on peut conclure que les deux valeurs $x_{(1)} = -7,61567$ et $x_{(2)} = -4,60385$ sont non-représentatives.

.....

2.4.2. Test de Grubbs pour deux valeurs non représentatives, une de chaque côté

On utilise la même idée que celle sur laquelle repose les statistiques de test L_k et $L_k^\star$. On compare la variation totale de l'échantillon où l'on a retiré les deux valeurs $x_{(1)}$ et $x_{(n)}$ à celle de l'échantillon complet en faisant le rapport suivant :

$$\frac{S_{1,n}^2}{S^2} = \frac{\displaystyle\sum_{i=2}^{n-1}\left(x_{(i)} - \bar{x}_{1,n}\right)^2}{(n-1)s^2}$$

où $\bar{x}_{1,n}$ est la moyenne des $n-2$ valeurs centrales de l'échantillon initial, c'est-à-dire $\bar{x}_{1,n} = \dfrac{1}{n-2}\displaystyle\sum_{i=2}^{n-1} x_{(i)}$.

Exemple 2.4.2.
On continue avec les mêmes données :

$$\begin{aligned}
\bar{x}_{1,10} &= \frac{1}{8}\sum_{i=2}^{9} x_{(i)} = -0,513268 \\
\frac{S_{1,n}^2}{S^2} &= \frac{\sum_{i=2}^{9}\left(x_{(i)} - \bar{x}_{1,10}\right)^2}{9s^2} \\
&= \frac{25,4295}{87,0345} = 0,292177.
\end{aligned}$$

La valeur critique pour $\alpha = 10$ % est de 0,246. Le test n'est donc pas significatif à ce niveau. Le minimum et le maximum de l'échantillon ne sont pas tous les deux des valeurs non représentatives.

.....

Il est bien entendu possible de généraliser cette approche au cas où l'on considérerait $2k$ valeurs, k de chaque côté de l'échantillon. Malheureusement vous ne disposez pas des valeurs critiques associées à ces tests.

2.4.3. Test de Tietjen-Moore pour k valeurs non représentatives de l'un ou des deux côtés

Là encore on reprend l'idée qui a permis de construire L_k. On construit r l'échantillon formé des valeurs absolues des déviations par rapport à la moyenne :

$$r_i = |x_i - \bar{x}|.$$

Puis on classe cet échantillon : $r_{(1)}, \ldots, r_{(n)}$. En conservant l'ordre ainsi obtenu on multiplie chaque $r_{(i)}$ par le signe de $x_{(i)} - \bar{x}$ et on note $z_{(i)}$ les valeurs ainsi construites. Les statistiques E_k sont alors :

$$E_k = \frac{\sum_{i=1}^{n-k}\left(z_{(i)} - \bar{z}_{n-k}\right)^2}{\sum_{i=1}^{n}\left(z_{(i)} - \bar{z}\right)^2},$$

où $\bar{z}$ est la moyenne de $z_{(1)}, \ldots, z_{(n)}$ et $\bar{z}_{n-k}$ est la moyenne de $z_{(1)}, \ldots, z_{(n-k)}$.
Les $z_{(i)}$ sont simplement les déviations par rapport à la moyenne $\bar{z}$ ordonnées par valeur absolue croissante.

Exemple 2.4.3.
Toujours avec le même exemple :

x_i	$x_{(i)}$	r_i	$r_{(i)}$	$z_{(i)}$
0,26787	-7,61567	1,13868	0,05009	0,05009
3,01367	-4,60385	3,88448	0,23979	-0,23979
-0,27047	-1,11060	0,60034	0,60034	0,60034
-7,61567	-0,82072	6,74486	0,76260	0,76260
-4,60385	-0,27047	3,73304	1,13868	1,13868
0,54445	-0,10821	1,41526	1,41526	1,41526
-0,10821	0,26787	0,76260	2,86620	2,86620
1,99539	0,54445	2,86620	3,73304	-3,73304
-1,11060	1,99539	0,23979	3,88448	3,88448
-0,82072	3,01367	0,05009	6,74486	-6,74486

$$\begin{aligned}
\bar{z}_9 &= 0,749428, \\
E_1 &= \frac{\sum_{i=1}^{9}\left(z_{(i)}-\bar{z}_9\right)^2}{\sum_1^{10}\left(z_{(i)}-\bar{z}\right)^2} \\
&= \frac{36,4867}{87,0345} = 0,41922. \\
\bar{z}_8 &= 0,357546, \\
E_2 &= \frac{\sum_{i=1}^{8}\left(z_{(i)}-\bar{z}_8\right)^2}{\sum_1^{10}\left(z_{(i)}-\bar{z}\right)^2} \\
&= \frac{25,4295}{87,0345} = 0,29218. \\
\bar{z}_7 &= 0,941915, \\
E_3 &= \frac{\sum_{i=1}^{7}\left(z_{(i)}-\bar{z}_7\right)^2}{\sum_1^{10}\left(z_{(i)}-\bar{z}\right)^2} \\
&= \frac{6,30625}{87,0345} = 0,072457.
\end{aligned}$$

pour $\alpha = 1$ % et dans l'ordre $k = 2, 3, 4$, les valeurs critiques sont, 0,235, 0,101, 0,048. Ainsi aucun des tests n'est significatif. Au seuil $\alpha = 5$ %, on a, toujours dans le même ordre, les valeurs critiques 0,356, 0,172, 0,083.

Les choses se compliquent puisque maintenant seul le test basé sur E_3 est significatif. Ceci est assez cohérent avec ce qui précède puisque les deux valeurs les plus faibles de l'échantillon sont associées à la première et à la troisième déviation par rapport à la moyenne les plus fortes en valeurs absolues. En d'autres termes $z_{(i)}$ est associé à $x_{(1)}$ et

$z_{(3)}$ à $x_{(2)}$. Or si l'on avait vu que $x_{(1)}$ et $x_{(2)}$ étaient non représentatives, aucun test jusqu'alors n'avait permis de conclure à la non représentativité de $x_{(10)}$.

.....

2.5. Procédures pour détecter un nombre de valeurs non représentatives non fixé à l'avance

2.5.1. La boîte à moustaches

On base une mesure de la dispersion de l'échantillon sur la longueur de l'intervalle interquartile. En procédant ainsi, on construit un indicateur qui ne dépend que des valeurs centrales de l'échantillon, la moitié du nombre totale d'entre elles pour être précis. De ce fait, l'étendue interquartile est un indicateur robuste de la dispersion de l'échantillon.

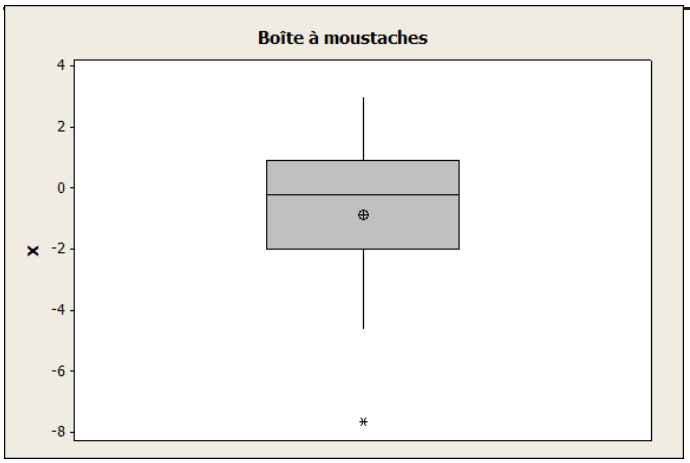
On note Q_1 le premier quartile de l'échantillon et Q_3 le troisième quartile, $IQR = Q_3 - Q_1$ est alors l'étendue interquartile. On dit qu'une valeur x de l'échantillon est une valeur extrême si $x \notin [Q_1 - 3/2IQR, Q_3 + 3/2IQR]$ et qu'il s'agit d'un type particulier de valeur extrême, une valeur très extrême, si $x \notin [Q_1 - 3IQR, Q_3 + 3IQR]$.

On a montré que si l'effectif n de l'échantillon est compris entre 20 et 75 on doit s'attendre à ce que 1 à 2 % des observations soient des valeurs extrêmes. Si l'effectif n est supérieur à 100, alors on doit s'attendre à ce que moins de 1 % des observations soient des valeurs extrêmes, la valeur limite, lorsque n tend vers l'infini convergeant vers $0,7$ %.

En ce qui concerne les valeurs très extrêmes, leur proportion est de moins de 1 % si $6 \leqslant n \leqslant 19$ et de moins de $0,1$ % si $n \geqslant 20$, la valeur limite lorsque n tend vers l'infini convergeant vers une proportion de $0,00023$ %.

L'intérêt de la boîte à moustaches est de proposer un outil visuel d'exploration des données, résistant aux effets d'écrantage car basé sur les valeurs centrales de l'échantillon et pour lequel on n'a pas à faire d'hypothèses sur la localisation ou le nombre de valeurs non représentatives présentes dans l'échantillon. Son principal défaut est qu'il ne permet pas de faire de test et donc de quantifier les risques associés aux décisions qui découlent de son utilisation. En tant que technique exploratoire, la boîte à moustaches permet de choisir des valeurs plausibles pour k et d'utiliser alors la panoplie de tests présentés dans les sections 2.3 et 2.4 précédentes et dans la section 2.6 à venir.

Exemple 2.5.1.
On continue avec les mêmes données.



On constate la présence d'une valeur extrême. Minitab ne distinguant pas les valeurs extrêmes des valeurs très extrêmes on va faire le calcul séparément.

$$\begin{aligned} Q_1 &= -1,98391 \\ Q_3 &= 0,90719 \\ IQR &= 2,89110 \\ Q_1 - 3/2IQR &= -6,32056 \geqslant x_{(1)} \\ Q_1 - 3IQR &= -10,6572 \leqslant x_{(1)} \end{aligned}$$

Ainsi $x_{(1)}$ n'est pas une valeur très extrême mais seulement une valeur extrême. De cette représentation graphique on tirerait l'information $k = 1$. Or on a vu qu'il ne s'agit pas de la seule situation qui soit envisageable.

.....

2.5.2. Test basé sur le coefficient d'asymétrie

La statistique b_1 du test basé sur le coefficient d'asymétrie est :

$$\sqrt{b_1} = \frac{m_3}{m_2^{\frac{3}{2}}},$$

où m_3 est le moment centré d'ordre 3 de la variable étudiée et m_2 est le moment centré d'ordre 2, c'est-à-dire la variance, de la variable étudiée.
Si $\sqrt{b_1} > 0$ les données sont plus localisées au delà de la moyenne. Si $\sqrt{b_1} < 0$ les données sont plus localisées en deçà de la moyenne.
On utilise généralement un test bilatéral sauf si l'on sait que les valeurs non représentatives ne sont présentes que d'un côté de l'échantillon.

Exemple 2.5.2.

$$\sqrt{b_1} = -1,25505$$

On conclut à l'aide des valeurs de la table.

.....

2.5.3. Test basé sur le coefficient d'aplatissement

La statistique b_2 du test basé sur le coefficient d'aplatissement est :

$$b_2 = \frac{m_4}{m_2^2}.$$

On utilise un test unilatéral dont la zone de rejet est située au voisinage de $+\infty$.

Exemple 2.5.3.

$$b_2 = 1,64728$$

On conclut à l'aide des valeurs de la table.

.....

2.6. Procédures séquentielles de détections de valeurs non représentatives

Comme le montre l'exemple 2.4.3, il est difficile de déterminer a priori le nombre de valeurs non représentatives ainsi que leur répartition de chaque côté. On se tourne alors vers des algorithmes itératifs pour détecter lesquelles parmi celles que l'on soupçonne sont des valeurs non représentatives. L'utilisation d'une boîte à moustaches, détaillée à la section 2.5, ou d'une droite de Henry, il s'agit du type de graphique qui apparaît dans l'exemple 2.3.1, permet généralement d'avoir une idée du nombre potentiel, donc maximal, de valeurs non représentatives présentes dans l'échantillon étudié.

2.6.1. Procédure séquentielle de Prescott

L'idée ici est de calculer sucessivement les rapports de sommes de carrés ce qui ressemble à ce que l'on a déjà utilisé pour les tests de Grubbs en retirant les valeurs potentiellement non représentatives l'une après l'autre. On spécifie le nombre maximal de valeurs non représentatives k à détecter ce qui permet de trouver les valeurs $\lambda_j(\beta)$ en utilisant une table. On s'en sert de la manière suivante. On commence par calculer les D_j :

$$D_j = \frac{S_{(j)}^2}{S_{(j-1)}^2} \quad \text{pour } 1, \ldots, k,$$

où $S_{(j)}^2$ est la somme des carrés de l'échantillon privé des j observations les plus éloignées de la moyenne, c'est-à-dire, en concervant les notations du 2.4.3, privé de $z_{(n)}, \ldots, z_{(n-j+1)}$.

Le nombre de valeurs non représentatives m est alors le plus grand entier $m \leqslant k$ tel que :

$$D_j < \lambda_j(\beta).$$

Exemple 2.6.1.
Attention, on calcule ici **la somme des carrés et non la variation**.

$$\begin{aligned}
S^2_{(0)} &= S^2 = 94,61771 \\
S^2_{(1)} &= 36,61928 \\
S^2_{(2)} &= 27,53707 \\
S^2_{(3)} &= 6,34164 \\
D_1 &= 0,38702 \\
D_2 &= 0,75198 \\
D_3 &= 0,23029.
\end{aligned}$$

Les tables disponibles ne contiennent les valeurs que pour un effectif de 10 que si $k = 2$. Alors même au niveau $\alpha = 10$ %, il n'y a pas de valeurs non représentatives, $\lambda_1(\beta) = 0,360$ et $\lambda_2(\beta) = 0,385$.
Par contre, en étudiant les tables pour $k = 2$ et $k = 3$ on se rend compte que même si les valeurs critiques pour $k = 2$ sont plus élévées que pour $k = 3$, elles restent semblables. Ainsi il est fort probable que $D_3 < \lambda_3(\beta)$ au seuil $\alpha = 2,5$ % et assurément au seuil $\alpha = 5$ %. De ce fait on détecte trois valeurs non représentatives sur le maximum de trois que l'on avait fixé ($k = 3$).

.....

2.6.2. Procédure RST de Rosner

À nouveau, on utilise une procédure séquentielle basée sur des statistiques R_i de Grubbs. On choisit à nouveau un nombre k, le nombre maximal de valeurs non repésentatives présentes dans l'échantillon. La différence avec la procédure de Prescott exposée au 2.6.1 est que l'on va utiliser une moyenne et une variance tronquée basée sur l'échantillon dans lequel on a supprimé les k valeurs les plus fortes et les k valeurs les plus faibles.

$$\begin{aligned}
a &= \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}, \\
b^2 &= \frac{1}{n-2k-1} \sum_{i=k+1}^{n-k} \left(x_{(i)} - a\right)^2.
\end{aligned}$$

Maintenant si on note E_0 l'échantillon complet on appelle x^0 l'élément tel que :

$$R_1 = \max_{E_0} \left| \frac{x_i - a}{b} \right| = \left| \frac{x^0 - a}{b} \right|.$$

Puis on considère $E_1 = E_0 - \{x^0\}$ et on cherche x^1 tel que :

$$R_2 = \max_{E_1} \left| \frac{x_i - a}{b} \right| = \left| \frac{x^1 - a}{b} \right|.$$

et ainsi de suite... On remarque que les R_i sont les déviations par rapport à la moyenne tronquée a et normalisées par b classées dans l'ordre décroissant.
On compare alors les valeurs de R_i avec les valeurs critiques qui dépendent de α, n et k, i observations étant non représentatives si R_i est supérieur à la valeur critique.

Exemple 2.6.2.

$$\begin{aligned}
a &= -0,23288 \\
b &= 0,45213 \\
R_1 &= \frac{|-7,61567 - (-0,23288)|}{0,45213} = 16,32894 \\
R_2 &= \frac{|-4,60385 - (-0,23288)|}{0,45213} = 9,66752 \\
R_3 &= \frac{|3,01367 - (-0,23288)|}{0,45213} = 7,18059.
\end{aligned}$$

Les tables disponibles ne sont utilisables que si $n \geqslant 20$.

.....

2.7. Conclusion

Avant tout il s'agit de réaliser les deux représentations graphiques des données que sont la boîte à moustaches et la droite de Henry. Ceci fait, on aura une idée de la situation dans laquelle l'on se trouve.
Même alors, les procédures proposées restent variées. Il serait intéressant de disposer d'études de puissance afin de comparer les différents choix qui s'offrent au praticien au sein d'une même problématique.
Lorsque l'on se demande si une valeur et une seule ($k = 1$) est une valeur non représentative, le test T de Grubbs se comporte toujours le mieux. Si $k > 1$, le mieux est de confronter les résultats des différentes procédures. Si l'on souhaite seulement savoir s'il y a des valeurs non représentatives dans une direction donnée, on utilisera un test sur le coefficient

d'asymétrie. Si au contraire on suspecte la présence de valeurs non représentatives dans chacune des directions, on utilsera un test basé sur le coefficient d'aplatissement. Si l'on cherche à déterminer le nombre de valeurs non représentatives on utilisera une procédure séquentielle.
D'autre part, il est conseillé, par Grubbs, de toujours procéder au test des valeurs non représentatives à un niveau α plus conservatif que le niveau communément utilisé de 5 %, c'est-à-dire avec $\alpha < 0,05$, par exemple $\alpha = 0,01$.

Chapitre 3

Mesures de liaison paramétriques.[1]

3.1. Coefficient de corrélation simple

Le coefficient de corrélation simple $\rho(X,Y)$ mesure le degré d'association linéaire entre deux variables aléatoires X et Y.

$$\rho(X,Y) = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}[X,Y]}{(\mathrm{Var}[X] \times \mathrm{Var}[Y])^{\frac{1}{2}}}.$$

On a toujours $-1 \leqslant \rho(X,Y) \leqslant 1$.

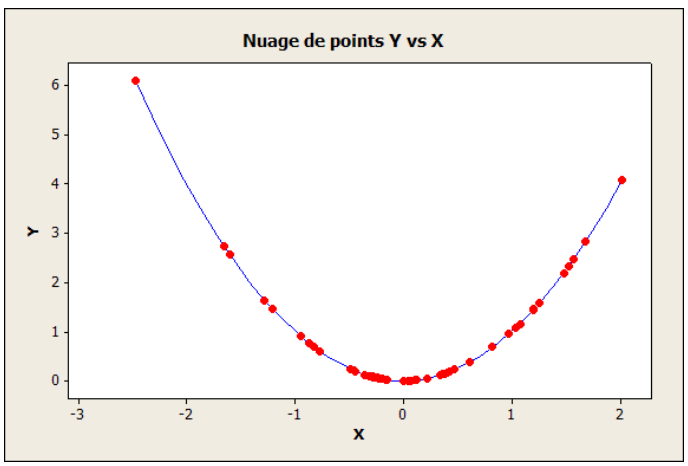
Lorsque $\rho(X,Y) = 0$, on dit que les deux variables aléatoires X et Y sont non corrélées. Attention, ne pas être corrélé ne signifie pas généralement être indépendants. Mais l'indépendance implique toujours l'absence de corrélation.

Exemple 3.1.1.
Prenons $X \sim \mathcal{N}(0,1)$ et $Y = X^2$, on sait alors que $Y \sim \chi^2_1$ une loi du χ^2 à un degré de liberté. Ainsi les propriétés classiques d'une loi du χ^2 impliquent que $\mathbb{E}[Y] = 1$ et $\mathrm{Var}[Y] = 2$. Remarquer que nous ne sommes donc pas dans le cas qui sera développé plus tard où le couple (X,Y) suit une loi normale bidimensionnelle. On a alors :

$$\begin{aligned}
\rho(X,Y) &= \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}[X,Y]}{(\mathrm{Var}[X] \times \mathrm{Var}[Y])^{\frac{1}{2}}} = \frac{\mathbb{E}\left[(X-0)(X^2-1)\right]}{(1 \times 2)^{\frac{1}{2}}} \\
&= \frac{1}{\sqrt{2}}\mathbb{E}\left[X^3 - X\right] = \frac{1}{\sqrt{2}}\left(\mathbb{E}\left[X^3\right] - \mathbb{E}\left[X\right]\right) = 0 - 0 = 0.
\end{aligned}$$

[1] Le cas paramétrique sous hypothèse de multinormalité

Car X suit une loi normale centrée donc symétrique par rapport à 0. Pourtant il est clair que les deux variables aléatoires X et $Y = X^2$ ne sont pas indépendantes !



En bleu la relation entre les deux variables aléatoires X et $Y = X^2$, en rouge 50 réalisations du couple (X, X^2). On calculera au paragraphe 3.2.2 la valeur de la statistique de corrélation à partir de l'échantillon formé des points en rouge.

. . . . . .

Il faut donc toujours garder en mémoire que le coefficient de corrélation ne mesure que le **degré d'association linéaire** entre deux variables aléatoires. Nous verrons dans les sections suivantes quels outils utiliser lorsque l'on suspecte que l'association n'est pas nécessairement linéaire.

Lorsque $|\rho(X, Y)| = 1$ alors, avec une probabilité de 1, $Y = aX + b$ pour des constantes réelles a, du même signe que $\rho(X, Y)$, et b.

Le coefficient de corrélation est invariant par transformation linéaire. Si $\mathrm{Cor}(X, Y) = \rho(X, Y)$ et que l'on pose :

$$\begin{aligned} X' &= aX + b, \\ Y' &= cY + d, \end{aligned}$$

alors $\mathrm{Cor}(X', Y') = \rho(X, Y)$. Ceci a une conséquence extrêmement importante : on peut aussi bien travailler avec les variables brutes qu'avec les variables centrées réduites pour l'étude de la corrélation.

3.2. Le cas bidimensionnel

3.2.1. Loi normale bidimensionnelle

Supposons que l'on dispose d'un vecteur aléatoire (X, Y) gaussien. Nous verrons dans la suite, au moment de tester cette hypothèse, qu'un vecteur aléatoire (X, Y) n'est pas nécéssairement gaussien si X et Y suivent des lois normales[2]. L'hypothèse que l'on fait porte sur le couple et est de ce fait plus délicate à tester que simplement tester la normalité de l'échantillon $\boldsymbol{x}$ issu de X et celle de l'échantillon $\boldsymbol{y}$ issu de Y. Commençons par rappeler ce que l'on entend par la loi d'un vecteur aléatoire gaussien à deux dimensions.

Définition 3.2.1. Loi normale bidimensionnelle
La densité de la loi normale à deux dimensions est :

$$f(x, y) = K \exp\left[-\frac{1}{2}P(x, y)\right]$$

où :

$$\begin{aligned} K &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho(X,Y)^2}}, \\ P(x, y) &= \frac{1}{1-\rho(X,Y)^2}\left[\frac{(x-m_X)^2}{\sigma_X^2} - 2\rho(X,Y)\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right]. \end{aligned}$$

Cette loi dépend donc des cinq paramètres : m_X, m_Y, σ_X, σ_Y et $\rho(X, Y)$.

Les quatre premiers sont les moyennes et les écarts types des deux lois marginales, la loi de X et celle de Y, qui sont des lois normales, $X \sim \mathcal{N}(m_X, \sigma_X^2)$ et $Y \sim \mathcal{N}(m_Y, \sigma_Y^2)$. Les lois liées, c'est-à-dire la probabilité d'obtenir une valeur $Y = y$ sachant que l'on a obtenu $X = x$ ou d'obtenir une valeur $X = x$ sachant que l'on a obtenu une valeur $Y = y$ sont également des lois normales dont les paramètres respectifs $\left(m_{Y|X=x}, \sigma_{Y|X=x}\right)$ et $\left(m_{X|Y=y}, \sigma_{X|Y=y}\right)$ s'expriment ainsi :

$$\begin{aligned} m_{Y|X=x} &= m_Y + \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(x - m_X), \\ \sigma^2_{Y|X=x} &= \sigma_Y^2\left(1-\rho(X,Y)^2\right). \\ m_{X|Y=y} &= m_X + \rho(X,Y)\frac{\sigma_X}{\sigma_Y}(y - m_Y), \\ \sigma^2_{X|Y=y} &= \sigma_X^2\left(1-\rho(X,Y)^2\right). \end{aligned}$$

On remarque que :

[2] Voir l'ouvrage de J.-Y. Ouvrard [11], exercice 13.1 page 276, pour un contre-exemple.

- La variance de chacune des lois liées est constante : elle ne dépend pas de la valeur de X ou de Y prise en compte pour la calculer, c'est-à-dire du nombre x, qui intervient dans $\sigma_{Y|X=x}$, respectivement y, qui intervient dans $\sigma_{X|Y=y}$, pour lequel elle est calculée.
- La moyenne de chacune des lois liées varie linéairement avec la variable liée. On peut donc parler de deux droites de régression :

$$\begin{aligned} D_{YX}: \quad y - m_Y &= \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(x - m_X), \\ D_{XY}: \quad x - m_X &= \rho(X,Y)\frac{\sigma_X}{\sigma_Y}(y - m_Y). \end{aligned}$$

- $\rho(X,Y)$ est le coefficient de corrélation linéaire :

$$\begin{aligned} \rho(X,Y) &= \frac{\mathrm{Cov}\,[X,Y]}{(\mathrm{Var}\,[X] \times \mathrm{Var}\,[Y])^{\frac{1}{2}}} \\ &= \frac{\mathbb{E}\,[(X - m_X) \times (Y - m_Y)]}{\left[\mathbb{E}\left[(X - m_X)^2\right] \times \mathbb{E}\left[(Y - m_Y)^2\right]\right]^{\frac{1}{2}}}. \end{aligned}$$

  $\rho(X,Y) = 0$ est une condition **nécessaire et suffisante** pour que les deux variables aléatoires X et Y soient indépendantes car ici on a une **hypothèse** de normalité bidimensionnelle **supplémentaire**, voir le paragraphe 3.1 pour la situation générale.
  Si $\rho(X,Y) = \pm 1$, les deux droites sont confondues et il y a une relation fonctionnelle entre X et Y. Plus précisement, $(X - m_X)$ et $(Y - m_Y)$ sont colinéaires, c'est-à-dire qu'il existe λ et μ deux nombres réels tels que $\lambda\,(X - m_X) + \mu\,(Y - m_Y) = 0$. On reconnaît dans la formule précédente l'équation d'une droite. Ainsi si $\rho(X,Y) = \pm 1$, les deux variables X et Y sont liées par une relation fonctionnelle linéaire.
  La valeur absolue de $\rho(X,Y)$ est toujours inférieure ou égale à 1. Plus $\rho(X,Y)$ est proche de cette valeur, plus la liaison entre X et Y est serrée. Si l'on considère un échantillon $((x_1,y_1),\ldots,(x_n,y_n))$ de réalisations du couple (X,Y), la dispersion des couples de points (x_i,y_i) autour des droites est d'autant plus faible que les deux droites sont plus proches l'une de l'autre.

- $\rho(X,Y)^2$ est égal au coefficient de détermination R^2 que l'on trouve dans le contexte de la régression linéaire simple.

3.2.2. Estimation des paramètres

On considère un $n-$échantillon $(X_1,Y_1),\ldots,(X_n,Y_n)$ indépendant et identiquement distribué suivant la loi de (X,Y), qui est une loi normale bidimensionnelle.

Les estimateurs des moyennes m_X et m_Y et des écarts types σ_X et σ_Y que nous allons

utiliser sont :

$$\begin{aligned}
\widehat{m_X} &= \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \\
\widehat{m_Y} &= \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \\
\widehat{\sigma_X^2} &= \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2, \\
\widehat{\sigma_Y^2} &= \frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2, \\
\widehat{\sigma_X} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}, \\
\widehat{\sigma_Y} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}.
\end{aligned}$$

Par convention et par soucis d'alléger les notations, on note $\bar{X}$ à la place de $\widehat{m_X}$.
Les quatre premiers estimateurs ci-dessus sont des estimateurs **sans biais** et convergents :

$$\begin{aligned}
\mathbb{E}\left[\widehat{m_X}\right] &= m_X \quad \mathrm{Var}\left[\widehat{m_X}\right] = \frac{\sigma_X^2}{n}, \\
\mathbb{E}\left[\widehat{m_Y}\right] &= m_Y \quad \mathrm{Var}\left[\widehat{m_Y}\right] = \frac{\sigma_Y^2}{n}, \\
\mathbb{E}\left[\widehat{\sigma_X^2}\right] &= \sigma_X^2 \quad \mathrm{Var}\left[\widehat{\sigma_X^2}\right] = \frac{2\sigma_X^4}{n-1}, \\
\mathbb{E}\left[\widehat{\sigma_Y^2}\right] &= \sigma_Y^2 \quad \mathrm{Var}\left[\widehat{\sigma_Y^2}\right] = \frac{2\sigma_Y^4}{n-1}.
\end{aligned}$$

Les deux derniers sont **biaisés** et convergents :

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma_X}\right] &= \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sigma_X = \left(1 - \frac{3}{4n} + o\left(\frac{1}{n}\right)\right)\sigma_X, \\
\mathrm{Var}\left[\widehat{\sigma_X}\right] &= \frac{1}{n-1}\left[n - 1 - \frac{2\Gamma\left(\frac{n}{2}\right)^2}{\Gamma\left(\frac{n-1}{2}\right)^2}\right]\sigma_X^2, \\
\mathbb{E}\left[\widehat{\sigma_Y}\right] &= \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sigma_Y = \left(1 - \frac{3}{4n} + o\left(\frac{1}{n}\right)\right)\sigma_Y,
\end{aligned}$$

$$\mathrm{Var}\left[\widehat{\sigma_Y}\right] = \frac{1}{n-1}\left[n-1-\frac{2\Gamma\left(\frac{n}{2}\right)^2}{\Gamma\left(\frac{n-1}{2}\right)^2}\right]\sigma_Y^2$$

où $o\left(1/n\right)$ est une fonction qui tend vers 0 plus vite que $1/n$ et $\Gamma\left(x\right)$ est une fonction mathématique définie à l'aide d'une intégrale !
Il serait alors possible de corriger le biais[3] de $\widehat{\sigma_X}$. La situation est beaucoup plus compliquée que pour les estimateurs ci-dessus. On comprend ainsi pourquoi on préfère estimer la variance d'une loi normale à la place de son écart type. « Heureusement », les deux paramètres qui ont été choisis pour définir une loi normale sont sa moyenne et sa variance. Au passage on remarque que la racine carrée d'un estimateur sans biais n'est pas nécessairement sans biais.

Le coefficient de corrélation linéaire $\rho(X,Y)$ est un rapport. Pour estimer ce rapport nous allons estimer le numérateur $\mathrm{Cov}\left[X,Y\right]$ et le dénominateur $\left(\mathrm{Var}\left[X\right]\times\mathrm{Var}\left[Y\right]\right)^{\frac{1}{2}}$. En ce qui concerne le dénominateur, nous verrons qu'il suffit d'utiliser les estimateurs $\widehat{\sigma_X}$ et $\widehat{\sigma_Y}$ ci-dessus. Un estimateur du numérateur peut être défini par la quantité :

$$\begin{aligned}\widehat{\mathrm{Cov}\left[X,Y\right]} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)\\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i\times Y_i\right)-\bar{X}\bar{Y}.\end{aligned}$$

C'est un estimateur sans biais et convergent de $\mathrm{Cov}\left[X,Y\right]$:

$$\mathbb{E}\left[\widehat{\mathrm{Cov}\left[X,Y\right]}\right]=\mathrm{Cov}\left[X,Y\right].$$

Notez la similitude avec l'estimateur corrigé de la variance. Ici aussi le dénominateur est $n-1$.

L'estimateur du coefficient de corrélation $\rho(X,Y)$, que nous allons utiliser, est $\widehat{\rho(X,Y)}$

[3] Il suffit de poser $\widehat{\sigma_X}_{\text{corrigé}}=\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\widehat{\sigma_X}$. Néanmoins l'obtention d'intervalle de confiance reste délicate puisque la loi de cet estimateur n'est pas classique. P. Chapouille, dans son livre [4], propose la formule approximative :

$$\widehat{\sigma_X}_{\text{corrigé}}\approx\widehat{\sigma_X}\sqrt{\frac{2n-2}{2n-3}}\approx\widehat{\sigma_X}\left(1+\frac{1}{4n-6}\right).$$

défini par :

$$\begin{aligned}
\widehat{\rho(X,Y)} &= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\times\left(Y_i-\bar{Y}\right)}{\left(\widehat{\sigma_X^2}\times\widehat{\sigma_Y^2}\right)^{\frac{1}{2}}} \\
&= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\times\left(Y_i-\bar{Y}\right)}{\widehat{\sigma_X}\times\widehat{\sigma_Y}} \\
&= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\times\left(Y_i-\bar{Y}\right)}{\left(\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\times\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\right)^{\frac{1}{2}}} \\
&= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\times\left(Y_i-\bar{Y}\right)}{\dfrac{1}{n-1}\left(\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\times\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\right)^{\frac{1}{2}}} \\
&= \frac{\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\times\left(Y_i-\bar{Y}\right)}{\left(\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\times\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\right)^{\frac{1}{2}}} \\
&= \frac{\displaystyle\sum_{i=1}^{n}\left(X_i\times Y_i\right)-\bar{X}\times\bar{Y}}{\left(\displaystyle\sum_{i=1}^{n}X_i^2-\bar{X}^2\right)^{\frac{1}{2}}\left(\displaystyle\sum_{i=1}^{n}Y_i^2-\bar{Y}^2\right)^{\frac{1}{2}}}.
\end{aligned}$$

L'étude des propriétés de cet estimateur est plus complexe puisqu'il s'agit d'un rapport de variables aléatoires qui ne sont pas indépendantes[4]. Il s'agit d'un estimateur **biaisé** et convergent de $\rho(X,Y)$:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\rho(X,Y)}\right] &= \rho(X,Y)-\frac{\rho(X,Y)(1-\rho(X,Y)^2)}{2n} \\
\mathrm{Var}\left[\widehat{\rho(X,Y)}\right] &= \frac{(1-\rho(X,Y)^2)^2}{n}+o\left(\frac{1}{n}\right),
\end{aligned}$$

[4]Voir la note 6

où $o\,(1/n)$ est une fonction qui tend vers 0 lorsque n tend vers $+\infty$ plus vite que $1/n$.
La distribution asymptotique de $\rho(X,Y)$ est :

$$\sqrt{n}(\widehat{\rho(X,Y)} - \rho(X,Y)) \approx \mathcal{N}\left(0, \left(1-\rho(X,Y)^2\right)^2\right).$$

Ce résultat n'est pas assez précis pour permettre de résoudre les problèmes qui nous intéressent généralement. On l'utilisera uniquement si l'on ne peut pas faire autrement et si $n \geqslant 30$.

Ainsi lorsque l'on dispose d'un échantillon $(x_1,y_1),\ldots,(x_n,y_n)$ de réalisations du couple (X,Y) on obtient les estimations des paramètres de la loi normale bidimensionnelle suivie par (X,Y) en utilisant les formules ci-dessous qui sont les réalisations des estimateurs ci-dessus sur notre échantillon. On note $\boldsymbol{x}$ l'échantillon $(x_1,\ldots,x_n)$, $\boldsymbol{y}$ l'échantillon $(y_1,\ldots,y_n)$ et $(\boldsymbol{x},\boldsymbol{y})$ l'échantillon $(x_1,y_1),\ldots,(x_n,y_n)$.

Les estimations des moyennes m_X et m_Y et des écarts types σ_X et σ_Y sont :

$$\begin{aligned}
\widehat{m_X}(\boldsymbol{x}) &= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\\
\widehat{m_Y}(\boldsymbol{y}) &= \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\\
\widehat{\sigma_X^2}(\boldsymbol{x}) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,\\
\widehat{\sigma_Y^2}(\boldsymbol{y}) &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2,\\
\widehat{\sigma_X}(\boldsymbol{x}) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2},\\
\widehat{\sigma_Y}(\boldsymbol{y}) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}.
\end{aligned}$$

On note $\bar{x}$ bien qu'il s'agisse de la moyenne de l'échantillon $\boldsymbol{x}$ et que de ce fait on pourrait également écrire $\bar{\boldsymbol{x}}$. Il s'agit là aussi d'une convention mais si l'on voulait être cohérent avec les notations utilisées pour les autres estimateurs on devrait écrire $\widehat{m_X}(\boldsymbol{x})$.
L'estimation de la covariance $\mathrm{Cov}\,[X,Y]$ est donnée par :

$$\widehat{\mathrm{Cov}\,[X,Y]}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y}).$$

L'estimation du coefficient de corrélation $\rho(X,Y)$ est donnée par :

$$\begin{aligned}
\widehat{\rho(X,Y)}(\boldsymbol{x},\boldsymbol{y}) &= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\left(\widehat{\sigma_X^2}(\boldsymbol{x})\times\widehat{\sigma_Y^2}(\boldsymbol{y})\right)^{\frac{1}{2}}} \\
&= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\widehat{\sigma_X}(\boldsymbol{x})\times\widehat{\sigma_Y}(\boldsymbol{y})} \\
&= \frac{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\left(\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2\times\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{\frac{1}{2}}}. \\
&= \frac{\displaystyle\sum_{i=1}^{n}(x_i\times y_i)-\bar{x}\times\bar{y}}{\left(\displaystyle\sum_{i=1}^{n}x_i^2-\bar{x}^2\right)^{\frac{1}{2}}\times\left(\displaystyle\sum_{i=1}^{n}y_i^2-\bar{y}^2\right)^{\frac{1}{2}}}.
\end{aligned}$$

Exemple 3.2.1.
On évalue la formule ci-dessus pour l'échantillon représenté en rouge au paragraphe 3.1 :

```
Pearson correlation of Y and X = 0,006
```

.....

3.2.3. Procédure de test

Sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes », la variable :

$$\sqrt{n-2}\frac{\widehat{\rho(X,Y)}}{\sqrt{1-\widehat{\rho(X,Y)}^2}}$$

suit une loi de Student T_{n-2} à $n-2$ degrés de liberté.
Ceci permet de réaliser le test suivant :

$$\boxed{\mathcal{H}_0 : \rho(X,Y)=0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho(X,Y)\neq 0.}$$

Pour prendre une décision associée à un test de niveau $(1-\alpha)$ %, il suffit donc de trouver la valeur critique d'une loi de Student à $n-2$ degrés de liberté pour $\alpha/2$.
Il est également possible, en réalisant un test unilatéral, de tester les hypothèses suivantes :

$$\boxed{\mathcal{H}_0 : \rho(X,Y) = 0}$$
contre
$$\boxed{\mathcal{H}_1 : \rho(X,Y) > 0.}$$

$$\boxed{\mathcal{H}_0 : \rho(X,Y) = 0}$$
contre
$$\boxed{\mathcal{H}_1 : \rho(X,Y) < 0.}$$

Notons que, si $\rho_0 \neq 0$, la propriété ci-dessus ne permet pas de tester des hypothèses du type :

$$\boxed{\mathcal{H}_0 : \rho(X,Y) = \rho_0}$$
contre
$$\boxed{\mathcal{H}_1 : \rho(X,Y) \neq \rho_0.}$$

Pour procéder à ces types de test, on utilise la transformation de Fisher[5]. En effet, Fisher, le premier, a démontré que si l'on introduit les deux variables :

$$\begin{aligned}
Z(X,Y) &= \frac{1}{2}\ln\left(\frac{1+\rho(X,Y)}{1-\rho(X,Y)}\right) \\
&= \tanh^{-1}\left(\rho(X,Y)\right), \\
\widehat{Z(X,Y)} &= \frac{1}{2}\ln\left(\frac{1+\widehat{\rho(X,Y)}}{1-\widehat{\rho(X,Y)}}\right) \\
&= \tanh^{-1}\left(\widehat{\rho(X,Y)}\right),
\end{aligned}$$

alors la différence $\widehat{Z(X,Y)} - Z(X,Y)$ suit une loi voisine de la loi normale de moyenne $\rho(X,Y)/(2(n-1))$ et de variance proche de $1/(n-3)$. On approximera donc dans le cas général la loi de cette différence par une loi $\mathcal{N}(\rho(X,Y)/(2(n-1)), 1/(n-3))$ et sous une hypothèse d'indépendance entre X et Y, qui implique donc $\rho(X,Y) = 0$, par une loi

[5]Il existe d'autres transformations pour obtenir des intervalles de confiance pour $\rho(X,Y)$, par exemple l'approximation de Ruben qui semble être plus précise mais encore plus complexe que celle de Fisher. On peut par exemple montrer que, quelque soit la valeur du coefficient de corrélation $\rho(X,Y)$, la variable aléatoire $(R-\rho(X,Y))\sqrt{(n-2)/\left[(1-R^2)(1-\rho(X,Y)^2)\right]}$ suit approximativement une loi $\mathcal{T}_{n-2}$ de Student à $n-2$ degrés de liberté. On retrouve le résultat du paragraphe 3.2.3 établi lorsque $\rho(X,Y) = 0$. Pour plus de détails voir [2] et [6].

$\mathcal{N}(0, 1/(n-3))$.
Pour passer des résultats obtenus pour $Z(X,Y)$ à des résultats concernant $\rho(X,Y)$, on utilise la transformation inverse de celle de Fisher :

$$\begin{aligned}\rho(X,Y) &= \frac{\exp(2Z(X,Y)) - 1}{\exp(2Z(X,Y)) + 1} \\ \rho(X,Y) &= \tanh(Z(X,Y)).\end{aligned}$$

La plupart des logiciels de statistique permettent de calculer les estimations par intervalles de $\rho(X,Y)$. Certains proposent même des intervalles de confiance exacts.

3.2.4. Remarques sur les liaisons

- L'existence de deux droites de régression dans les études de corrélation est embarassante car elle amène à la question suivante : laquelle des deux droites représente-t-elle le mieux la réalité ? On ne peut trancher sans utiliser la notion de causalité : si l'on pense que X est la cause de Y, on retiendra la droite D_{YX} qui est la régression de la variable dépendante par rapport à la variable cause.
  Dans les cas où l'on ne peut établir une relation de cause à effet, il n'y a pas lieu de fixer son attention sur l'une des droites de régression plutôt que sur l'autre. Le coefficient de corrélation linéaire $\rho(X,Y)$ est alors le paramètre le plus intéressant ; quelque soit sa valeur, positive ou négative, il ne préjuge en rien d'un quelconque rapport de cause à effet entre les variables étudiées.
  Dans tous les cas, l'interprétation de toute étude de liaison mérite beaucoup de réflexions, et seules les raisons physiques ou biologiques et non statistiques pourront permettre de porter des jugements de cause à effet.

- Lorsque l'on trouve une courbe de régression qui serre de près le phénomène étudié, il faut se garder de conclure que cette formule traduit la loi exacte qui gouverne le phénomène.

- Enfin signalons que souvent deux variables X et Y sont fortement corrélées avec une troisième Z, le temps par exemple, et on peut conclure à une corrélation significative entre X et Y, alors qu'à priori il n'y a aucune relation entre ces deux grandeurs si ce n'est leur liaison avec Z. On sent alors la nécessité d'introduire une mesure de liaison qui permettra de connaître l'association de X et Y en éliminant l'influence de Z : il s'agit de la notion de corrélation partielle qui sera étudiée aux paragraphes 3.5, 4.3 et 4.5.

3.3. Le cas général ($n \geqslant 2$)

3.3.1. Loi multinormale

On considère un vecteur gaussien de dimension $p \geqslant 1$: $\boldsymbol{X} = (X_1, \ldots, X_p)$. Un tel vecteur est entièrement déterminé par la connaissance de sa moyenne $\boldsymbol{\mu}(\boldsymbol{X})$ et de sa matrice de variance-covariance $\boldsymbol{\Sigma}(\boldsymbol{X})$. En termes plus clairs, il suffit de connaître :

$$\begin{aligned}
\boldsymbol{\mu}(\boldsymbol{X}) &= (\mu_1, \ldots, \mu_p) = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p]), \\
\boldsymbol{\Sigma}(\boldsymbol{X}) &= ((\sigma_{i,j} = \mathrm{Cov}\,[X_i, X_j]))_{1 \leqslant i,j \leqslant p} = \begin{pmatrix} \mathrm{Var}\,[X_1] & \ldots & \mathrm{Cov}\,[X_1, X_p] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}\,[X_p, X_1] & \cdots & \mathrm{Var}\,[X_p] \end{pmatrix}.
\end{aligned}$$

Il s'agit d'une extension de ce que nous venons de voir pour le cas bidimensionnel.
On introduit alors la matrice des corrélations $\boldsymbol{\rho}(\boldsymbol{X})$ entre les composantes X_i, $1 \leqslant i \leqslant p$ de $\boldsymbol{X}$:

$$\begin{aligned}
\boldsymbol{\rho}(\boldsymbol{X}) &= ((\rho_{i,j} = \mathrm{Cor}\,[X_i, X_j]))_{1 \leqslant i,j \leqslant p} \\
&= \begin{pmatrix} 1 & \ldots & \dfrac{\mathrm{Cov}\,[X_1, X_p]}{(\mathrm{Var}\,[X_1]\,\mathrm{Var}\,[X_p])^{\frac{1}{2}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\mathrm{Cov}\,[X_p, X_1]}{(\mathrm{Var}\,[X_p]\,\mathrm{Var}\,[X_1])^{\frac{1}{2}}} & \ldots & 1 \end{pmatrix}.
\end{aligned}$$

3.3.2. Estimation

Afin de construire un estimateur simple de la matrice $\boldsymbol{\rho}(\boldsymbol{X})$ on doit disposer d'un $n-$échantillon $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ indépendant et identiquement distribué du vecteur $\boldsymbol{X}$.
Ce $n-$échantillon est donc composé de n éléments, qui sont des vecteurs de taille p, que l'on note $(X_{1,1}, \ldots, X_{p,1}), \ldots, (X_{1,n}, \ldots, X_{p,n})$. $X_{i,j}$ est la variable aléatoire associée à la valeur prise par la $i-$ème composante X_i lors de la réalisation de la $j-$ème expérience. Par exemple $X_{1,3}$ serait la variable aléatoire associée à la valeur prise par X_1 lors de la troisième expérience et $X_{5,2}$ serait la variable aléatoire associée à la valeur prise par X_5 lors de la seconde expérience.

Un estimateur $\widehat{\boldsymbol{\mu}}$ de la moyenne $\boldsymbol{\mu}$ est :

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= \frac{\boldsymbol{X_1} + \cdots + \boldsymbol{X_n}}{n} \\
&= \frac{\displaystyle\sum_{i=1}^{n} \boldsymbol{X_i}}{n}.
\end{aligned}$$

On note, comme d'habitude, $\widehat{\boldsymbol{\mu}}$ par $\bar{\boldsymbol{\mu}}$. Il s'agit d'un estimateur sans biais et convergent, comme au paragraphe 3.2.2 :

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\mu}}\right] &= \mathbb{E}\left[\frac{\boldsymbol{X_1} + \cdots + \boldsymbol{X_n}}{n}\right] \\
&= \frac{\mathbb{E}\left[\sum_{i=1}^{n} \boldsymbol{X_i}\right]}{n} \\
&= \frac{\sum_{i=1}^{n} \mathbb{E}\left[\boldsymbol{X_i}\right]}{n} \\
&= \frac{\sum_{i=1}^{n} \boldsymbol{\mu}}{n} \\
&= \boldsymbol{\mu}. \\
\mathrm{Var}\left[\widehat{\boldsymbol{\mu}}\right] &= \frac{\boldsymbol{\Sigma}}{n}.
\end{aligned}$$

Un estimateur $\widehat{\boldsymbol{\rho}(\boldsymbol{X})}$ de la matrice des corrélations $\boldsymbol{\rho}(\boldsymbol{X})$ est alors donné par :

$$\begin{aligned}
\widehat{\boldsymbol{\rho}(\boldsymbol{X})} &= \left(\left(\widehat{\rho_{i,j}} = \widehat{\rho_{i,j}} = \widehat{\mathrm{Cor}\left[X_i, X_j\right]}\right)\right)_{1 \leqslant i,j \leqslant p} \\
&= \begin{pmatrix} 1 & \cdots & \dfrac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_{1,i} - \bar{X}_1\right) \times \left(X_{p,i} - \bar{X}_p\right)}{\widehat{\sigma_1} \times \widehat{\sigma_p}} \\ \vdots & \ddots & \vdots \\ \dfrac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_{p,i} - \bar{X}_p\right) \times \left(X_{1,i} - \bar{X}_1\right)}{\widehat{\sigma_p} \times \widehat{\sigma_1}} & \cdots & 1 \end{pmatrix}.
\end{aligned}$$

On obtient alors la forme explicite suivante :

$$\begin{pmatrix} 1 & \cdots & \dfrac{\displaystyle\sum_{i=1}^{n}\left(X_{1,i}-\bar{X}_1\right)\times\left(X_{p,i}-\bar{X}_p\right)}{\left(\displaystyle\sum_{i=1}^{n}\left(X_{1,i}-\bar{X}_1\right)^2\times\sum_{i=1}^{n}\left(X_{i,p}-\bar{X}_p\right)^2\right)^{\frac{1}{2}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\displaystyle\sum_{i=1}^{n}\left(X_{p,i}-\bar{X}_p\right)\times\left(X_{1,i}-\bar{X}_1\right)}{\left(\displaystyle\sum_{i=1}^{n}\left(X_{p,i}-\bar{X}_p\right)^2\times\sum_{i=1}^{n}\left(X_{1,i}-\bar{X}_1\right)^2\right)^{\frac{1}{2}}} & \cdots & 1 \end{pmatrix}.$$

La distribution asymptotique de ρ_{ij} est :

$$\sqrt{n}(\widehat{\rho_{ij}}-\rho_{ij}) \approx \mathcal{N}\left(0,\left(1-\rho_{ij}^2\right)^2\right).$$

Ce résultat n'est pas assez précis pour permettre de résoudre les problèmes qui nous intéressent généralement. On l'utilisera uniquement si l'on ne peut pas faire autrement et si $n \geqslant 30$.

On a donc une estimation $\widehat{\boldsymbol{\rho}(\boldsymbol{X})}(\boldsymbol{x})$ de $\boldsymbol{\rho}(\boldsymbol{X})$ à l'aide d'un échantillon $\boldsymbol{x}$ de n réalisations de $\boldsymbol{X}$. En notant $\boldsymbol{x_1}$ l'échantillon formé par les n réalisations, $(x_{1,1},\ldots,x_{1,n})$, de $X_1,\ldots$, $\boldsymbol{x_p}$ l'échantillon formé par les n réalisations, $(x_{p,1},\ldots,x_{p,n})$, de X_p on a :

$$\begin{aligned}\widehat{\boldsymbol{\rho}(\boldsymbol{X})}(\boldsymbol{x}) &= \left(\left(\widehat{\rho(\boldsymbol{X})}(\boldsymbol{x})_{i,j}=\widehat{\rho_{i,j}}(\boldsymbol{x_i},\boldsymbol{x_j})=\widehat{\mathrm{Cor}\,[X_i,X_j]}(\boldsymbol{x_i},\boldsymbol{x_j})\right)\right)_{1\leqslant i,j\leqslant p} \\ &= \begin{pmatrix} 1 & \cdots & \dfrac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_{1,i}-\bar{x_1})\times(x_{p,i}-\bar{x_p})}{\widehat{\sigma_1}(\boldsymbol{x_1})\times\widehat{\sigma_p}(\boldsymbol{x_p})} \\ \vdots & \ddots & \vdots \\ \dfrac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_{p,i}-\bar{x_p})\times(x_{1,i}-\bar{x_1})}{\widehat{\sigma_p}(\boldsymbol{x_p})\times\widehat{\sigma_1}(\boldsymbol{x_1})} & \cdots & 1 \end{pmatrix}.\end{aligned}$$

L'estimation $\widehat{\boldsymbol{\rho}}(\boldsymbol{X})(\boldsymbol{x})$ de $\boldsymbol{\rho}(\boldsymbol{X})$ est alors :

$$\begin{pmatrix} 1 & \cdots & \dfrac{\displaystyle\sum_{i=1}^{n}(x_{1,i}-\bar{x_1})\times(x_{p,i}-\bar{x_p})}{\left(\displaystyle\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)^2\times\sum_{i=1}^{n}(x_{i,p}-\bar{x}_p)^2\right)^{\frac{1}{2}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\displaystyle\sum_{i=1}^{n}(x_{p,i}-\bar{x_p})\times(x_{1,i}-\bar{x_1})}{\left(\displaystyle\sum_{i=1}^{n}(x_{p,i}-\bar{x}_p)^2\times\sum_{i=1}^{n}(x_{i,1}-\bar{x}_1)^2\right)^{\frac{1}{2}}} & \cdots & 1 \end{pmatrix}.$$

3.3.3. Test de l'hypothèse $\rho_{ij} = 0$

On peut maintenant s'intéresser au test d'indépendance des composantes d'un vecteur gaussien.

$$\boxed{\mathcal{H}_0 : \rho_{ij} = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{ij} \neq 0.}$$

Sous l'hypothèse nulle $\mathcal{H}_0$, alors :

$$t_{ij} = \sqrt{n-2}\frac{\widehat{\rho_{ij}}(\boldsymbol{x_i}, \boldsymbol{x_j})}{\sqrt{1-\widehat{\rho_{ij}}(\boldsymbol{x_i}, \boldsymbol{x_j})^2}}$$

est la réalisation d'une variable aléatoire T qui suit une loi de Student à $n-2$ degrés de liberté, i.e. $T \sim \mathcal{T}_{n-2}$.
On rejette l'hypothèse nulle $\mathcal{H}_0$ au niveau $(1-\alpha)$ % lorsque $|t_{ij}| > \mathcal{T}_{n-2,\alpha/2}$.

3.3.4. Test de l'hypothèse $\rho_{ij} = \rho_0$, $\rho_0 \neq 0$

Si $\rho_0 \neq 0$, la propriété du paragraphe 3.3.3 ci-dessus ne permet pas de tester des hypothèses du type :

$$\boxed{\mathcal{H}_0 : \rho_{ij} = \rho_0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{ij} \neq \rho_0.}$$

.

Il existe alors deux possibilités :

- Si l'on souhaite faire un test de ce type à la main, on utilise la transformation de Fisher introduite ci-dessus, au paragraphe 3.2.3, pour obtenir un intervalle de confiance pour ρ_{ij} de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

  où $\widehat{z} = \tanh\left(\widehat{\rho_{ij}}(\boldsymbol{x_i}, \boldsymbol{x_j})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Certains logiciels, comme SPSS, R , SAS ou StatXact, proposent des versions *exactes* de l'intervalle de confiance dont une approximation est donnée ci-dessus. En effet la distribution exacte de ρ_{ij} n'est pas une distribution de probabilité classique[6]. Si historiquement, les approximations présentées ci-dessus étaient les seules possibilités auxquelles l'expérimentateur pouvaient recourir pour obtenir des intervalles de confiance pour ρ_{ij}, il est devenu possible avec l'augmentation des capacités de calcul dont vous disposez désormais de déterminer les intervalles de confiance exacts pour ρ_{ij}. Ceci présente l'avantage de ne pas avoir à utiliser des résultats asymptotiques qui peuvent s'avérer incorrects si la taille de l'échantillon est petite.

Exemple 3.3.1.

Considérons deux échantillons $\boldsymbol{x_1}$ et $\boldsymbol{x_2}$.

Procédons au test de l'hypothèse $\boldsymbol{x_1}$ est issu d'une variable aléatoire X_1 qui suit une loi normale et au test de l'hypothèse $\boldsymbol{x_2}$ est issu d'une variable aléatoire X_2 qui suit une loi normale. Compte tenu des effectifs on procède au test de **Shapiro-Wilk** ou plutôt de sa variante disponible dans Minitab : le test de **Ryan-Joiner**.

$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
0,19577	-0,81685
0,92204	0,58257
-0,04690	1,24359
0,45863	0,31088
-0,58454	0,95124
1,51722	0,02028
-0,97862	-0,91823
-0,04557	-0,18670
-0,03707	-0,96033
0,84505	1,11031

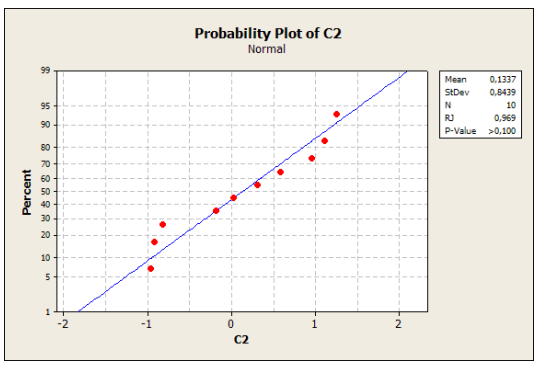
On commence par tester la normalité du couple (X_1, X_2).

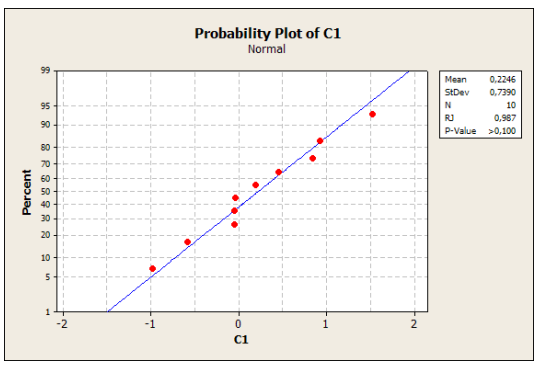
$$\boxed{\mathcal{H}_0 : (X_1, X_2) \text{ suit une loi normale bidimensionnelle}}$$

contre

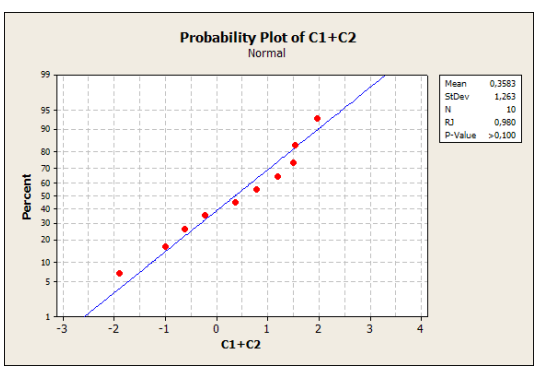
$$\boxed{\mathcal{H}_1 : (X_1, X_2) \text{ ne suit pas une loi bidimensionnelle.}}$$

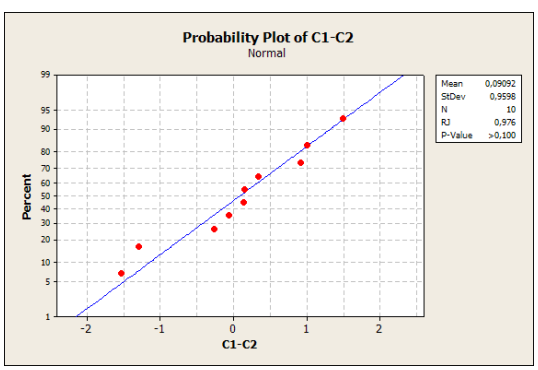
[6]La densité du coefficient de corrélation $\rho\,(X,Y)$ d'un couple gaussien (X,Y) au point $-1 \leqslant r \leqslant 1$ est $\frac{1}{\pi}(n-2)(1-r^2)^{(n-4)/2)}(1-\rho^2)^{(n-1)/2)} \int_0^{+\infty} \frac{\mathrm{d}\beta}{(\cosh\beta - \rho r)^{n-1}}$, voir [15] pour plus de détails.





Les $p-$valeurs sont toutes les deux supérieures à $0,1$: on ne peut donc pas rejeter l'hypothèse de normalité pour X_1 et X_2.
Notons que l'on a uniquement vérifié la normalité de X_1 et celle de X_2. Ceci n'implique pas celle du couple (X_1, X_2). Il s'agit néanmoins d'une **condition nécessaire**. Nous verrons au paragraphe 5.1 comment vérifier l'hypothèse de normalité du couple (X_1, X_2). Une des premières choses à faire est de tester la normalité de $X_1 + X_2$ et de $X_1 - X_2$, voir [5].





Les $p-$valeurs sont supérieures à $0,100$, on ne peut donc pas rejeter l'hypothèse de normalité de $X_1 + X_2$ et de $X_1 - X_2$.
Attention, nous testions ici la multinormalité de (X_1, X_2) et nous avons procédé à quatre tests de Shapiro-Wilk. Nous avions, comme d'habitude, fixé un seuil de $\alpha_{global} = 5$ % pour le test global de multinormalité. Pour garantir ce risque global de 5 %, il faut fixer un risque différent de 5 % pour chaque test de Shapiro-Wilk, vous avez déjà rencontré ce problème dans le contexte des comparaisons multiples en analyse de la variance. En effet si le risque pour chacun des tests de Shapiro-Wilk est fixé à $\alpha_{ind} = 5$ %, alors le risque global est au maximum de $\alpha_{global} = 1 - (1 - 0,05)^4 = 0,185 > 0,05$. Ainsi on doit fixer un seuil de $\alpha_{ind} = 1 - \sqrt[4]{1 - \alpha_{global}} = 1 - \sqrt[4]{0,95} = 0,013$ pour chacun des tests de Shapiro-Wilk, ce qui revient à être moins exigeant pour chacun des tests individuels puisque $0,013 < 0,05$ et donc à accepter la normalité dans plus de cas.
Dans cet exemple cette correction ne change rien à la conclusion puisqu'aucune des $p-$valeurs n'étaient comprises entre $0,013$ et $0,05$.
Ainsi on ne peut pas rejeter l'hypothèse nulle $\mathcal{H}_0$ de normalité du couple (X_1, X_2).

Testons maintenant l'hypothèse d'indépendance de X_1 et X_2. Puisque le couple (X_1, X_2) suit une loi normale, l'indépendance de X_1 et X_2 est équivalente à l'absence de corrélation linéaire entre X_1 et X_2. Il suffit donc de tester :

$$\boxed{\mathcal{H}_0 : \rho(X_1, X_2) = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho(X_1, X_2) \neq 0.}$$

```
Correlations: X1; X2

Pearson correlation of X1 and X2 = 0,270
P-Value = 0,450
```

Avec SPSS 13.0 on obtient :

```
Correlations
                            X1    X2
X1   Pearson Correlation    1     ,270
     Sig. (2-tailed)             ,450
     N                      10    10
X2   Pearson Correlation    ,270  1
     Sig. (2-tailed)        ,450
     N                      10    10
```

En utilisant le module Tests Exacts de SPSS 13.0 :

```
                   Value   Approx. Sig.   Exact Sig.
Pearson's R        0,270   0,450          0,446
N of Valid Cases   10
```

Notons que dans notre cas la significativité exacte est proche de la significativité approchée. Néanmoins, dans de nombreux cas il n'en va pas de même.
Ainsi avec une $p-$valeur de $0,446$, le test n'est pas significatif au seuil $\alpha = 5$ %. On ne peut rejeter l'hypothèse nulle $\mathcal{H}_0$ d'absence de corrélation entre X_1 et X_2. On en déduit que l'on ne peut rejeter l'hypothèse d'indépendance de X_1 et X_2.

.....

3.4. Corrélation multiple

3.4.1. Définition

Le coefficient de corrélation multiple R est la corrélation maximale possible entre une variable réelle X_1 et toutes les combinaisons linéaires de composantes d'un vecteur aléatoire

$\boldsymbol{X_2}$.

Rappelons rapidement ce que l'on entend par **combinaison linéaire**.
Si $\boldsymbol{U}$ et $\boldsymbol{V}$ sont deux vecteurs de $\mathbb{R}^k$, $k \geqslant 1$, une combinaison linéaire de $\boldsymbol{U}$ et $\boldsymbol{V}$ est un vecteur $\mathrm{CL}_{(\boldsymbol{U},\boldsymbol{V})}(\alpha, \beta)$ défini comme la somme $\alpha\boldsymbol{U} + \beta\boldsymbol{V}$ avec $(\alpha, \beta) \in \mathbb{R}^2$.
Cette notion s'étend au cas de n vecteurs : si $\boldsymbol{U_1}, \ldots, \boldsymbol{U_n}$ sont n vecteurs de $\mathbb{R}^k$, $k \geqslant 1$, une combinaison linéaire des $(\boldsymbol{U_i})_{1 \leqslant i \leqslant n}$ est un vecteur $\mathrm{CL}_{(\boldsymbol{U_1},\ldots,\boldsymbol{U_n})}(\alpha_1, \ldots, \alpha_n)$ défini comme la somme $\sum_{i=1}^n \alpha_i \boldsymbol{U_i}$ avec $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.

Mise en équation : on dispose de $X_1 \in \mathbb{R}$ et $\boldsymbol{X_2} \in \mathbb{R}^{p-1}$ tels que $(X_1, \boldsymbol{X_2}) \in \mathbb{R}^p$ ait une distribution dans $\mathbb{R}^p$ de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$.
La moyenne $\boldsymbol{\mu}$ du couple $(X_1, \boldsymbol{X_2})$ est reliée à la moyenne μ_1 de X_1 et $\boldsymbol{\mu_2}$ de $\boldsymbol{X_2}$ par :

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \boldsymbol{\mu_2} \end{pmatrix}.$$

On adopte une écriture particulière pour $\boldsymbol{\Sigma}$ pour faire ressortir la variance de X_1, notée $\sigma_{11}(X_1)$, la matrice de variance-covariance de $\boldsymbol{X_2}$, notée $\boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})$, et les covariances des composantes de $\boldsymbol{X_2}$ avec X_1, qui sont les composantes du vecteur $\boldsymbol{\Sigma_{21}}(X_1, \boldsymbol{X_2})$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}(X_1) & \boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})^T \\ \boldsymbol{\Sigma_{21}}(X_1, \boldsymbol{X_2}) & \boldsymbol{\Sigma_{22}}(\boldsymbol{X_2}) \end{pmatrix}$$

où $\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})^T$ est la transposée de $\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})$. Remarquons que $\mathrm{Var}[X_1] = \sigma_{11}(X_1) = \mathrm{Cov}[X_1, X_1] = \sigma_{X_1}^2$ avec les notations du paragraphe 3.2.2.
On montre par un peu de calcul matriciel que la corrélation maximale R vérifie :

$$R^2(X_1, \boldsymbol{X_2}) = \frac{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})^T \boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})^{-1} \boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}{\sigma_{11}(X_1)}.$$

On peut vérifier qu'il s'agit bien d'un nombre réel.

Le cas de la corrélation simple est légèrement différent de celui-ci puisqu'ici on ne peut trouver facilement que la valeur absolue de $R(X_1, \boldsymbol{X_2})$. A nouveau on doit faire le lien entre le coefficient de détermination en régression linéaire multiple et $R^2(X_1, \boldsymbol{X_2})$: ils ont la même valeur.

3.4.2. Estimation

On se donne, comme précemment, un $n-$échantillon indépendant et identiquement distribué $((X_{1,1}, \boldsymbol{X_{2,1}}), \ldots, (X_{1,n}, \boldsymbol{X_{2,n}}))$. En utilisant les estimateurs définis au paragra-

phe 3.3.2, un estimateur du coefficient $R^2(X_1, \boldsymbol{X_2})$ est donné par la formule suivante :

$$\begin{aligned} \widehat{R^2(X_1, \boldsymbol{X_2})} &= \frac{\widehat{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}^T \widehat{\boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})}^{-1} \widehat{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}}{\widehat{\sigma_{11}(X_1)}} \\ &= \frac{\widehat{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}^T \widehat{\boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})}^{-1} \widehat{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}}{\widehat{\sigma_{11}(X_1)}}. \end{aligned}$$

On note désormais $\boldsymbol{x_1}$ un échantillon de n réalisations de X_1 et $\boldsymbol{x_2}$ un échantillon de n réalisations de $\boldsymbol{X_2}$.
Une estimation de $R^2(X_1, \boldsymbol{X_2})$ est alors :

$$\widehat{R^2(X_1, \boldsymbol{X_2})}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\left(\widehat{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}(\boldsymbol{x_1}, \boldsymbol{x_2})\right)^T \left(\widehat{\boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})}(\boldsymbol{x_2})\right)^{-1} \widehat{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}(\boldsymbol{x_1}, \boldsymbol{x_2})}{\widehat{\sigma_{11}(X_1)}(\boldsymbol{x_1})}.$$

3.4.3. Asymptotique

Sans hypothèse supplémentaire on ne pourrait rien dire[7] sur la distribution de $\widehat{R^2(X_1, \boldsymbol{X_2})}$ et on ne pourrait donc pas effectuer de test.
Supposons désormais que nos $X_1 \in \mathbb{R}$ et $\boldsymbol{X_2} \in \mathbb{R}^{p-1}$ soient tels que $(X_1, \boldsymbol{X_2}) \in \mathbb{R}^p$ ait une distribution **multinormale** dans $\mathbb{R}^p$ de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$. En conservant les notations du paragraphe 3.4.1 on peut écrire :

$$(X_1, \boldsymbol{X_2}) \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu_2} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma_{21}}^T \\ \boldsymbol{\sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}\right).$$

La distribution asymptotique de $R(X_1, \boldsymbol{X_2})$ est alors identique à celle du coefficient de corrélation simple :

$$\sqrt{n}(\widehat{R(X_1, \boldsymbol{X_2})} - R(X_1, \boldsymbol{X_2})) \approx \mathcal{N}\left(0, \left(1 - R(X_1, \boldsymbol{X_2})^2\right)^2\right).$$

Ce résultat n'est pas assez précis pour permettre de résoudre les problèmes qui nous intéressent généralement. On l'utilisera uniquement si l'on ne peut pas faire autrement et si $n \geqslant 30$.

3.4.4. Test de l'hypothèse $R(X_1, \boldsymbol{X_2}) = 0$

Toujours sous l'hypothèse de multinormalité du vecteur $(X_1, \boldsymbol{X_2})$ comme précisé au paragraphe 3.4.3, on peut connaître la loi exacte d'une fonction de l'estimateur $\widehat{R^2}(X_1, \boldsymbol{X_2})$

[7] Le théorème central limite permet d'avoir des informations sur la loi limite de $\widehat{R(X_1, \boldsymbol{X_2})}$ mais ces renseignements ne sont valables que pour des effectifs très importants, $n \to +\infty$ et de ce fait ne présentent que peu d'intérêt dans la pratique.

sous l'hypothèse nulle « $\mathcal{H}_0$: Le coefficient de corrélation multiple $R(X_1, \boldsymbol{X_2})$ de X_1 et $\boldsymbol{X_2}$ est nul » :

$$\frac{n-p-1}{p}\frac{R^2\widehat{(X_1,\boldsymbol{X_2})}}{1-R^2\widehat{(X_1,\boldsymbol{X_2})}^2} \sim \mathcal{F}_{p,n-p-1}$$

où $\mathcal{F}_{p,n-p-1}$ est une loi de Fisher à p et $n-p-1$ degrés de liberté[8].
On se sert de cette propriété pour tester les hypothèses :

$$\boxed{\mathcal{H}_0 : \text{Le coefficient de corrélation multiple } R(X_1, \boldsymbol{X_2}) \text{ de } X_1 \text{ et } \boldsymbol{X_2} \text{ est nul}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Le coefficient de corrélation multiple } R(X_1, \boldsymbol{X_2}) \text{ de } X_1 \text{ et } \boldsymbol{X_2} \text{ est non nul.}}$$

Ces hypothèses sont équivalentes aux hypothèses :

$$\boxed{\mathcal{H}_0 : X_1 \text{ et } \boldsymbol{X_2} \text{ sont indépendants}}$$

contre

$$\boxed{\mathcal{H}_1 : X_1 \text{ et } \boldsymbol{X_2} \text{ sont liés.}}$$

Pour obtenir des intervalles de confiance pour $R(X_1, \boldsymbol{X_2})$ on peut aussi utiliser la transformation en « argument tangente hyperbolique », i.e. en $\tanh^{-1}$, voir le paragraphe 3.2.3.

3.4.5. Test de l'hypothèse $R(X_1, \boldsymbol{X_2}) = R_0$, $R_0 \neq 0$

On se place encore sous l'hypothèse de multinormalité du vecteur $(X_1, \boldsymbol{X_2})$ comme précisé au paragraphe 3.4.3. Notons que la propriété du paragraphe 3.4.4 ci-dessus ne permet pas de tester, si $R_0 \neq 0$, des hypothèses du type :

[8] Si $p = 1$, on retrouve exactement le cas de la corrélation simple. En effet dans cette situation, on étudie la corrélation entre deux variables X_1 et X_2 à valeurs réelles. La formule ci-dessus pour $R^2(X_1, X_2)$ est alors identique à celle de $\rho(X_1, X_2)^2$. De plus, on sait que $\frac{\sqrt{n-2}\rho\widehat{(X_1,X_2)}}{\sqrt{1-\rho\widehat{(X_1,X_2)}^2}}$ suit une loi de Student à $n-2$ degrés de liberté. On montre qu'alors son carré $\left(\sqrt{n-2}\frac{\rho\widehat{(X_1,X_2)}}{\sqrt{1-\rho\widehat{(X_1,X_2)}^2}}\right)^2 = (n-2)\frac{\rho\widehat{(X_1,X_2)}^2}{1-\rho\widehat{(X_1,X_2)}^2} = (n-2)\frac{R^2\widehat{(X_1,X_2)}}{1-R^2\widehat{(X_1,X_2)}^2}$ suit une loi de Fisher à 1 et $n-2$ degrés de liberté. Ceci est bien conforme au résultat de ce paragraphe puisque si $p = 1$ on a vu que $\frac{n-p-1}{p}\frac{R^2\widehat{(X_1,X_2)}}{1-R^2\widehat{(X_1,X_2)}^2} = (n-2)\frac{R^2\widehat{(X_1,X_2)}}{1-R^2\widehat{(X_1,X_2)}^2}$ suit une loi de Fisher à p et $n-p-1$, c'est-à-dire 1 et $n-2$, degrés de liberté. Réciproquement, si $(n-2)\frac{R^2\widehat{(X_1,X_2)}}{1-R^2\widehat{(X_1,X_2)}^2}$ suit une loi de Fisher à 1 et $n-2$, on sait qu'alors $\sqrt{(n-2)\frac{R^2\widehat{(X_1,X_2)}}{1-R^2\widehat{(X_1,X_2)}^2}} = \frac{\sqrt{n-2}\rho\widehat{(X_1,X_2)}}{\sqrt{1-\rho\widehat{(X_1,X_2)}^2}}$ suit une loi de Student à $n-2$ degrés de liberté.

$$\boxed{\mathcal{H}_0 : R(X_1, \boldsymbol{X_2}) = R_0}$$

contre

$$\boxed{\mathcal{H}_1 : R(X_1, \boldsymbol{X_2}) \neq R_0.}$$

.

La démarche est alors la même que pour le coefficient de corrélation simple, voir paragraphe 3.3.4. Deux approches sont possibles :

- Utiliser la transformation de Fisher introduite ci-dessus, au paragraphe 3.2.3, pour obtenir un intervalle de confiance pour $R(X_1, \boldsymbol{X_2})$ de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

  où $\widehat{z} = \tanh\left(\widehat{R(X_1, \boldsymbol{X_2})}(\boldsymbol{x_1}, \boldsymbol{x_2})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Utiliser des logiciels, comme SPSS, R , SAS ou StatXact, qui proposent des versions *exactes* de l'intervalle de confiance dont une approximation est donnée ci-dessus. Cette approche est particulièrement intéressante si l'effectif commun n aux échantillons $\boldsymbol{x_1}$ et $\boldsymbol{x_2}$ est petit.

3.5. Corrélation partielle

Pour introduire formellement le concept de corrélation partielle, un petit détour mathématique est nécessaire. Les applications seront généralement beaucoup plus simples.

3.5.1. Définition

Considérons deux vecteurs $\boldsymbol{X_1} \in \mathbb{R}^q$ et $\boldsymbol{X_2} \in \mathbb{R}^{m-q}$ qui ont une distribution conjointe multinormale :

$$(\boldsymbol{X_1}, \boldsymbol{X_2}) \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}\right).$$

La distribution conditionnelle de $\boldsymbol{X_1}$ sachant $\boldsymbol{X_2}$, on étend les résultats du paragraphe 3.2.1 qui concernaient le cas de deux variables réelles au cas de deux vecteurs, est :

$$\boldsymbol{X_1}|\boldsymbol{X_2} \sim \mathcal{N}\left(\boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}(\boldsymbol{X_2} - \boldsymbol{\mu_2}), \boldsymbol{\Sigma_{11|2}}\right),$$

où $\boldsymbol{\Sigma_{11|2}} = \boldsymbol{\Sigma_{1,1}} - \boldsymbol{\Sigma_{1,2}}\boldsymbol{\Sigma_{2,2}}^{-1}\boldsymbol{\Sigma_{2,1}}$.
En écrivant $\boldsymbol{\Sigma_{11|2}} = \left(\left(\sigma_{i,j|q+1,\ldots,m}\right)\right)_{1 \leqslant i,j \leqslant p}$, il s'agit d'une matrice carrée de taille p, on

définit le coefficient de corrélation partielle entre les composantes i et j de $\boldsymbol{X_1}$ sachant $\boldsymbol{X_2}$ comme :

$$\rho_{i,j|q+1,\ldots,m} = \frac{\sigma_{i,j|q+1,\ldots,m}}{\sigma_{i,i|q+1,\ldots,m}^{\frac{1}{2}}\sigma_{j,j|q+1,\ldots,m}^{\frac{1}{2}}}.$$

Son interprétation pratique est la suivante :
Le coefficient de corrélation partielle entre les composantes i et j du vecteur $\boldsymbol{X_1}$ sachant $\boldsymbol{X_2}$ représente la corrélation entre les composantes i et j après avoir éliminé l'effet des variables de $\boldsymbol{X_2}$ sur les composantes i et j du vecteur $\boldsymbol{X_1}$.

3.5.2. Estimation

On se donne, comme précemment, un $n-$échantillon indépendant et identiquement distribué $((\boldsymbol{X_{1,1}}, \boldsymbol{X_{2,1}}), \ldots, (\boldsymbol{X_{1,n}}, \boldsymbol{X_{2,n}}))$. En utilisant les estimateurs définis au paragraphe 3.3.2, un estimateur du coefficient $\rho_{i,j|q+1,\ldots,m}$ est donné par la formule suivante :

$$\widehat{\rho_{i,j|q+1,\ldots,m}} = \frac{\widehat{\sigma_{i,j|q+1,\ldots,m}}}{\widehat{\sigma_{i,i|q+1,\ldots,m}}^{\frac{1}{2}}\widehat{\sigma_{j,j|q+1,\ldots,m}}^{\frac{1}{2}}},$$

où $\left(\left(\widehat{\sigma_{i,j|q+1,\ldots,m}}\right)\right)_{1\leqslant i,j\leqslant p} = \widehat{\boldsymbol{\Sigma_{11|2}}} = \widehat{\boldsymbol{\Sigma_{1,1}}} - \widehat{\boldsymbol{\Sigma_{1,2}}}\widehat{\boldsymbol{\Sigma_{2,2}}}^{-1}\widehat{\boldsymbol{\Sigma_{2,1}}}$.
On note désormais $\boldsymbol{x_1}$ un échantillon de n réalisations de $\boldsymbol{X_1}$ et $\boldsymbol{x_2}$ un échantillon de n réalisations de $\boldsymbol{X_2}$.
Une estimation de $\rho_{i,j|q+1,\ldots,m}$ est alors :

$$\widehat{\rho_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\widehat{\sigma_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})}{\widehat{\sigma_{i,i|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})^{\frac{1}{2}}\widehat{\sigma_{j,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})^{\frac{1}{2}}},$$

où

$$\begin{aligned}\left(\left(\widehat{\sigma_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})\right)\right)_{1\leqslant i,j\leqslant p} &= \widehat{\boldsymbol{\Sigma_{11|2}}}(\boldsymbol{x_1}, \boldsymbol{x_2}) \\ &= \widehat{\boldsymbol{\Sigma_{1,1}}}(\boldsymbol{x_1}, \boldsymbol{x_2}) - \widehat{\boldsymbol{\Sigma_{1,2}}}(\boldsymbol{x_1}, \boldsymbol{x_2})\left(\widehat{\boldsymbol{\Sigma_{2,2}}}(\boldsymbol{x_1}, \boldsymbol{x_2})\right)^{-1}\widehat{\boldsymbol{\Sigma_{2,1}}}(\boldsymbol{x_1}, \boldsymbol{x_2}).\end{aligned}$$

3.5.3. Asymptotique

La distribution asymptotique d'un coefficient de corrélation partielle est la même que pour une corrélation simple, c'est-à-dire :

$$\sqrt{n}\left(\widehat{\rho_{i,j|q+1,\ldots,m}} - \rho_{i,j|q+1,\ldots,m}\right) \approx \mathcal{N}\left(0, (1-\rho^2_{i,j|q+1,\ldots,m})^2\right).$$

Ce résultat n'est applicable que si $n \geqslant 30$. Ne s'en servir que si l'on ne peut pas faire autrement.

3.5.4. Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$

On peut maintenant s'intéresser au test de nullité du coefficient de corrélation partielle des composantes du vecteur gaussien $\boldsymbol{X_1}$ connaissant le vecteur $\boldsymbol{X_2}$.

$$\boxed{\mathcal{H}_0 : \rho_{i,j|q+1,\ldots,m} = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{i,j|q+1,\ldots,m} \neq 0.}$$

Ces hypothèses sont équivalente à :

$$\boxed{\mathcal{H}_0 : \text{Les composantes } X_{1,i} \text{ et } X_{1,j} \text{ de } \boldsymbol{X_1} \text{ sont indépendantes sans l'effet de } \boldsymbol{X_2}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les composantes } X_{1,i} \text{ et } X_{1,j} \text{ de } \boldsymbol{X_1} \text{ sont liées sans l'effet de } \boldsymbol{X_2}.}$$

Sous l'hypothèse nulle $\mathcal{H}_0$, alors :

$$t_{i,j|q+1,\ldots,m}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sqrt{n-m+q-2}\frac{\widehat{\rho_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})}{\sqrt{1-\widehat{\rho_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})^2}}$$

est la réalisation d'une variable aléatoire T qui suit une loi de Student à $n-m+q-2$ degrés de liberté, i.e. $T \sim \mathcal{T}_{n-m+q-2}$.
On rejette l'hypothèse nulle $\mathcal{H}_0$ au niveau $(1-\alpha)$ % lorsque $\left|t_{i,j|q+1,\ldots,m}(\boldsymbol{x_1}, \boldsymbol{x_2})\right| > t_{n-m+q-2,\alpha/2}$.
On peut aussi appliquer la transformation z dans ce contexte, voir le paragraphe 3.2.3.

3.5.5. Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$

Si $\rho_0 \neq 0$, la propriété du paragraphe 3.5.4 ci-dessus ne permet pas de tester des hypothèses du type :

$$\boxed{\mathcal{H}_0 : \rho_{i,j|q+1,\ldots,m} = \rho_0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{i,j|q+1,\ldots,m} \neq \rho_0.}$$

Il existe alors deux possibilités :

- Si l'on souhaite faire un test de ce type à la main, on utilise la transformation de Fisher introduite ci-dessus, au paragraphe 3.2.3, pour obtenir un intervalle de confiance pour $\rho_{i,j|q+1,\ldots,m}$ de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

où $\widehat{z} = \tanh\left(\widehat{\rho_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1}, \boldsymbol{x_2})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Certains logiciels, comme SPSS, R , SAS ou StatXact, proposent des versions *exactes* de l'intervalle de confiance.

3.5.6. Cas de trois variables réelles

Intéressons-nous au cas le plus simple. Soient X_1, X_2 et X_3 trois variables aléatoires réelles telles que la loi jointe de (X_1, X_1, X_3) soit multinormale de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$. Ainsi on a ici $m = 3$ et $q = 2$. On montre alors la relation suivante :

$$\rho_{12|3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

où ρ_{12} est le coefficient de corrélation simple entre les variables X_1 et X_2, ρ_{13} est le coefficient de corrélation simple entre les variables X_1 et X_3 et ρ_{23} est le coefficient de corrélation simple entre les variables X_2 et X_3, voir le paragraphe 3.3.1.
Dans la plupart des situations expérimentales que vous rencontrerez cette formule sera suffisante. On remarque également que la définition est symétrique en X_1 et X_2, c'est-à-dire :

$$\rho_{12|3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}} = \frac{\rho_{21} - \rho_{23}\rho_{13}}{\sqrt{1-\rho_{23}^2}\sqrt{1-\rho_{13}^2}} = \rho_{21|3}.$$

La corrélation partielle de X_1 et X_2 sachant X_3 est heureusement la même que celle de X_2 et X_1 sachant X_3.

On considère un échantillon $\boldsymbol{x}$ indépendant et identiquement distribué suivant la loi de X_1, X_2, X_3. On note $\boldsymbol{x_1}$ l'échantillon des réalisations de X_1, $\boldsymbol{x_2}$ l'échantillon des réalisations de X_2 et $\boldsymbol{x_3}$ l'échantillon des réalisations de X_3.

On tire de $\boldsymbol{x}$ une estimation de $\rho_{12|3}$ en calculant des estimations de ρ_{12}, ρ_{13}, ρ_{23} comme expliqué au paragraphe 3.3.2.
On peut maintenant s'intéresser au test de nullité du coefficient de corrélation partielle de X_1 et X_2 connaissant la variable X_3.

$$\boxed{\mathcal{H}_0 : \rho_{12|3} = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{12|3} \neq 0.}$$

Ces hypothèses sont équivalente à :

$$\boxed{\mathcal{H}_0 : \text{Les variables } X_1 \text{ et } X_2 \text{ sont indépendantes sans l'effet de } X_3}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les variables } X_1 \text{ et } X_2 \text{ sont liées sans l'effet de } X_3.}$$

Sous l'hypothèse nulle $\mathcal{H}_0$, alors :

$$t_{12|3}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}) = \sqrt{n-3}\frac{\widehat{\rho_{12|3}}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})}{\sqrt{1-\widehat{\rho_{12|3}}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})^2}}$$

est la réalisation d'une variable aléatoire T qui suit une loi de Student à $n-3$ degrés de liberté, i.e. $T \sim \mathcal{T}_{n-3}$.
On rejette l'hypothèse nulle $\mathcal{H}_0$ au niveau $(1-\alpha)$ % lorsque $\left|t_{12|3}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})\right| > t_{n-3,\alpha/2}$.
On peut aussi appliquer la transformation z dans ce contexte, voir le paragraphe 3.2.3.

Si $\rho_0 \neq 0$, la propriété du paragraphe ci-dessus ne permet pas de tester des hypothèses du type :

$$\boxed{\mathcal{H}_0 : \rho_{12|3} = \rho_0}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{12|3} \neq \rho_0.}$$

Il existe alors deux possibilités :

- Si l'on souhaite faire un test de ce type à la main, on utilise la transformation de Fisher introduite ci-dessus, au paragraphe 3.2.3, pour obtenir un intervalle de confiance pour $\rho_{12|3}$ de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

  où $\widehat{z} = \tanh\left(\widehat{\rho_{12|3}}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Certains logiciels, comme SPSS, R , SAS ou StatXact, proposent des versions *exactes* de l'intervalle de confiance.

Exemple 3.5.1.

Détaillons le traitement de la corrélation partielle dans la situation suivante où l'on dispose de trois variables continues réelles.

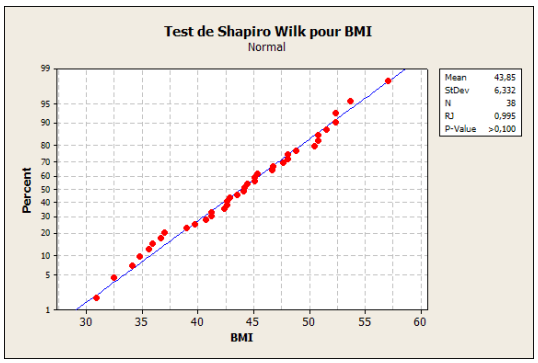
- Le BMI X_1.
- Le poids X_2.
- La taille X_3.

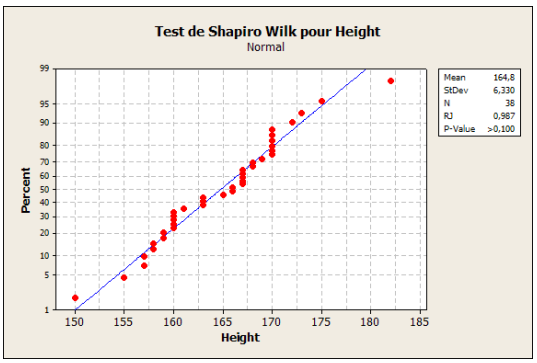
On rappelle que le BMI d'un individu est un indice qui se calcule à partir de la taille et du poids par la formule suivante :

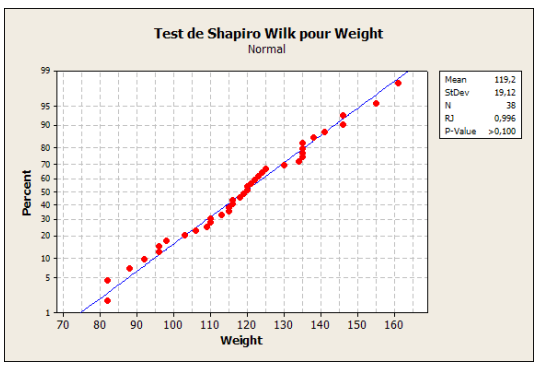
$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}$$

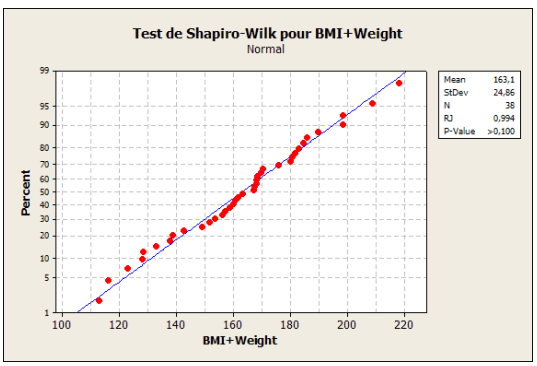
où la masse s'exprime en kg et la taille en m.
L'échantillon étudié a un effectif $n = 38$ et est exclusivement constitué d'individus de sexe féminin.

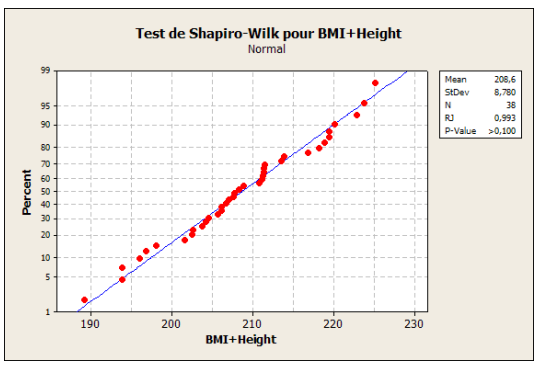
Comme nous l'avons fait remarquer, l'hypothèse nécessaire à l'approche paramétrique de la corrélation simple, multiple ou partielle qui vous a été présentée plus haut sont que la loi du vecteur (X_1, X_2, X_3) est une loi multinormale sur $\mathbb{R}^3$.
Ceci a pour conséquence qu'il faut que X_1, X_2 et X_3 aient une distribution normale mais comme nous l'avons déjà souligné ce n'est pas suffisant. On doit par exemple aussi tester la normalité de $X_1 + X_2$, $X_2 + X_3$, $X_1 + X_3$ et $X_1 + X_2 + X_3$. Se référer au paragraphe 5.1 pour un exposé détaillé des tests de multinormalité ou au livre de R. Christensen [5].

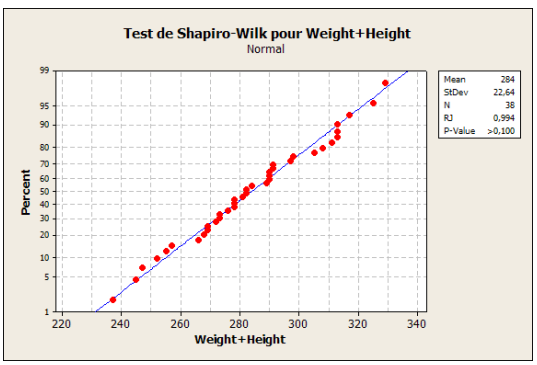


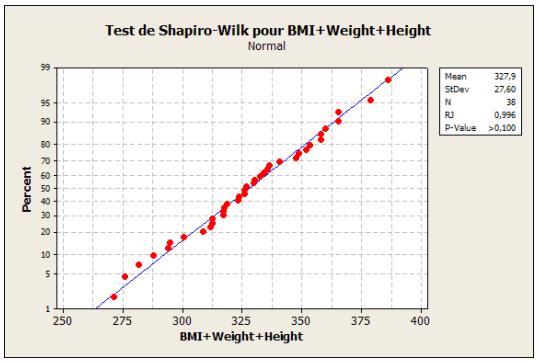












Les $p-$valeurs sont toutes supérieures à $0,100$, on ne peut donc pas rejeter l'hypothèse de normalité de X_1, X_2, X_3, X_1+X_2, X_1+X_3, X_2+X_3 et de $X_1+X_2+X_3$.
Attention, nous testions ici la multinormalité de (X_1, X_2, X_3) et nous avons procédé à sept tests de Shapiro-Wilk. Nous avions, comme d'habitude, fixé un seuil de $\alpha_{global} = 5$ % pour le test global de multinormalité. Pour garantir ce risque global de 5 %, il faut fixer un risque différent de 5 % pour chaque test de Shapiro-Wilk, nous avons déjà évoqué ce problème à l'exemple 3.3.1. En effet si le risque pour chacun des tests de Shapiro-Wilk est fixé à $\alpha_{ind} = 5$ %, alors le risque global est au maximum de $\alpha_{global} = 1-(1-0,05)^7 = 0,302 \gg 0,05$. Ainsi on doit fixer un seuil de $\alpha_{ind} = 1-\sqrt[7]{1-\alpha_{global}} = 1-\sqrt[7]{0,95} = 0,007$ pour chacun des sept tests de Shapiro-Wilk, ce qui revient à être moins exigeant pour chacun des tests individuels puisque $0,007 < 0,05$ et donc à accepter la normalité dans plus de cas.
Dans cet exemple cette correction ne change rien à la conclusion puisqu'aucune des $p-$valeurs n'étaient comprises entre $0,007$ et $0,05$.
Suite à cette analyse sommaire rien ne permet de rejeter l'hypothèse de multinormalité de (X_1, X_2, X_3).

On peut donc utiliser une approche paramétrique. Commençons par étudier les coefficients de corrélation simple, dits de Pearson.

```
Correlations: BMI; Weight; Height

           BMI  Weight
Weight   0,878
         0,000

Height  -0,038   0,441
         0,820   0,006

Cell Contents: Pearson correlation
               P-Value
```

On constate ainsi que :

- Au niveau $\alpha = 5$ %, on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables BMI et Weight, car $0,000 < 0,05$.
- Au niveau $\alpha = 5$ %, on ne peut pas rejetter l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables BMI et Height, car $0,820 > 0,05$.
- Au niveau $\alpha = 5$ %, on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables Height et Weight, car $0,006 < 0,05$.

Ces résultats sont en accord avec la formule permettant de calculer le BMI à l'aide de la taille et de la masse.

- La liaison est linéaire et positive pour la masse, ce que l'on retrouve bien ici avec la valeur de $\rho_{BW} = 0,878$ et une $p-$valeur de 0,000.
- La liaison est négative pour la taille mais, vu sa faiblesse, $\rho_{BH} = -0,038$, on ne peut pas rejeter l'hypothèse d'indépendance entre ces deux variables, $p-$valeur de 0,820. La formule ci-dessus implique une **relation non linéaire** entre le BMI et la taille c'est pourquoi le coefficient de corrélation linéaire ne peut la mettre en évidence.
- La liaison positive entre la taille et la masse, $\rho_{WH} = 0,441$ et $p-$valeur de 0,006, est en accord avec notre connaissance a priori de l'existence d'une relation linéaire positive entre la taille et la masse d'un individu.

Passons maintenant au calcul de la corrélation multiple R du BMI sur (Height, Weight), de la corrélation multiple R du Height sur (BMI, Weight) et de la corrélation multiple R du Weight sur (Height, BMI).

Pour effectuer ce calcul on utilise le fait que R^2 est égal au coefficient de détermination obtenu lorsque l'on fait la régression linéaire multiple de BMI sur (Height, Weight). On obtient par exemple avec Minitab :

```
Regression Analysis: BMI versus Weight; Height

The regression equation is
BMI = 87,1 + 0,368 Weight - 0,529 Height

Predictor       Coef    SE Coef       T      P
Constant      87,090      1,899   45,85  0,000
Weight      0,367955   0,004146   88,75  0,000
Height      -0,52855    0,01252  -42,21  0,000

S = 0,432656   R-Sq = 99,6%   R-Sq(adj) = 99,5%
```

Pour calculer les autres coefficients de corrélation multiple, on procède de même en changeant les rôles de BMI, Height et Weight.

```
Regression Analysis: Weight versus BMI; Height

The regression equation is
Weight = - 236 + 2,71 BMI + 1,44 Height

Predictor      Coef  SE Coef       T      P
Constant   -236,083    5,253  -44,95  0,000
BMI         2,70570  0,03049   88,75  0,000
Height      1,43599  0,03050   47,09  0,000

S = 1,17324   R-Sq = 99,6%   R-Sq(adj) = 99,6%

Regression Analysis: Height versus BMI; Weight

The regression equation is
Height = 164 - 1,86 BMI + 0,686 Weight

Predictor      Coef  SE Coef       T      P
Constant    164,436    0,933  176,17  0,000
BMI        -1,85552  0,04396  -42,21  0,000
Weight      0,68556  0,01456   47,09  0,000

S = 0,810650   R-Sq = 98,4%   R-Sq(adj) = 98,4%
```

On a ainsi trouvé successivement :

$$\begin{aligned}
\rho_{B/^s(H,W)} &= 99,6\ \% \\
\rho_{W/^s(B,H)} &= 99,6\ \% \\
\rho_{H/^s(B,W)} &= 98,4\ \%
\end{aligned}$$

Calculons désormais les trois corrélations partielles $\rho_{(\text{Height},\text{Weight})|\text{BMI}}$, $\rho_{(\text{Weight},\text{BMI})|\text{Height}}$ et $\rho_{(\text{Height},\text{BMI})|\text{Weight}}$.

On applique la formule ci-dessus :

$$\begin{aligned}
\rho_{HW|B} &= \frac{\rho_{HW} - \rho_{HB}\rho_{WB}}{\sqrt{1-\rho_{HB}^2}\sqrt{1-\rho_{WB}^2}} \approx 0,992 \\
\rho_{WB|H} &= \frac{\rho_{WB} - \rho_{WH}\rho_{BH}}{\sqrt{1-\rho_{WH}^2}\sqrt{1-\rho_{BH}^2}} \approx 0,998 \\
\rho_{HB|W} &= \frac{\rho_{HB} - \rho_{HW}\rho_{BW}}{\sqrt{1-\rho_{HW}^2}\sqrt{1-\rho_{BW}^2}} \approx -0,990.
\end{aligned}$$

Ces résultats sont à comparer avec les valeurs de corrélation simple. En particulier la corrélation entre le poids et la taille en éliminant l'effet du poids semble significative. Pour aller plus loin et décider de la significativité de ces corrélations, on peut se réferrer à une table ou utiliser la loi du coefficient de corrélation partielle énoncée aux paragraphes 3.5.3 et 3.5.6.
Certains logiciels, comme SPSS, nous renseignent directement sur les $p-$valeurs associées aux corrélations partielles :

Partial Correlations

Control Variables			Weight	Height
BMI	Weight	Correlation	1,000	,992
		Significance (2-tailed)	.	,000
		df	0	35
	Height	Correlation	,992	1,000
		Significance (2-tailed)	,000	.
		df	35	0

Partial Correlations

Control Variables			Weight	BMI
Height	Weight	Correlation	1,000	,998
		Significance (2-tailed)	.	,000
		df	0	35
	BMI	Correlation	,998	1,000
		Significance (2-tailed)	,000	.
		df	35	0

Partial Correlations

Control Variables			BMI	Height
Weight	BMI	Correlation	1,000	-,990
		Significance (2-tailed)	.	,000
		df	0	35
	Height	Correlation	-,990	1,000
		Significance (2-tailed)	,000	.
		df	35	0

On constate ainsi que les trois tests sont significatifs au seuil $\alpha = 5$ %.
On ne peut conserver l'hypothèse d'indépendance de deux variables sachant la troisième et ce dans les trois cas qui se présentent ici : $(\text{Height}, \text{Weight})|\text{BMI}$, $(\text{Weight}, \text{BMI})|\text{Height}$ et $(\text{Height}, \text{BMI})|\text{Weight}$.

Avez-vous une critique à formuler quant au modèle que nous avons utilisé ici ? Par exemple l'hypothèse de multinormalité est-elle vraisemblable ? Bien que les tests utilisés ne l'infirment pas, notre connaissance de la relation entre le BMI et le couple $(\text{Height}, \text{Weight})$ ne devait-elle pas nous faire exclure ce modèle ?

.....

Exemple 3.5.2.
On réalise dans un collège deux tests d'évaluation communs à tous les élèves quel que soit leur âge mais exclusivement de sexe masculin. L'un porte sur leur compétence en mathématiques et l'autre en sport, le temps mis à parcourir 100 m en course à pied. Les performances ont été notées sur 100 dans les deux tests puis ramenées à des notes sur 20. On dispose donc de deux variables aléatoires :
- l'une appelée Math et associée à la note en mathématiques qu'a reçu un élève,
- l'autre appelée Math et associée à la note en sport qu'a reçu un élève.

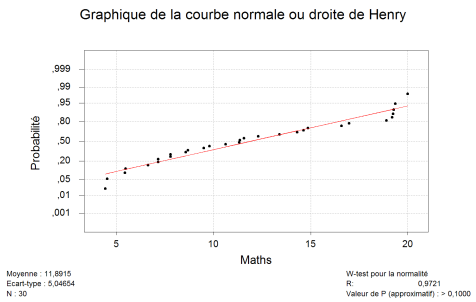
On les abrégera parfois en M et S.

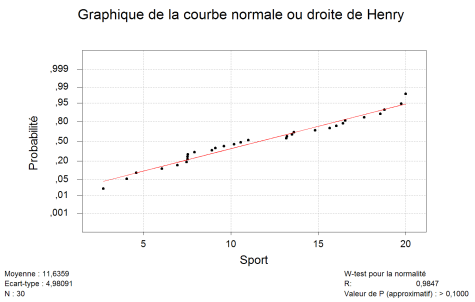
Les données ont été reportées dans le tableau suivant :

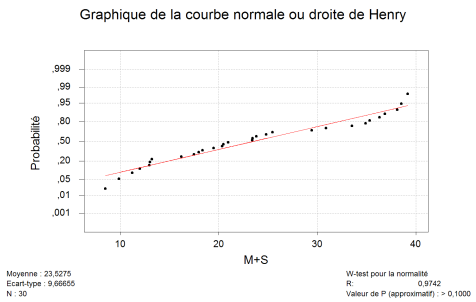
Élève	Math	Sport	Élève	Math	Sport
1	7,15	2,69	16	12,29	13,16
2	7,79	10,56	17	11,56	7,92
3	8,57	8,91	18	14,31	9,11
4	7,14	6,04	19	11,33	13,49
5	5,47	7,53	20	9,49	10,99
6	4,43	4,04	21	16,99	16,54
7	4,52	7,45	22	14,86	16,03
8	6,63	4,58	23	18,91	16,41
9	5,43	7,50	24	19,35	14,81
10	7,78	10,18	25	19,24	17,63
11	9,79	13,59	26	14,62	18,78
12	13,40	6,92	27	20,00	18,54
13	8,68	7,51	28	19,19	20,00
14	11,35	9,59	29	19,28	15,64
15	10,62	13,19	30	16,59	19,73

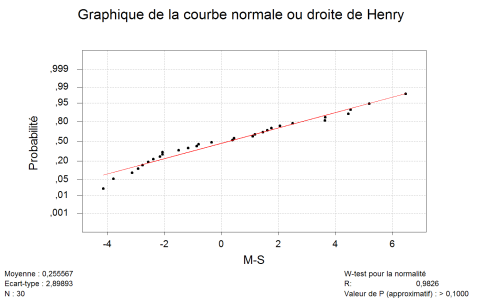
L'échantillon étudié a un effectif $n = 30$ et est exclusivement constitué d'individus de sexe masculin.

Comme nous l'avons fait remarquer, l'hypothèse nécessaire à l'approche paramétrique de la corrélation simple, multiple ou partielle qui vous a été présentée plus haut est que la loi du vecteur (M, S) est une loi multinormale sur $\mathbb{R}^2$.
Ceci a pour conséquence qu'il faut que M, S aient une distribution normale mais comme nous l'avons déjà souligné ce n'est pas suffisant. On doit par exemple aussi tester la normalité de $\mathrm{M} + \mathrm{S}$ et $\mathrm{M} - \mathrm{S}$. Se référer au paragraphe 5.1 pour un exposé détaillé des tests de multinormalité ou au livre de R. Christensen [5].









Les $p-$valeurs sont toutes supérieures à $0,100$, on ne peut donc pas rejeter l'hypothèse de normalité de M, S, M + S et M − S.
Attention, nous testions ici la multinormalité de (M, S) et nous avons procédé à quatre tests de Shapiro-Wilk. Attention, nous testions ici la multinormalité de (X, Y) et nous avons procédé à quatre tests de Shapiro-Wilk. Nous avions, comme d'habitude, fixé un seuil de $\alpha_{global} = 5$ % pour le test global de multinormalité. Pour garantir ce risque global de 5 %, il faut fixer un risque différent de 5 % pour chaque test de Shapiro-Wilk. En effet si le risque pour chacun des tests de Shapiro-Wilk est fixé à $\alpha_{ind} = 5$ %, alors le risque global est au maximum de $\alpha_{global} = 1 - (1 - 0,05)^4 = 0,185 > 0,05$. Ainsi on doit fixer un seuil de $\alpha_{ind} = 1 - \sqrt[4]{1 - \alpha_{global}} = 1 - \sqrt[4]{0,95} = 0,013$ pour chacun des tests de Shapiro-Wilk, ce qui revient à être moins exigeant pour chacun des tests individuels puisque $0,013 < 0,05$ et donc à accepter la normalité dans plus de cas.
Dans cet exemple cette correction ne change rien à la conclusion puisqu'aucune des $p-$valeurs n'étaient comprises entre $0,013$ et $0,05$. Suite à cette analyse sommaire rien ne permet de rejeter l'hypothèse de multinormalité de (M, S).

On peut donc utiliser une approche paramétrique et utiliser le coeficient de corrélation de M et S.

```
Corrélation de Pearson de Maths et Sport = 0,833
Valeur de p = 0,000
```

On constate ainsi qu'au seuil $\alpha = 5$ %, on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables Math et Sport, car $0,000 < 0,05$. On décide donc qu'il y a une corrélation significative entre les résultats en mathématiques et en sport. La valeur de l'estimation, $0,833$, du coefficient de corrélation étant positive, l'association est positive, c'est-à-dire, plus on est fort en sport plus on est fort en mathématiques. Qu'en pensez-vous ? **De quel autre facteur faudrait-il tenir compte ?**

Il est évident que les résultats des élèves à ces tests dépend de leur âge puisque le même questionnaire est posé à tous les élèves quelque soit leur niveau, donc aussi bien en classe de $6^{\text{ème}}$ qu'en classe de $3^{\text{ème}}$.

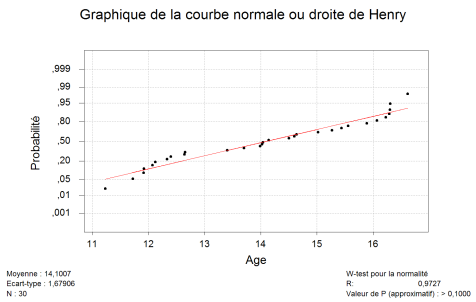
3.5 Corrélation partielle

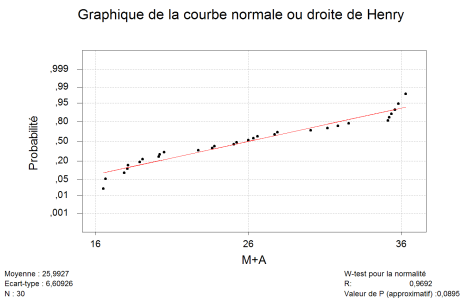
Fort heureusement il a été possible de retrouver l'âge, et même une information encore plus précise, la date de naissance, de chacun des élèves qui a participé à l'évaluation. On a alors décidé de coder l'âge comme une variable aléatoire quantitative continue, appelée Age et parfois abrégée en A. Cette décision correspond à l'idée, relativement raisonnable, suivante : le développement d'un enfant ne serait pas exactement le même si celui-ci est né en début ou en fin d'année civile.

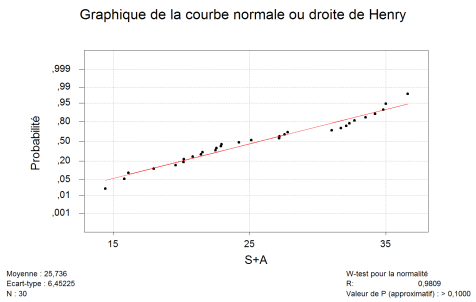
Élève	Math	Sport	Age	Élève	Math	Sport	Age
1	7,15	2,69	11,72	16	12,29	13,16	14,02
2	7,79	10,56	12,33	17	11,56	7,92	15,02
3	8,57	8,91	11,91	18	14,31	9,11	13,40
4	7,14	6,04	11,92	19	11,33	13,49	13,70
5	5,47	7,53	12,64	20	9,49	10,99	14,14
6	4,43	4,04	12,07	21	16,99	16,54	15,56
7	4,52	7,45	12,12	22	14,86	16,03	16,31
8	6,63	4,58	11,23	23	18,91	16,41	16,30
9	5,43	7,50	12,65	24	19,35	14,81	16,23
10	7,78	10,18	12,40	25	19,24	17,63	15,89
11	9,79	13,59	13,99	26	14,62	18,78	15,44
12	13,40	6,92	14,50	27	20,00	18,54	16,29
13	8,68	7,51	14,04	28	19,19	20,00	16,62
14	11,35	9,59	14,64	29	19,28	15,64	16,07
15	10,62	13,19	14,60	30	16,59	19,73	15,27

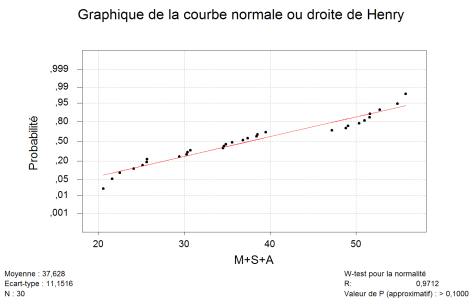
Puisque nous introduisons une nouvelle variable aléatoire Age, pour étudier les corrélations du vecteur (Math, Sport, Age) nous devons tester la normalité du vecteur (Math, Sport, Age).

On procède comme précedemment. En s'appuyant sur les résultats des tests de normalité ci-dessus, on voit qu'il ne reste plus qu'à s'intéresser à la normalité de Math + Age, Sport + Age et de Math + Sport + Age.









Aucun de ces tests n'est significatif au seuil 5 % et donc **a fortiori** ils ne seront pas significatifs au seuil inférieur qu'il faudrait fixer pour garantir un risque global de 5 %. En ajoutant ces résultats aux précédents, rien ne permet de rejeter l'hypothèse de normalité du vecteur (Math, Sport, Age).

On calcule donc les corrélations simples entre les composantes de ce vecteur :

```
Corrélations : Maths; Sport; Age

          Maths    Sport
Sport     0,833
          0,000

Age       0,909    0,837
          0,000    0,000

Contenu de la cellule : corrélation de Pearson
               Valeur de p
```

On note une corrélation positive et significative, au seuil $\alpha = 5$ % pour chacune des associations possibles. Ainsi, comme on pouvait le prévoir, plus l'on est âgé, mieux l'on réussit en mathématiques et en sport.

Exerçons-nous encore au calcul de coefficients de corrélation multiple. Ainsi déterminons la corrélation multiple R du Math sur (Sport, Age), la corrélation multiple R du Sport sur (Math, Age) et la corrélation multiple R du Age sur (Math, Sport).

Pour effectuer ce calcul on utilise le fait que R^2 est égal au coefficient de détermination obtenu lorsque l'on fait la régression linéaire multiple de Math sur (Sport, Age). On obtient par exemple avec Minitab :

```
Analyse de régression : Maths en fonction de Sport; Age

L'équation de régression est
```

```
Maths = - 20,9 + 0,244 Sport + 2,13 Age

Régresseur       Coef   Er-T coef          T        P
Constante     -20,919       4,635      -4,51    0,000
Sport          0,2437      0,1412       1,73    0,096
Age            2,1258      0,4189       5,07    0,000

S = 2,071       R-carré = 84,3%  R-carré (ajust) = 83,2%
```

Pour calculer les autres coefficients de corrélation multiple, on procède de même en changeant les rôles de Math, Sport et Age.

```
Analyse de régression : Sport en fonction de Maths; Age

L'équation de régression est
Sport = - 12,5 + 0,408 Maths + 1,37 Age

Régresseur       Coef   Er-T coef          T        P
Constante     -12,533       7,565      -1,66    0,109
Maths          0,4077      0,2362       1,73    0,096
Age            1,3701      0,7099       1,93    0,064

S = 2,679       R-carré = 73,1%  R-carré (ajust) = 71,1%

Analyse de régression : Age en fonction de Maths; Sport

L'équation de régression est
Age = 10,3 + 0,230 Maths + 0,0885 Sport

Régresseur       Coef   Er-T coef          T        P
Constante     10,3401      0,3338      30,97    0,000
Maths         0,22965     0,04525       5,07    0,000
Sport         0,08848     0,04585       1,93    0,064

S = 0,6807      R-carré = 84,7%  R-carré (ajust) = 83,6%
```

On a ainsi trouvé successivement :

$$\begin{aligned}
\rho_{M/^s(S,A)} &= 84,3\ \% \\
\rho_{S/^s(M,A)} &= 73,1\ \% \\
\rho_{A/^s(M,S)} &= 84,7\ \%.
\end{aligned}$$

Calculons désormais les trois corrélations partielles $\rho_{(\text{Maths,Sport})|\text{Age}}$, $\rho_{(\text{Maths,Age})|\text{Sport}}$ et $\rho_{(\text{Sport,Age})|\text{Maths}}$.

On applique la formule ci-dessus :

$$\begin{aligned}
\rho_{MS|A} &= \frac{\rho_{MS} - \rho_{MA}\rho_{SA}}{\sqrt{1-\rho_{MA}^2}\sqrt{1-\rho_{SA}^2}} \approx 0,315 \\
\rho_{MA|S} &= \frac{\rho_{MA} - \rho_{MS}\rho_{SA}}{\sqrt{1-\rho_{MS}^2}\sqrt{1-\rho_{SA}^2}} \approx 0,699 \\
\rho_{SA|M} &= \frac{\rho_{SA} - \rho_{MS}\rho_{MA}}{\sqrt{1-\rho_{MS}^2}\sqrt{1-\rho_{MA}^2}} \approx 0,348.
\end{aligned}$$

Ces résultats sont à comparer avec les valeurs de corrélation simple. En particulier la corrélation entre le poids et la taille en éliminant l'effet du poids semble significative. Pour aller plus loin et décider de la significativité de ces corrélations on peut se réferrer à une table ou utiliser la loi du coefficient de corrélation partielle énoncée aux paragraphes 3.5.3 et 3.5.6.
Certains logiciels, comme SPSS, nous renseignent directement sur les $p-$valeurs associées aux corrélations partielles :

```
Correlations
 | ----------------- | ----- | ------------------------ | ----- | ----- |
 | Control Variables |       |                          | Maths | Sport |
 | ----------------- | ----- | ------------------------ | ----- | ----- |
 | Age               | Maths | Correlation              | 1,000 | ,315  |
 |                   |       | ------------------------ | ----- | ----- |
 |                   |       | Significance (2-tailed)  | .     | ,096  |
 |                   |       | ------------------------ | ----- | ----- |
 |                   |       | df                       | 0     | 27    |
 |                   | ----- | ------------------------ | ----- | ----- |
 |                   | Sport | Correlation              | ,315  | 1,000 |
 |                   |       | ------------------------ | ----- | ----- |
 |                   |       | Significance (2-tailed)  | ,096  | .     |
 |                   |       | ------------------------ | ----- | ----- |
 |                   |       | df                       | 27    | 0     |
 | ----------------- | ----- | ------------------------ | ----- | ----- |

Correlations
 | ----------------- | ----- | ------------------------ | ----- | ----- |
 | Control Variables |       |                          | Age   | Maths |
 | ----------------- | ----- | ------------------------ | ----- | ----- |
 | Sport             | Age   | Correlation              | 1,000 | ,699  |
 |                   |       | ------------------------ | ----- | ----- |
 |                   |       | Significance (2-tailed)  | .     | ,000  |
 |                   |       | ------------------------ | ----- | ----- |
 |                   |       | df                       | 0     | 27    |
```

```
 |                   | ----- | ----------------------- | ----- | ----- |
 |                   | Maths | Correlation             | ,699  | 1,000 |
 |                   |       | ----------------------- | ----- | ----- |
 |                   |       | Significance (2-tailed) | ,000  | .     |
 |                   |       | ----------------------- | ----- | ----- |
 |                   |       | df                      | 27    | 0     |
 | ----------------- | ----- | ----------------------- | ----- | ----- |

Correlations
 | ----------------- | ----- | ----------------------- | ----- | ----- |
 | Control Variables |       |                         | Age   | Sport |
 | ----------------- | ----- | ----------------------- | ----- | ----- |
 | Maths             | Age   | Correlation             | 1,000 | ,348  |
 |                   |       | ----------------------- | ----- | ----- |
 |                   |       | Significance (2-tailed) | .     | ,064  |
 |                   |       | ----------------------- | ----- | ----- |
 |                   |       | df                      | 0     | 27    |
 |                   | ----- | ----------------------- | ----- | ----- |
 |                   | Sport | Correlation             | ,348  | 1,000 |
 |                   |       | ----------------------- | ----- | ----- |
 |                   |       | Significance (2-tailed) | ,064  | .     |
 |                   |       | ----------------------- | ----- | ----- |
 |                   |       | df                      | 27    | 0     |
 | ----------------- | ----- | ----------------------- | ----- | ----- |
```

On constate ainsi que les trois tests ne sont pas significatifs au seuil $\alpha = 5$ %.
On conserve l'hypothèse d'indépendance de deux variables sachant la troisième et ce dans les trois cas qui se présentent ici : (Maths, Sport)|Age, (Maths, Age)|Sport et (Sport, Age)|Maths.

On constate que les résultats à l'épreuve de sport et à celle de mathématiques sont indépendants si l'on tient compte de l'âge des élèves, ce qui est bien plus conforme à ce que l'on pouvait penser avant de réaliser cette expérience.

.....

Chapitre 4

Mesures de liaison non paramétriques

4.1. Mesures non paramétriques

Les techniques développées dans ce paragraphe permettent de tester l'indépendance ou de mesurer le degré de dépendance entre deux variables aléatoires appariées continues quelconques. Certains de ces tests permettent de faire apparaître des relations en tendance entre les variables et non pas des relations de type fonctionnel comme c'est le cas avec les techniques présentées précédemment. En effet celles-ci se placent systématiquement dans un cadre d'étude d'une liaison de type linéaire entre les différentes variables mises en jeu. Ainsi le non paramétrique apparaît à deux niveaux : en premier lieu dans le fait que l'on n'aura pas à vérifier d'hypothèse sur la loi des variables étudiées à partir du moment où celles-ci sont quantitatives, continues ou discrètes, ou qualitatives dans certains cas, et en deuxième lieu dans le fait que l'on ne fait d'hypothèse sur le type de relation pouvant exister entre les variables étudiées.
Comme à l'accoutumée, on utilisera, lorsque cela est possible des tests paramétriques conjointement aux tests non paramétriques.

4.2. La statistique $\rho_{\mathcal{S},n}$ de Spearman

Le coefficient de corrélation de Spearman est basé sur l'étude de la corrélation des rangs.

4.2.1. Cadre d'application

La mesure de la dépendance au sens de Spearman entre deux variables aléatoires continues X et Y sera notée $\rho_{\mathcal{S}}(X, Y)$. On note F la fonction de répartition de X et G celle de Y.

$\rho_{\mathcal{S}}(X,Y)$ pourra être estimé par la statistique de Spearman $\rho_{\mathcal{S},n}(X,Y)$ définie sur un $n-$échantillon $((X_1,Y_1),\ldots,(X_n,Y_n))$ indépendant et identiquement distribué suivant la loi de (X,Y).
La statistique $\rho_{\mathcal{S},n}(X,Y)$ permet de réaliser plusieurs tests :

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}}$$

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les valeurs prises par } X \text{ et } Y \text{ ont tendance à être concordantes.}}$$

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les valeurs prises par } X \text{ et } Y \text{ ont tendance à être discordantes.}}$$

On se reportera au paragraphe 4.4.2 pour une définition des termes concordants et discordants.
Ce coefficient a le même champ d'application que la statistique de Kendall $\tau_n(X,Y)$ qu'on lui préférera généralement.

4.2.2. Le coefficient de corrélation $\rho_{\mathcal{S}}(X,Y)$ de Spearman

Le coefficient de corrélation de Spearman est un nombre associé à (X,Y) qui sert à mesurer le degré de dépendance qui lie X et Y. Il peut être défini comme étant le coefficient de corrélation simple du couple aléatoire $(F(X),G(Y))$:

$$\rho_{\mathcal{S}}(X,Y) = \rho(F(X),G(Y)).$$

Il possède les propriétés suivantes :
- $-1 \leqslant \rho_{\mathcal{S}}(X,Y) \leqslant 1$,
- « X et Y indépendantes » implique $\rho_{\mathcal{S}}(X,Y) = 0$,
- $\rho_{\mathcal{S}}(X,Y) = 1$ (resp. $\rho_{\mathcal{S}}(X,Y) = -1$) si et seulement si il existe une fonction ϕ croissante (resp. décroissante) de $\mathbb{R}$ dans $\mathbb{R}$ telle que $Y = \phi(X)$,
- Si ϕ et ψ désignent deux fonctions croissantes de $\mathbb{R}$ dans $\mathbb{R}$ alors $\rho_{\mathcal{S}}(\phi(X),\psi(Y)) = \rho_{\mathcal{S}}(X,Y)$.

4.2.3. Estimation de $\rho_{\mathcal{S}}(X,Y)$

À chaque couple (x_i, y_i) de l'échantillon on associe le couple d'entiers (r_i, s_i) où r_i est le rang de x_i dans $x_1, \ldots, x_n$ et s_i est le rang de y_i dans $y_1, \ldots, y_n$. On appelle R_i la valeur aléatoire associée au rang d'un X_i, et S_i celle associée au rang d'un Y_i, puis on pose $R = (R_1, \ldots, R_n)$ et $S = (S_1, \ldots, S_n)$. On calcule alors simplement le coefficient de corrélation des rangs :

$$\begin{aligned}
\rho_{\mathcal{S},n}(X,Y) &= \widehat{\rho(R,S)} = \frac{\displaystyle\sum_{i=1}^{n} \left(R_i - \bar{R}\right) \times \left(S_i - \bar{S}\right)}{\widehat{\sigma_R}\widehat{\sigma_S}} \\
&= \frac{12}{n^3 - n} \sum_{i=1}^{n} \left[\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)\right] \\
&= 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} \left(R_i - S_i\right)^2 .
\end{aligned}$$

La réalisation $\rho_{\mathcal{S},n}(X,Y)(\boldsymbol{x},\boldsymbol{y})$ de la statistique $\rho_{\mathcal{S},n}(X,Y)$ sur l'échantillon $((x_1,y_1),\ldots,(x_n,y_n))$, noté $(\boldsymbol{x},\boldsymbol{y})$, possède les propriétés suivantes :
- $-1 \leqslant \rho_{\mathcal{S},n}(X,Y)(\boldsymbol{x},\boldsymbol{y}) \leqslant 1$,
- $\rho_{\mathcal{S},n}(X,Y)(\boldsymbol{x},\boldsymbol{y}) = 1$ si et seulement si $\forall i = 1 \ldots n, r_i = s_i$,
- $\rho_{\mathcal{S},n}(X,Y)(\boldsymbol{x},\boldsymbol{y}) = -1$ si et seulement si $\forall i = 1 \ldots n, r_i = n + 1 - s_i$.

On introduit souvent une variable auxiliaire $D^2_{\mathcal{S},n}(X,Y) = \sum_{i=1}^{n} \left(R_i - S_i\right)^2$. On a alors la relation suivante :

$$\rho_{\mathcal{S},n}(X,Y) = 1 - \frac{6D^2_{\mathcal{S},n}(X,Y)}{n^3 - n}.$$

Ainsi dans certaines tables vous trouverez les valeurs de $D_{\mathcal{S},n}$.

La statistique $\rho_{\mathcal{S},n}(X,Y)$ possède les propriétés suivantes :
- $\rho_{\mathcal{S},n}(X,Y)$ est un estimateur convergent avec biais de $\rho_{\mathcal{S}}(X,Y)$.

$$\mathbb{E}\left[\rho_{\mathcal{S},n}(X,Y)\right] = \rho_{\mathcal{S}}(X,Y) + \frac{3\left(\tau(X,Y) - \rho_{\mathcal{S}}(X,Y)\right)}{n+1}.$$

  Pour la définition de $\tau(X,Y)$, voir la section suivante. On note donc désormais $\rho_{\mathcal{S},n}(X,Y)$ par $\widehat{\rho_{\mathcal{S}}(X,Y)}$.
- Sous l'hypothèse nulle $\mathcal{H}_0$ « X et Y sont indépendantes » on a :
  - ⋆ $\mathbb{E}\left[\widehat{\rho_{\mathcal{S}}(X,Y)}\right] = 0$,
  - ⋆ $\mathrm{Var}\left[\widehat{\rho_{\mathcal{S}}(X,Y)}\right] = \dfrac{1}{n-1}$.
  - ⋆ Pour $4 \leqslant n \leqslant 19$, la distribution de $\widehat{\rho_{\mathcal{S}}(X,Y)}$ se déduit de celle de $D^2_{\mathcal{S},n}(X,Y)$ qui est tabulée.

- ⋆ Pour $20 \leqslant n \leqslant 30$, on utilise la variable $R_n(X,Y) = \sqrt{n-2}\dfrac{\widehat{\rho_{\mathbb{S}}(X,Y)}}{\sqrt{1-\widehat{\rho_{\mathbb{S}}(X,Y)}^2}}$ qui suit approximativement une loi de Student à $n-2$ degrés de liberté.
- ⋆ Pour $n > 30$, on a l'approximation normale $\sqrt{n-1}\widehat{\rho_{\mathbb{S}}(X,Y)} \approx \mathcal{N}(0,1)$.

On rappelle qu'un estimateur $\widehat{\rho_{\mathbb{S}}(X,Y)}$ de $\rho_{\mathbb{S}}(X,Y)$ est une variable aléatoire $\widehat{\rho_{\mathbb{S}}(X,Y)}$ qui prend pour valeur une estimation $\widehat{\rho_{\mathbb{S}}(X,Y)}(\boldsymbol{x},\boldsymbol{y})$ sur chaque échantillon $(\boldsymbol{x},\boldsymbol{y})$ de taille n. Ainsi un estimateur est une variable aléatoire alors qu'une estimation est un nombre réel construit pour être proche de la vraie valeur du paramètre inconnu $\rho_{\mathbb{S}}(X,Y)$.

4.2.4. Procédure de test

La statistique $\widehat{\rho_{\mathbb{S}}(X,Y)}$ permet de tester l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendants » contre plusieurs alternatives, cf plus haut.

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}}$$

Si l'on cherche un niveau de signification de α, on cherche r_α tel que :

$$\mathbb{P}\left[\widehat{\rho_{\mathbb{S}}(X,Y)} \geqslant r_\alpha\right] \leqslant \frac{\alpha}{2}.$$

On rejette l'hypothèse nulle $\mathcal{H}_0$ si, pour notre échantillon $(\boldsymbol{x},\boldsymbol{y})$, $\widehat{\rho_{\mathbb{S}}(X,Y)}(\boldsymbol{x},\boldsymbol{y}) \notin \,]-r_\alpha, r_\alpha[$.

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{\mathbb{S}}(X,Y) > 0, \text{ i.e. les valeurs prises par } X \text{ et } Y \text{ ont tendance à être concordantes.}}$$

Si l'on cherche un niveau de signification de α, on cherche r_α tel que :

$$\mathbb{P}\left[\widehat{\rho_{\mathbb{S}}(X,Y)} \geqslant r_\alpha\right] \leqslant \alpha.$$

On rejette l'hypothèse nulle $\mathcal{H}_0$ si, pour notre échantillon $(\boldsymbol{x},\boldsymbol{y})$, $\widehat{\rho_{\mathbb{S}}(X,Y)}(\boldsymbol{x},\boldsymbol{y}) \geqslant r_\alpha$.

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \rho_{\mathbb{S}}(X,Y) < 0, \text{ i.e. les valeurs prises par } X \text{ et } Y \text{ ont tendance à être discordantes.}}$$

Si l'on cherche un niveau de signification de α, on cherche r_α tel que :

$$\mathbb{P}\left[\widehat{\rho_{\mathcal{S}}(X,Y)} \leqslant r_\alpha\right] \leqslant \alpha.$$

On rejette l'hypothèse nulle $\mathcal{H}_0$ si, pour notre échantillon $(\boldsymbol{x},\boldsymbol{y})$, $\widehat{\rho_{\mathcal{S}}(X,Y)}(\boldsymbol{x},\boldsymbol{y}) \leqslant r_\alpha$.

Exemple 4.2.1.
On reprend les données de l'exemple 3.5.2. On remarque qu'il n'y a pas d'ex æquo. On peut donc calculer les estimations des coefficients de corrélation de Spearman $\rho_{\mathcal{S}}\,(\text{Maths}, \text{Sport})$, $\rho_{\mathcal{S}}\,(\text{Maths}, \text{Age})$ et $\rho_{\mathcal{S}}\,(\text{Sport}, \text{Age})$.

En utilisant SPSS 13.0 on obtient :

Correlations

			Maths	Sport	Age
Spearman's rho	Maths	Correlation Coefficient	1,000	,819(**)	,890(**)
		Sig. (2-tailed)	.	,000	,000
		N	30	30	30
	Sport	Correlation Coefficient	,819(**)	1,000	,818(**)
		Sig. (2-tailed)	,000	.	,000
		N	30	30	30
	Age	Correlation Coefficient	,890(**)	,818(**)	1,000
		Sig. (2-tailed)	,000	,000	.
		N	30	30	30

** Correlation is significant at the 0.01 level (2-tailed).

Tous les tests sont significatifs au seuil $\alpha = 5$ %. Il y a donc une association significative entre toutes les variables prises deux à deux. Voir le pragraphe 4.4.1 pour un calcul *exact* des $p-$valeurs avec SPSS 13.0 et le module Tests Exacts.

.....

4.2.5. Cas des ex æquo

Deux méthodes sont adaptées au cas où la variable n'est pas continue ou au cas où l'on a observé des ex æquo :

- on modifie la valeur de $\widehat{\rho_{\mathcal{S}}(X,Y)}$,
- on départage les ex æquo à l'aide d'une table de nombres aléatoires.

4.2.6. Statistique corrigée $\rho^{\star}_{\mathcal{S},n}$

On se donne un échantillon $(\boldsymbol{x},\boldsymbol{y})$ qui sont les réalisations d'un $n-$échantillon indépendant et identiquement distribué suivant la loi de (X,Y).
Les n valeurs observées $x_1,\ldots,x_n$ sont regroupées en h classes d'ex æquo $C_1,\ldots,C_h$. Certaines de ces classes peuvent ne comporter qu'un seul élément, si cet élément n'a pas d'ex æquo. On regroupe de même les valeurs y_i en k classes d'ex æquo $C_i^{'}$. Au couple de rangs réels (r_i,s_i) associé à (x_i,y_i) on substitue le couple de rangs fictifs $(r_i^{\star},s_i^{\star})$ où $r_i^{\star}$ est le rang moyen du groupe d'ex æquo auquel appartient x_i et $s_i^{\star}$ est le rang moyen du groupe d'ex æquo auquel appartient y_i. On note $d_i=\mathrm{Card}(C_i)$ et $d_i^{'}=\mathrm{Card}(C_i^{'})$.

On calcule alors $\delta(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}(d_i^3-d_i)$ et $\delta^{'}(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}((d_i^{'})^3-d_i^{'})$.

Dans le cas d'ex æquo la valeur prise par la statistique $\rho^{\star}_{\mathcal{S},n}$ est :

$$\begin{aligned}\rho^{\star}_{\mathcal{S},n}(X,Y)(\boldsymbol{x},\boldsymbol{y}) &= \frac{12}{\sqrt{(n^3-n-\delta)(n^3-n-\delta^{'})}}\sum_{i=1}^{n}\left[\left(r_i^{\star}-\frac{n+1}{2}\right)\left(s_i^{\star}-\frac{n+1}{2}\right)\right]\\ &= \frac{12\sum_{i=1}^{n}(r_i\times s_i)-3n(n+1)^2}{\sqrt{(n^3-n-\delta)(n^3-n-\delta^{'})}}.\end{aligned}$$

$\rho^{\star}_{\mathcal{S},n}(X,Y)$ est donc la variable aléatoire associée à $\rho^{\star}_{\mathcal{S},n}(X,Y)(\boldsymbol{x},\boldsymbol{y})$.
Lorsque $n>20$ et $(\delta(\boldsymbol{x},\boldsymbol{y})+\delta^{'}(\boldsymbol{x},\boldsymbol{y}))/n^3<0,1$ on utilise l'approximation normale :

$$\sqrt{n-1}\rho_{\mathcal{S},n}(X,Y)^{\star}\approx\mathcal{N}(0,1).$$

Dans les autres situations, il n'y a pas de table numérique.

Exemple 4.2.2.
On dispose de 7 couples de valeurs entières prises par le couple de variables aléatoires discrètes (X,Y). On a classé ces valeurs suivant les valeurs croissantes de X. Les couples (x_i,y_i), (r_i,s_i) et $(r_i^{\star},s_i^{\star})$ ont été reportés dans le tableau ci-dessous :

x_i	0	1	1	1	2	2	5
r_i	1	2	2	2	5	5	7
$r_i^\star$	1	3	3	3	5,5	5,5	7
y_i	3	2	0	3	4	5	3
s_i	3	2	1	3	6	7	3
$s_i^\star$	4	2	1	4	6	7	4

On obtient :

$$\begin{aligned}
\sum_{i=1}^{7}(r_i^\star \times s_i^\star) &= 124,5 \\
\delta &= (3^3-3)+(2^3-2)=30 \\
\delta' &= (3^3-3)=24 \\
\rho_{\mathcal{S}}^\star &= 0,485.
\end{aligned}$$

Comme $n = 7 < 20$, on ne peut utiliser l'approximation du paragraphe 4.2.6. On ne dispose pas non plus de table. On ne peut rien conclure quant à la significativité d'un test de corrélation partielle.

.....

4.2.7. Départition des ex æquo

On départage les ex æquo à l'aide d'une table de nombres aléatoires de la manière suivante : si on a r répétitions, on lit r nombres de suite dans la table de nombres aléatoires. La valeur des nombres permettent alors de départager les ex æquo, et donc de se ramener au cas où il n'y a pas de répétition.

Exemple 4.2.3.
Si par exemple on est face au cas où $x_1 = x_4 = x_7 = x_9 = 1$. On lit dans une table de nombres aléatoires 13407, 50270, 84980 et 22116. On décide alors $x_1 < x_9 < x_4 < x_7$ et l'on note $x_1 = 1^+$, $x_4 = 1^{+++}$, $x_7 = 1^{++++}$ et $x_9 = 1^{++}$. On aura alors, par exemple, $(x_7 - x_4) > 0$ et $(x_1 - x_9) < 0$.

.....

Exemple 4.2.4.
Appliquer ce procédé à l'exemple 4.2.2 ci-dessus.

.....

4.3. Corrélation partielle de Spearman

L'observation d'une corrélation entre deux variables X et Y peut être due à leur association avec une troisième Z.

4.3.1. Coefficient de corrélation partielle de Spearman $\rho_{\mathcal{S}}(X, Y|Z)$

Le coefficient de corrélation de Spearman est lié à l'analyse de la corrélation de Pearson des rangs d'un $n-$échantillon.
Si l'on a trois variables aléatoires X, Y et Z réelles et continues, on définit le coefficient de corrélation partielle de Spearman comme étant le coefficient de corrélation partielle de Pearson des rangs :

$$\frac{\rho_{\mathcal{S}}(X,Y) - \rho_{\mathcal{S}}(X,Z) \times \rho_{\mathcal{S}}(Y,Z)}{\sqrt{(1-\rho_{\mathcal{S}}(X,Z)^2)\,(1-\rho_{\mathcal{S}}(Y,Z)^2)}}.$$

4.3.2. Estimation de $\rho_{\mathcal{S}}\,(X, Y|Z)$

On considère trois variables aléatoires X, Y et Z ainsi qu'un $n-$échantillon indépendant identiquement distribué suivant la loi de (X, Y, Z), noté $((X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n))$. On suppose que (X, Y, Z) suit une loi continue.

On définit la statistique de corrélation partielle de Spearman $\rho_{\mathcal{S},n}\,(X, Y|Z)$, qui s'exprime à l'aide des statistiques de corrélation de Spearman des n couples (X_i, Y_i), des n couples (X_i, Z_i) et des n couples (Y_i, Z_i), par l'expression suivante :

$$\rho_{\mathcal{S},n}\,(X, Y|Z) = \frac{\rho_{\mathcal{S},n}(X,Y) - \rho_{\mathcal{S},n}(X,Z) \times \rho_{\mathcal{S},n}(Y,Z)}{\sqrt{(1-\rho_{\mathcal{S},n}(X,Z)^2)\,(1-\rho_{\mathcal{S},n}(Y,Z)^2)}}.$$

$\rho_{\mathcal{S},n}\,(X, Y|Z)$ a les propriétés suivantes :
- -1 $\leqslant \rho_{\mathcal{S},n}\,(X, Y|Z) \leqslant 1$,
- $\rho_{\mathcal{S},n}\,(X, Y|Z)$ est un estimateur convergent du coefficient de corrélation partielle de rang de Spearman $\rho_{\mathcal{S}}\,(X, Y|Z)$. On notera donc désormais $\rho_{\mathcal{S},n}\,(X, Y|Z)$ par $\rho_{\mathcal{S}}\,\widehat{(X, Y|Z)}$.

Sous l'hypothèse nulle $\mathcal{H}_0$ « $\rho_{\mathcal{S},n}(X, Y|Z) = 0$ », la distribution de $\rho_{\mathcal{S}}\,\widehat{(X, Y|Z)}$ est tabulée.

4.3.3. Méthode d'utilisation

Identique au cas non partiel.

Exemple 4.3.1.
On souhaite étudier la corrélation de Spearman des variables de l'exemple 3.5.2, Maths

et Sport, en éliminant l'influence de la variable Age. On calcule donc une estimation de la corrélation partielle de Spearman $\rho_S(\text{Maths}, \text{Sport}|\text{Age})$:

$$\widehat{\rho_S(\text{Maths}, \text{Sport}|\text{Age})}(\boldsymbol{x}) = \frac{0,819 - 0,890 * 0,818}{\sqrt{(1-0,890^2)(1-0,818^2)}} \approx 0,347.$$

On doit alors chercher dans une table adéquate la valeur critique pour $\alpha = 5$ %. Si vous parvenez à en trouver une, comparer ce résultat à celui obtenu dans le cadre paramétrique à l'exemple 3.5.2.

.....

4.4. La statistique τ_n de Kendall

4.4.1. Cadre d'application

La mesure de la dépendance au sens de Kendall entre deux variables aléatoires continues X et Y sera notée $\tau(X, Y)$.
Ce nombre pourra être estimé par la statistique de Kendall τ_n définie sur un échantillon $((x_1, y_1), \ldots, (x_n, y_n))$ indépendant et identiquement distribué.
La statistique τ_n permet de réaliser plusieurs tests :

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}}$$

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les valeurs prises par } X \text{ et } Y \text{ ont tendance à être concordantes.}}$$

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les valeurs prises par } X \text{ et } Y \text{ ont tendance à être discordantes.}}$$

4.4.2. Le coefficient de corrélation $\tau(X, Y)$ de Kendall

Considérons deux paires (x_i, y_i) et (x_j, y_j) issues de l'échantillon. Elles sont dites :

- concordantes si $(x_i - x_j)(y_i - y_j) > 0$, c'est-à-dire si l'on a simultanément $(x_i > x_j)$ et $(y_i > y_j)$ ou $(x_i < x_j)$ et $(y_i < y_j)$.
- discordantes si $(x_i - x_j)(y_i - y_j) < 0$, c'est-à-dire si l'on a simultanément $(x_i > x_j)$ et $(y_i < y_j)$ ou $(x_i < x_j)$ et $(y_i > y_j)$.

Considérons deux couples de variables aléatoires (X_1, Y_1) et (X_2, Y_2) de même loi que celle du couple étudié (X, Y).
Une concordance parfaite est telle que $X_2 > X_1 \longleftrightarrow Y_2 > Y_1$, c'est-à-dire :

$$\begin{aligned} \mathbb{P}\left[(X_2 > X_1) \text{ et } (Y_2 > Y_1)\right] + \mathbb{P}\left[(X_2 < X_1) \text{ et } (Y_2 < Y_1)\right] &= 1 \\ &\Updownarrow \\ \mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_1) > 0\right] &= 1. \end{aligned}$$

Une discordance parfaite est telle que $X_2 > X_1 \longleftrightarrow Y_2 < Y_1$, c'est-à-dire :

$$\begin{aligned} \mathbb{P}\left[(X_2 > X_1) \text{ et } (Y_2 < Y_1)\right] + \mathbb{P}\left[(X_2 < X_1) \text{ et } (Y_2 > Y_1)\right] &= 1 \\ &\Updownarrow \\ \mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_1) < 0\right] &= 1. \end{aligned}$$

On introduit donc $\tau^+(X, Y)$ et $\tau^-(X, Y)$ qui mesurent respectivement la concordance et la discordance du couple (X, Y) :

$$\begin{aligned} \tau^+(X, Y) &= \mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_1) > 0\right] \\ \tau^-(X, Y) &= \mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_1) < 0\right] \end{aligned}$$

Le coefficient $\tau(X, Y)$ de Kendall est défini par :

$$\tau(X, Y) = \tau^+(X, Y) - \tau^-(X, Y).$$

Il mesure le degré de concordance si $\tau(X, Y) > 0$ ou au contraire le degré de discordance si $\tau(X, Y) < 0$.

Il possède les propriétés suivantes :
- $-1 \leqslant \tau(X, Y) \leqslant 1$,
- X et Y indépendantes implique $\tau(X, Y) = 0$,
- $\tau(X, Y) = 1$ (resp. $\tau(X, Y) = -1$) si et seulement si il existe une fonction ϕ croissante (resp. décroissante) de $\mathbb{R}$ dans $\mathbb{R}$ telle que $Y = \phi(X)$,
- Si ϕ et ψ désignent deux fonctions croissantes de $\mathbb{R}$ dans $\mathbb{R}$ alors $\tau(\phi(X), \psi(Y)) = \tau(X, Y)$,
- Si (X, Y) suit une loi normale bivariée alors $\tau(X, Y)$ et $\rho(X, Y)$ sont liés par la relation : $\rho(X, Y) = \sin(\pi/2 \times \tau(X, Y))$.

4.4.3. Estimation de $\tau(X, Y)$

L'estimation se fait de manière « naturelle » : on commence par compter le nombre de paires de couples concordants c et le nombre de paires de couples discordants d dans l'échantillon $(x_1, y_1), \ldots, (x_n, y_n)$. Il apparaît ici une difficulté supplémentaire par rapport à la théorie. En effet la loi de (X, Y) est continue dont la probabilité que $X_1 = X_2$ ou que $Y_1 = Y_2$ est nulle. Mais dans l'échantillon il se peut néanmoins que vous observiez

plusieurs fois la même valeur. On note alors e le nombre de ces paires de couples. Ce sont celles pour lesquelles on a $(x_j - x_i)(y_j - y_i) = 0$. Dans un premier temps on supposera qu'il n'y a pas d'ex æquo.

$$\begin{aligned}
c &= \text{le nombre de paires de couples } (x_i, y_i), (y_i, y_j) \text{ tels que } (x_j - x_i)(y_j - y_i) > 0 \\
&\quad \text{avec } 1 \leqslant i < j \leqslant n. \\
d &= \text{le nombre de paires de couples } (x_i, y_i), (y_i, y_j) \text{ tels que } (x_j - x_i)(y_j - y_i) < 0 \\
&\quad \text{avec } 1 \leqslant i < j \leqslant n. \\
e &= \text{le nombre de paires de couples } (x_i, y_i), (y_i, y_j) \text{ tels que } (x_j - x_i)(y_j - y_i) = 0 \\
&\quad \text{avec } 1 \leqslant i < j \leqslant n.
\end{aligned}$$

Rappelons que pour le moment on suppose que $e = 0$. Toute paire de couples est forcément du type c ou du type d :

$$c + d = \frac{n(n-1)}{2}$$

qui est le nombre total de paires de couples qu'il est possible de faire.
On note C_n, D_n et E_n les variables aléatoires associées à c, d et e.
On définit alors :

$$\begin{aligned}
\tau_n(X, Y) &= \frac{2C_n - \dfrac{n(n-1)}{2}}{\dfrac{n(n-1)}{2}} \\
&= \frac{4C_n}{n(n-1)} - 1.
\end{aligned}$$

Sous l'hypothèse nulle $\mathcal{H}_0$ « X et Y sont indépendantes » la distribution de τ_n a les propriétés suivantes :

- $-1 \leqslant \tau_n \leqslant 1$, une valeur proche de 1 suggérant une forte concordance entre les valeurs prises par X et Y, une valeur proche de -1 suggérant une forte discordance entre les valeurs prises par X et Y.
- $\tau_n(X, Y)$ est un estimateur sans biais et convergent de $\tau(X, Y)$. On note donc désormais :

$$\tau_n(X, Y) = \widehat{\tau(X, Y)}$$

- Sous l'hypothèse nulle $\mathcal{H}_0$ « X et Y sont indépendantes », la distribution de $\widehat{\tau(X, Y)}$ a les caractéristiques suivantes :
  - ⋆ $\mathbb{E}\left[\widehat{\tau(X, Y)}\right] = 0$,
  - ⋆ $\mathrm{Var}\left[\widehat{\tau(X, Y)}\right] = \dfrac{2(2n+5)}{9n(n-1)}$,
  - ⋆ $\dfrac{\widehat{\tau(X,Y)}}{\sqrt{\mathrm{Var}\left[\widehat{\tau(X,Y)}\right]}} \approx \mathcal{N}(0, 1)$.

Pour réaliser des tests avec des effectifs inférieurs à 20, on se reportera donc à des tables spécifiques ou l'on utilisera un logiciel disposant de statistiques exactes.

4.4.4. Procédure de test

La statistique $\widehat{\tau_(X,Y)}$ permet de tester l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes » contre plusieurs alternatives, cf plus haut.

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}}$$

Si l'on cherche un niveau de signification de α, on cherche t_α tel que :

$$\mathbb{P}\left[\widehat{\tau(X,Y)} \geqslant t_\alpha\right] \leqslant \frac{\alpha}{2}.$$

On rejette l'hypothèse nulle $\mathcal{H}_0$ si $\widehat{\tau(X,Y)} \notin \,]-t_\alpha, t_\alpha[$.

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \tau(X,Y) > 0, \text{ i.e. les valeurs prises par } X \text{ et } Y \text{ ont tendance à être concordantes.}}$$

Si l'on cherche un niveau de signification de α, on cherche t_α tel que :

$$\mathbb{P}\left[\widehat{\tau(X,Y)} \geqslant t_\alpha\right] \leqslant \alpha.$$

On rejette l'hypothèse nulle $\mathcal{H}_0$ si $\widehat{\tau(X,Y)} \geqslant t_\alpha$.

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : \tau(X,Y) < 0, \text{ i.e. les valeurs prises par } X \text{ et } Y \text{ ont tendance à être discordantes.}}$$

Si l'on cherche un niveau de signification de α, on cherche t_α tel que :

$$\mathbb{P}\left[\widehat{\tau(X,Y)} \leqslant t_\alpha\right] \leqslant \alpha.$$

On rejette l'hypothèse nulle $\mathcal{H}_0$ si $\widehat{\tau(X,Y)} \leqslant t_\alpha$.

Exemple 4.4.1.
On reprend les données de l'exemple 3.5.2. On remarque qu'il n'y a pas d'ex æquo. On peut donc calculer les estimations des coefficients de corrélation de Kendall $\tau\,(\text{Maths}, \text{Sport})$, $\tau\,(\text{Maths}, \text{Age})$ et $\tau\,(\text{Sport}, \text{Age})$.

En utilisant SPSS 13.0 on obtient :

Correlations

			Maths	Sport	Age
Kendall's tau_b	Maths	Correlation Coefficient	1,000	,605(**)	,701(**)
		Sig. (2-tailed)	.	,000	,000
		N	30	30	30
	Sport	Correlation Coefficient	,605(**)	1,000	,600(**)
		Sig. (2-tailed)	,000	.	,000
		N	30	30	30
	Age	Correlation Coefficient	,701(**)	,600(**)	1,000
		Sig. (2-tailed)	,000	,000	.
		N	30	30	30

** Correlation is significant at the 0.01 level (2-tailed).

Tous les tests sont significatifs au seuil $\alpha = 5$ %. Il y a donc une association significative entre toutes les variables prises deux à deux.

On calcule alors les $p-$valeurs exactes associées aux estimations des corrélations de Spearman et de Kendall. Or l'obtention de ces valeurs serait trop longue pour nos moyens de calcul actuels et l'on doit donc utiliser une méthode de Monte-Carlo pour trouver un intervalle de confiance à 99 % sur ces $p-$valeurs. Si le seuil α du test n'appartient pas l'intervalle obtenu on conclut comme à l'accoutumée en fonction de la place de l'intervalle par rapport à α. Si α appartient à l'intervalle obtenu, il faut relancer le calcul en augmentant la taille de l'échantillon utilisé pour calculer l'intervalle de confiance sur la $p-$valeur par la méthode de Monte Carlo.

Symmetric Measures

```
 |                      | Monte Carlo Sig.                                |
 |                      | ---------- | ---------------------------------- |
 |                      | Sig.       | 99% Confidence Interval            |
 |                      |            | -------------------- | ----------- |
 |                      |            | Lower Bound          | Upper Bound |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Kendall's tau-b      | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Spearman Correlation | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Pearson's R          | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
c Based on 10000 sampled tables with starting seed 2000000.

Symmetric Measures
 | -------------------- | ---------------------------------------------- |
 |                      | Monte Carlo Sig.                                |
 |                      | ---------- | ---------------------------------- |
 |                      | Sig.       | 99% Confidence Interval            |
 |                      |            | -------------------- | ----------- |
 |                      |            | Lower Bound          | Upper Bound |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Kendall's tau-b      | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Spearman Correlation | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Pearson's R          | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
c Based on 10000 sampled tables with starting seed 1314643744.

Symmetric Measures
 | -------------------- | ---------------------------------------------- |
 |                      | Monte Carlo Sig.                                |
 |                      | ---------- | ---------------------------------- |
 |                      | Sig.       | 99% Confidence Interval            |
 |                      |            | -------------------- | ----------- |
 |                      |            | Lower Bound          | Upper Bound |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Kendall's tau-b      | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Spearman Correlation | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
 | Pearson's R          | ,000(c)    | ,000                 | ,000        |
 | -------------------- | ---------- | -------------------- | ----------- |
c Based on 10000 sampled tables with starting seed 1502173562.
```

Ici, pour $\alpha = 5$ %, tous les intervalles sont entièrement inclus dans $[0, \alpha]$ et donc tous les tests sont significatifs au seuil $\alpha = 5$ %.

.....

4.4.5. Cas des ex æquo

Deux méthodes sont adaptées au cas où la variable n'est pas continue ou au cas où l'on a observé des ex æquo :

- on modifie la valeur de $\widehat{\tau(X, Y)}$,
- on départage les ex æquo à l'aide d'une table de nombres aléatoires.

4.4.6. Statistique corrigée $\tau_n^\star$

Les n valeurs observées $x_1, \ldots, x_n$ sont regroupées en h classes d'ex æquo $C_1, \ldots, C_h$. Certaines de ces classes peuvent ne comporter qu'un seul élément, si cet élément n'a pas d'ex æquo. On regroupe de même les valeurs y_j en k classes d'ex æquo $C_j^{'}$. On note $d_i = \text{Card}(C_i)$ et $d_j^{'} = \text{Card}(C_j^{'})$. Enfin $d = \sum_{i=1}^{h} d_i(d_i - 1)$ et $d^{'} = \sum_{j=1}^{k} d_j^{'}(d_j^{'} - 1)$.

On calcule alors :

$$\begin{aligned} s^\star &= 1 \times \text{Nombre de concordants} + (-1) \times \text{Nombre de discordants} \\ &\quad +0 \times \text{Nombre de cas d'égalité.} \end{aligned}$$

Puis on pose :

$$\tau^\star = \frac{2s^\star}{\sqrt{(n(n-1) - d)(n(n-1) - d^{'})}}.$$

$\tau_n^\star$ est alors la variable aléatoire associée à $\tau^\star$ et $S_n^\star$ celle associée à $s^\star$.
Lorsque $n > 10$, $d/n^2 < 0,1$ et $d^{'}/n^2 < 0,1$ on utilise l'approximation normale :

$$\frac{\sqrt{18} S_n^\star}{\sqrt{n(n-1)(2n+5) - \delta - \delta^{'}}} \approx \mathcal{N}(0, 1)$$

où $\delta = \sum_{i=1}^{h} d_i(d_i - 1)(2d_i + 5)$ et $\delta^{'} = \sum_{j=1}^{k} d_j^{'}(d_j^{'} - 1)(2d_j^{'} + 5)$.

Exemple 4.4.2.
Appliquer ce procédé à l'exemple 4.2.2.

.....

4.4.7. Départition des ex æquo

On départage les ex æquo à l'aide d'une table de nombres aléatoires de la manière suivante : si on a r répétitions, on lit r nombres de suite dans la table de nombres aléatoires. La valeur des nombres permettent alors de départager les ex æquo, et donc de se ramener au cas où il n'y a pas de répétition.

Exemple 4.4.3.
Si par exemple on est face au cas où $x_1 = x_4 = x_7 = x_9 = 1$. On lit dans une table de nombres aléatoires 13407, 50270, 84980 et 22116. On décide alors $x_1 < x_9 < x_4 < x_7$ et l'on note $x_1 = 1^+$, $x_4 = 1^{+++}$, $x_7 = 1^{++++}$ et $x_9 = 1^{++}$. On aura alors, par exemple, $(x_7 - x_4) > 0$ et $(x_1 - x_9) < 0$.

.....

Exemple 4.4.4.
Appliquer ce procédé à l'exemple 4.2.2.

.....

4.5. Corrélation partielle de Kendall

L'observation d'une corrélation entre deux variables X et Y peut être due à leur association avec une troisième Z.

4.5.1. Coefficient de corrélation partiel de Kendall $\tau(X, Y|Z)$

On commence par chercher les probabilités de concordance et discordance pour deux triplets $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2)$ indépendants et identiquement distribués suivant la loi de (X, Y, Z). On suppose que la loi de (X, Y, Z) est continue.

- La concordance de X et Z sans tenir compte de Y est :

$$p_{1,\bullet} = \mathbb{P}\left[(X_2 - X_1)(Z_2 - Z_1) > 0\right].$$

- La discordance de X et Z sans tenir compte de Y est :

$$p_{2,\bullet} = \mathbb{P}\left[(X_2 - X_1)(Z_2 - Z_1) < 0\right].$$

- La concordance de Y et Z sans tenir compte de X est :

$$p_{\bullet,1} = \mathbb{P}\left[(Y_2 - Y_1)(Z_2 - Z_1) > 0\right].$$

- La discordance de Y et Z sans tenir compte de X est :

$$p_{\bullet,2} = \mathbb{P}\left[(Y_2 - Y_1)(Z_2 - Z_1) < 0\right].$$

- La concordance simultanée de X et Y avec Z est :

$$p_{1,1} = \mathbb{P}\left[(X_2 - X_1)(Z_2 - Z_1) > 0 \text{ et } (Y_2 - Y_1)(Z_2 - Z_1) > 0\right].$$

- La probabilité que, simultanément, (X_1, Z_1) et (X_2, Z_2) soient concordants et que (Y_1, Z_1) et (Y_2, Z_2) soient discordants est :

$$p_{2,1} = \mathbb{P}\left[(X_2 - X_1)(Z_2 - Z_1) < 0 \text{ et } (Y_2 - Y_1)(Z_2 - Z_1) > 0\right].$$

- La probabilité que, simultanénement, (X_1, Z_1) et (X_2, Z_2) soient discordants et que (Y_1, Z_1) et (Y_2, Z_2) soient concordants est :

$$p_{1,2} = \mathbb{P}\left[(X_2 - X_1)(Z_2 - Z_1) > 0 \text{ et } (Y_2 - Y_1)(Z_2 - Z_1) < 0\right].$$

- La discordance simultanée de X et Y avec Z est :

$$p_{2,2} = \mathbb{P}\left[(X_2 - X_1)(Z_2 - Z_1) < 0 \text{ et } (Y_2 - Y_1)(Z_2 - Z_1) < 0\right].$$

Le coefficient de corrélation partielle de Kendall est alors :

$$\tau(X, Y|Z) = \frac{p_{1,1}p_{2,2} - p_{1,2}p_{2,1}}{\sqrt{(p_{1,\bullet}p_{2,\bullet}p_{\bullet,1}p_{\bullet,2})}}.$$

Ses propriétés sont les suivantes :
- $-1 \leqslant \tau(X, Y|Z) \leqslant 1$,
- on peut exprimer le coefficient de corrélation partielle de Kendall du couple (X, Y) connaissant Z, $\tau(X, Y|Z)$, à l'aide des coefficients de corrélation de Kendall $\tau(X, Y)$, $\tau(X, Z)$ et $\tau(Y, Z)$ des couples (X, Y), (X, Z) et (Y, Z) :

$$\tau(X, Y|Z) = \frac{\tau(X, Y) - \tau(X, Z) \times \tau(Y, Z)}{\sqrt{(1 - \tau(X, Z)^2)\,(1 - \tau(Y, Z)^2)}},$$

- « (X, Y) et Z indépendants » donne $\tau(X, Y|Z) = \tau(X, Y)$,
- « X, Y et Z indépendantes » donne $\tau(X, Y|Z) = 0$.

4.5.2. Notations

On considère trois variables aléatoires X, Y et Z ainsi qu'un $n-$échantillon indépendant identiquement distribué de (X, Y, Z), noté $((X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n))$. On suppose que (X, Y, Z) suit une loi continue.

On va à nouveau utiliser la notion de paire de couples concordants ou discordants mais en la modifiant pour tenir compte de la présence de la troisième variable aléatoire Z. On dispose de $n(n-1)/2$ paires de triplets $(X_i, Y_i, Z_i), (X_j, Y_j, Z_j)$ avec $1 \leqslant i < j \leqslant n$.

On considère donc pour commencer :
- $C_{x,z}$ l'ensemble des paires de triplets pour lesquels X et Z sont concordants.
- $D_{x,z}$ l'ensembles des paires de triplets pour lesquels X et Z sont discordants.

On note $N_{1\bullet}$ le nombre d'éléments de $C_{x,z}$ et $N_{2\bullet}$ le nombre d'éléments $D_{x,z}$.

Puis on s'intéresse à :
- $C_{y,z}$ l'ensemble des paires de triplets pour lesquels Y et Z sont concordants.
- $D_{y,z}$ l'ensemble des paires de triplets pour lesquels Y et Z sont discordants.

On note $N_{\bullet 1}$ le nombre d'éléments de $C_{y,z}$ et $N_{\bullet 2}$ le nombre d'éléments $D_{y,z}$.

Pour prendre en compte simultanément les variables X et Y on partage les $n(n-1)/2$ paires de triplets en quatre groupes disjoints.

- $C_{1,1}$ correspond donc aux paires de triplets pour lesquelles à la fois X et Z sont concordants et Y et Z sont concordants :

$$C_{1,1} = C_{x,z} \cap C_{y,z}.$$

  On note $N_{1,1}$ le nombre d'éléments de $C_{1,1}$.
- $C_{1,2}$ correspond donc aux paires de triplets pour lesquelles à la fois X et Z sont concordants et Y et Z sont discordants :

$$C_{1,2} = C_{x,z} \cap D_{y,z}.$$

  On note $N_{2,1}$ le nombre d'éléments de $C_{2,1}$.
- $C_{2,1}$ correspond donc aux paires de triplets pour lesquelles à la fois X et Z sont discordants et Y et Z sont concordants :

$$C_{2,1} = D_{x,z} \cap C_{y,z}.$$

  On note $N_{1,2}$ le nombre d'éléments de $C_{1,2}$.
- $C_{2,2}$ correspond donc aux paires de triplets pour lesquelles à la fois X et Z sont discordants et Y et Z sont discordants :

$$C_{2,2} = D_{x,z} \cap D_{y,z}.$$

  On note $N_{2,2}$ le nombre d'éléments de $C_{2,2}$.

On résume la situation dans le tableau suivant :

	$C_{x,z}$	$D_{x,z}$	Total
$C_{y,z}$	$N_{1,1}$	$N_{1,1}$	$N_{\bullet,1}$
$D_{y,z}$	$N_{1,1}$	$N_{1,1}$	$N_{\bullet,2}$
Total	$N_{1,\bullet}$	$N_{2,\bullet}$	$n(n-1)/2$

4.5.3. Estimateur de $\tau\left(X, Y|Z\right)$

On définit la statistique de corrélation partielle de Kendall $\tau_{XY|Z,n}$ par l'expression suivante :

$$\tau_{XY|Z,n} = \frac{N_{1,1}N_{2,2} - N_{1,2}N_{2,1}}{\sqrt{(N_{1,\bullet}N_{2,\bullet}N_{\bullet,1}N_{\bullet,1})}}.$$

On peut exprimer cette statistique à l'aide des statistiques de corrélation de Kendall des n couples (X_i, Y_i), des n couples (X_i, Z_i) et des n couples (Y_i, Z_i) :

$$\tau_{XY|Z,n} = \frac{\tau_n(X,Y) - \tau_n(X,Z) \times \tau_n(Y,Z)}{\sqrt{(1-\tau_n(X,Z)^2)\,(1-\tau_n(Y,Z)^2)}}.$$

$\tau_{XY|Z,n}$ a les propriétés suivantes :
- -1 $\leqslant \tau_{XY|Z,n} \leqslant 1$,
- $\tau_{XY|Z,n}$ est un estimateur convergent du coefficient de corrélation partiel de rang de Kendall $\tau(X, Y|Z)$.

  On notera donc désormais $\tau_{XY|Z,n}$ par $\tau(\widehat{X, Y|Z})$.

Sous l'hypothèse nulle « $\mathcal{H}_0$: $\tau(X, Y|Z) = 0$ », la distribution de $\tau(\widehat{X, Y|Z})$ a les caractéristiques suivantes :
- $\mathbb{E}\left[\tau(\widehat{X, Y|Z})\right] = 0$, $\mathrm{Var}\left[\tau(\widehat{X, Y|Z})\right] = \dfrac{2(2n+5)}{9n(n-1)}$,
- elle est symétrique $\mathbb{P}\left[\tau(\widehat{X, Y|Z}) \leqslant t\right] = \mathbb{P}\left[\tau(\widehat{X, Y|Z}) \geqslant t\right]$, pour tout $t \in \mathbb{R}$,
- Pour $n < 30$, on se reporte à une table. Sinon on utilise l'approximation suivante :

$$3\sqrt{\frac{(n^2-n)}{(4n+10)}}\tau(\widehat{X, Y|Z}) \approx \mathcal{N}\left(0,1\right).$$

4.5.4. Méthode d'utilisation

Identique au cas non partiel.

Exemple 4.5.1.
On souhaite étudier la corrélation de Kendall des variables Maths et Sport en éliminant

l'influence de la variable Age. On calcule donc une estimation de la corrélation partielle de Kendall $\tau(\text{Maths}, \text{Sport}|\text{Age})$:

$$\widehat{\tau(\text{Maths}, \text{Sport}|\text{Age})}(\boldsymbol{x}) = \frac{0,605 - 0,701 * 0,600}{\sqrt{(1-0,701^2)(1-0,600^2)}} \approx 0,323.$$

On trouve dans la table que la valeur critique pour $\alpha = 5$ % est de $0,253$ qui est inférieure à notre estimation et de ce fait le test est significatif. On rejette donc l'hypothèse nulle $\mathcal{H}_0$ d'indépendance Maths et Sport lorsque l'on élimine l'influence de la variable Age et on décide qu'il y a une association significative de Maths et Sport lorsque l'on élimine l'influence de la variable Age. Comparer ce résultat à celui obtenu dans le cadre paramétrique à l'exemple 3.5.2. Lequel des deux faut-il préférer ?

.....

4.6. Test du χ^2 d'indépendance

4.6.1. Les classes

Ce test peut être réalisé quelque soit la loi du couple (X, Y), discrète, continue ou mixte. Toutefois il est **peu puissant**.
Le domaine $\mathcal{D}_X$ des valeurs de X est divisé en h classes adjacentes $C_{i,\bullet}$. Le domaine $\mathcal{D}_Y$ des valeurs de Y est divisé en k classes adjacentes $C_{\bullet,}$.
Ainsi les valeurs de l'échantillon se répartissent dans les $h \times k$ classes $C_{i,j} = C_{i,\bullet} \times C_{\bullet,j}$.
Le nombre de couples (x_l, y_l) tels que $(x_l, y_l) \in C_{i,j}$ est noté $n_{i,j}$. Le nombre de valeurs x_i appartenant à la classe $C_{i,\bullet}$ est noté $n_{i,\bullet}$ et le nombre de valeurs x_j appartenant à la classe $C_{\bullet,j}$ est noté $n_{i,\bullet}$. On a $n_{i,\bullet} = \sum_{j=1}^{k} n_{i,j}$ et $n_{\bullet,j} = \sum_{i=1}^{h} n_{i,j}$.

$X \backslash Y$	1	$\cdots$	j	$\cdots$	k	Total
1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,k}$	$n_{1,\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,k}$	$n_{i,\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
h	$n_{h,1}$	$\cdots$	$n_{h,j}$	$\cdots$	$n_{h,k}$	$n_{h,\bullet}$
Total	$n_{\bullet,1}$	$\cdots$	$n_{\bullet,j}$	$\cdots$	$n_{k,\bullet}$	n

4.6.2. Statistique du test

Pour tester l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes », on introduit la statistique χ_n^2 suivante :

$$\chi_n^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(N_{i,j} - \dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}\right)^2}{\dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}}$$

où $N_{i,j}$ est le nombre aléatoire de couple (X_i, Y_j) qui dans un échantillon indépendant identiquement distribué de taille n appartient à la classe $C_{i,j}$, $N_{i,\bullet} = \sum_{j=1}^{k} N_{i,j}$ et $N_{\bullet,j} = \sum_{i=1}^{h} N_{i,j}$.

Sous l'hypothèse nulle $\mathcal{H}_0$ la loi de χ_n^2 converge vers la loi du χ^2 à $(h-1)(k-1)$ degrés de liberté lorsque que n tend vers $+\infty$.

Il est raisonnable de faire cette approximation lorsque $n \geqslant 50$ et que chaque effectif théorique $n_{i,j}^{\star} = n\frac{n_{i,\bullet}}{n}\frac{n_{\bullet,j}}{n}$ est supérieur à 5. Dans la situation où les effectifs théoriques ne seraient pas tous supérieurs à 5, il est nécessaire de procéder au regroupement de classes adjacentes, C_i et C_{i+1} ou $C_j^{'}$ et $C_{j+1}^{'}$ pour $1 \leqslant i \leqslant h-1$ et $1 \leqslant j \leqslant k-1$.

4.6.3. Méthode d'utilisation

La valeur critique χ_α^2 du test au seuil α % est le plus petit nombre χ_α^2 tel que $\mathbb{P}\left[\chi_n^2 \geqslant \chi_\alpha^2\right] = \alpha$ où χ_n^2 suit une loi $\chi_{(h-1)(k-1)}^2$.
Si la valeur $\chi_{\text{obs},n}^2$, qui est la réalisation de la statistique du test sur l'échantillon :

$$\chi_{\text{obs},n}^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{i,j} - \dfrac{n_{i,\bullet} \times n_{\bullet,j}}{n}\right)^2}{\dfrac{n_{i,\bullet} \times n_{\bullet,j}}{n}}$$

est supérieure à χ_α^2 on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance.

Ce test pose plusieurs difficultés :

- Comment faire les regroupements en classe lorsque cela s'avère nécessaire, par exemple si les variables étudiées sont continues ?
- Les conditions d'application sont très contraignantes. Un effectif total n **supérieur à 50** et des fréquences d'apparition toutes supérieures à 5. Dans le livre de J. Bouyer, [3], ainsi que dans celui de G. Pupion et P.-C. Pupion, [13], il est indiqué que le test est encore utilisable si les effectifs théoriques sont tous supéreurs à 3. J. Bouyer, [3], évoque même la possibilité de se contenter du fait qu'il y ait moins de 20 % des cellules

pour lesquelles les effectifs théoriques soient inférieurs à 5. Néanmoins tous les auteurs s'entendent pour dire que si dans une telle situation vous obteniez des valeurs proches de la significativité, il est impératif de compléter l'étude par l'utilisation de certains des autres tests présentés ici.

- La nécessité de fondre plusieurs modalités en une seule pour que les conditions d'applications, mentionnées au paragraphe ci-dessus, soient remplies modifie les variables sur lesquelles porte le test.
- Ce test ne tient pas compte de l'éventuelle présence d'un ordre sur les lignes ou les colonnes du tableau de contingence. Si l'on peut ordonner les modalités de l'un des deux facteurs, on préférera utiliser un test de Kruskal-Wallis et si l'on peut ordonner les modalités des deux facteurs on utilisera un test de Jonckheere-Terpstra. On pourra alors, si l'on rejette l'hypothèse nulle $\mathcal{H}_0$: « Le facteur X n'a pas d'effet sur la réponse Y », étudier les raisons à l'origine de la non-indépendance à l'aide d'un test post-hoc. Le cas d'un tableau à deux lignes et k colonnes ou à h lignes et deux colonnes peut également être étudié à l'aide d'un test de Mann-Whitney.

4.6.4. Correction de Yates

Lorsque l'on étudie l'indépendance de deux variables et que certaines des fréquences attendues sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes » sont inférieures à 5, on peut corriger la statistique du test pour prendre en compte cette situation. Attention il faut néanmoins que toutes les fréquences attendues soient supérieures à 3. La correction de Yates est une correction de continuité qui consiste à utiliser la statistique de test modifiée de la manière suivante :

$$\chi_n^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{\left(\left|N_{i,j} - \dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}\right| - \dfrac{1}{2}\right)^2}{\dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}}$$

où $N_{i,j}$ est le nombre aléatoire de couple (X_i, Y_j) qui dans un échantillon indépendant identiquement distribué de taille n appartient à la classe $C_{i,j}$, $N_{i,\bullet} = \sum_{j=1}^{k} N_{i,j}$ et $N_{\bullet,j} = \sum_{i=1}^{h} N_{i,j}$.

Attention, on n'utilisera la correction que si l'une des fréquences attendues est strictement inférieure à 5. Si au contraire elles sont toutes supérieures ou égales à 5, on montre que la correction de Yates ne modifie que peu la valeur de la réalisation de la statistique du test. Comme le signale J Bouyer, [3], ce point de vue, c'est-à-dire le fait de réserver cette correction à de petits échantillons, n'est pas partagé par tous les auteurs. Si l'un des effectifs théoriques est inférieur à 3, on n'a pas d'autre solution que d'appliquer le test exact de Fisher décrit au paragraphe 4.7. Plus généralement, P. Dagnélie, [6], conseille l'utilisation systématique du test exact de Fisher lorsque l'effectif total n est inférieur ou égal à 40.

4.6.5. Étude des résidus

Lorsque **l'hypothèse d'indépendance est vérifiée**, les termes dont les carrés sont les contributions à la valeur χ_n^2 :

$$\frac{N_{i,j} - \dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}}{\sqrt{\dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}}}$$

sont approximativement des variables aléatoires qui suivent des lois normales centrées réduites. On a montré qu'une meilleure approximation de la loi normale centrée réduite est obtenue si l'on considère les valeurs ci-dessus et que l'on les divise par des estimations des écarts types correspondants :

$$\sqrt{\left(1 - \frac{N_{i,\bullet}}{n}\right)\left(1 - \frac{N_{\bullet,j}}{n}\right)}.$$

On obtient finalement les **écarts réduits** :

$$\frac{N_{i,j} - \dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}}{\sqrt{\left(1 - \dfrac{N_{i,\bullet}}{n}\right)\left(1 - \dfrac{N_{\bullet,j}}{n}\right)\dfrac{N_{i,\bullet} \times N_{\bullet,j}}{n}}}$$

On peut alors étudier les écarts réduits comme on le ferait pour des résidus obtenus après une régression par exemple : étude de la normalité, via une droite de Henry par exemple, identification d'éventuelles valeurs aberrantes ou non représentatives, voir le cours sur ce sujet.

Remarques :

- Il existe une version « exacte » de ce test.
- Pour des modélisations plus complexes et, par exemple, l'étude de corrélations partielles, on utilise le **modèle log-linéaire**, voir le cours associé.

4.7. Test exact de Fisher

Pour présenter le principe de ce test, on commence par étudier le cas de deux variables X et Y, ayant deux modalités, puis on traite le cas général. Pour plus de détails sur les tests exacts sur les tableaux de contingence, on se reportera à l'article de A. Agresti [1].

4.7.1. Deux variables à deux modalités

Soit X et Y deux variables quantitatives discrètes ou qualitatives. On suppose dans ce paragraphe que X et Y ne peuvent prendre que 2 valeurs différentes. On se donne un

$n-$échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendant et identiquement distribué suivant la loi de (X, Y) ainsi qu'un échantillon $(\boldsymbol{x}, \boldsymbol{y})$ formé d'une réalisation de chaque (X_i, Y_i), $1 \leqslant i \leqslant n$. Enfin on note $\boldsymbol{x}$ l'échantillon des réalisations de X et $\boldsymbol{y}$ l'échantillon des réalisations de Y. On considère le tableau des effectifs $n_{i,j}$ suivant chacune des 2 modalités de X et des 2 modalités de Y apparaissant dans l'échantillon $(\boldsymbol{x}, \boldsymbol{y})$:

$X \backslash Y$	1	2	Total
1	$n_{1,1}$	$n_{1,2}$	$n_{1,\bullet}$
2	$n_{2,1}$	$n_{2,2}$	$n_{2,\bullet}$
Total	$n_{\bullet,1}$	$n_{\bullet,2}$	n

Dans ce tableau les marges $(n_{1,\bullet}, n_{2,\bullet}, n_{\bullet,1}, n_{\bullet,2})$ n'apportent pas d'information sur l'éventuelle dépendance de X et de Y. En effet elles n'indiquent que la répartition des effectifs entre les deux modalités de X, indépendemment de la valeur de Y, et la répartition des effectifs entre les deux modalités de Y, indépendemment de la valeur de X. Ce sont les valeurs prisent par $n_{1,1}, n_{1,2}, n_{2,1}$ et $n_{2,2}$ qui servent pour étudier la dépendance de X et de Y.
L'idée du test exact de Fisher est de considérer l'ensemble Γ des tableaux ayant les mêmes marges $(n_{1,\bullet}, n_{2,\bullet}, n_{\bullet,1}, n_{\bullet,2})$ que le tableau des effectifs observés ci-dessus :

$$\Gamma_{(n_{\bullet,1}, n_{\bullet,2})}^{(n_{1,\bullet}, n_{2,\bullet})} = \left\{ \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}, \text{ avec } a+b = n_{1,\bullet}, c+d = n_{2,\bullet}, a+c = n_{\bullet,1} \text{ et } b+d = n_{\bullet,2} \right\}.$$

Les marges $(n_{1,\bullet}, n_{2,\bullet}, n_{\bullet,1}, n_{\bullet,2})$ étant fixées, la connaissance de a détermine par différence les valeurs de b, c et d :

$$\Gamma_{(n_{\bullet,1}, n_{\bullet,2})}^{(n_{1,\bullet}, n_{2,\bullet})} = \left\{ \gamma(a) = \begin{array}{|c|c|} \hline a & n_{1,\bullet} - a \\ \hline n_{\bullet,1} - a & n_{\bullet,2} - n_{\bullet,1} + a \\ \hline \end{array}, \text{ avec } 0 \leqslant a \leqslant \inf(n_{1,\bullet}, n_{\bullet,1}) \right\}.$$

On montre que, sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes », la probabilité d'obtenir un tableau de type $\gamma(a)$ suit une loi hypergéométrique, voir [7] et [6] :

$$\begin{aligned}
\mathbb{P}[n_{1,1} = a] &= \frac{\mathrm{C}_{n_{1,\bullet}}^{a} \times \mathrm{C}_{n_{2,\bullet}}^{b}}{\mathrm{C}_{n_{\bullet,2}+n_{\bullet,1}}^{n_{1,\bullet}}} \\
&= \frac{\mathrm{C}_{a+c}^{a} \times \mathrm{C}_{b+d}^{b}}{\mathrm{C}_{a+b+c+d}^{a+b}} \\
&= \frac{(a+c)! \times (b+d)! \times (a+b)! \times (c+d)!}{a! \times b! \times c! \times d! \times (a+b+c+d)!} \\
&= \frac{n_{\bullet,1}! \times n_{\bullet,2}! \times n_{1,\bullet}! \times n_{2,\bullet}!}{a! \times (n_{\bullet,1} - a)! \times (n_{1,\bullet} - a)! \times (n_{\bullet,2} - n_{\bullet,1} + a)! \times n!},
\end{aligned}$$

où $z!$ désigne la factorielle du nombre entier positif z qui vaut $z! = z \times (z-1) \times \cdots \times 1$ avec la convention $0! = 1$.

Procédure de test :
On teste l'hypothèse suivante :

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}}$$

Le principe du test consiste à évaluer la probabilité de rencontrer une distribution aussi anormale ou plus anormale que celle que l'on a observée.

Supposons par exemple que $\frac{a}{a+c} \leqslant \frac{b}{b+d}$. On additionne alors les probabilités d'obtenir des tableaux ayant des valeurs de $n_{1,1}$ comprises en 0 et a.

Supposons par contre que $\frac{a}{a+c} \geqslant \frac{b}{b+d}$. On additionne alors les probabilités d'obtenir des tableaux ayant des valeurs de $n_{1,1}$ comprises en b et $\inf(a+b, a+c)$.

Le test bilatéral de niveau α conduit alors au rejet de l'hypothèse nulle $\mathcal{H}_0$ quand la probabilité totale calculée par la méthode ci-dessus est inférieure ou égale à $\alpha/2$.

Il s'agit en fait d'un niveau de signification maximal α car la distribution hypergéométrique sur laquelle repose le test est discrète.

- La dénomination « exact » du test vient du fait que l'on ne fait appel à aucune approximation pour calculer la $p-$valeur du test.
- Le niveau de signification du test est d'au plus α, il se peut qu'il soit beaucoup plus faible, ce qui fait de ce test un test conservatif.
- Ce test peut également être utilisé pour comparer deux pourcentages.

Exemple 4.7.1. On cherche à savoir s'il existe une association entre la couleur des yeux, clairs ou foncés, et la couleur des cheveux, blonds ou bruns, chez les êtres humains. On considère un groupe de 20 individus de sexe féminin ou masculin.

Cheveux \ Yeux	Clairs	Foncés	Total
Blonds	5	1	6
Bruns	0	14	14
Total	5	15	20

On souhaite tester l'hypothèse :

$$\boxed{\mathcal{H}_0 : \text{La couleur des yeux est indépendante de celle des cheveux}}$$

contre

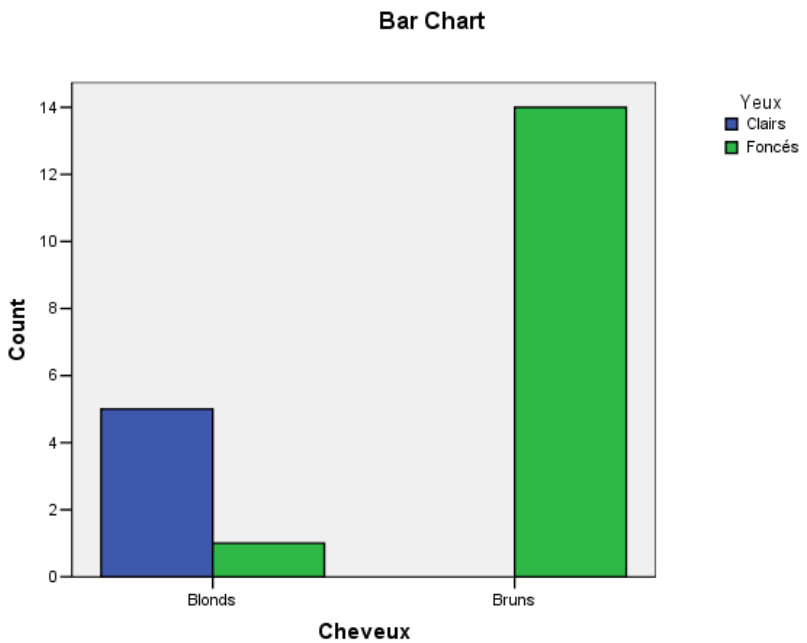
$$\boxed{\mathcal{H}_1 : \text{La couleur des yeux n'est pas indépendante de celle des cheveux.}}$$

On est dans la situation où $\frac{a}{a+c} \geqslant \frac{b}{b+d}$ car $\frac{5}{5} = 1 \geqslant \frac{1}{15}$. On cherche ainsi la $p-$valeur associée au test en additionnant la probabilité d'obtenir des tableaux pour lesquels $n_{1,1}$ est compris entre 5 et $\inf(6,5) = 5$. La $p-$valeur est donc égale à la probabilité d'obtenir un tableau pour lequel $n_{1,1} = 5$ connaissant les marges $(6, 14, 5, 15)$:

$$\begin{aligned}
\mathbb{P}[n_{1,1} = 5] &= \frac{(5+0)! \times (1+14)! \times (5+1)! \times (0+14)!}{5! \times 1! \times 0! \times 14! \times 20!} \\
&= \frac{5! \times 15! \times 6! \times 14!}{5! \times 1! \times 0! \times 14! \times 20!} \\
&= \frac{5 \times 4 \times 3 \times 2 \times 1 \times 6}{20 \times 19 \times 18 \times 17 \times 16} \\
&= \frac{1}{2 \times 19 \times 17 \times 4} \\
&= \frac{1}{2584}
\end{aligned}$$

Donc la $p-$valeur associée au test vaut : $\frac{1}{2584}$. Or pour un niveau de signification de $\alpha = 5$ %, c'est-à-dire, $\alpha = \frac{1}{20}$ on a obtenu une $p-$valeur inférieure ou égale au seuil du test : le test est significatif. On rejette l'hypothèse nulle $\mathcal{H}_0$ et on décide l'hypothèse alternative $\mathcal{H}_1$: « La couleur des yeux n'est pas indépendante de celle des cheveux ». Il existe une association significative au seuil $\alpha = 5$ % entre la couleur des yeux et celle des cheveux.

Voici ce que l'on obtient lorsque l'on utilise SPSS 13.0 :



Cheveux * Yeux Crosstabulation

Count

		Yeux		Total
		Clairs	Foncés	
Cheveux	Blonds	5	1	6
	Bruns	0	14	14
Total		5	15	20

Chi-Square Tests

	Value	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Fisher's Exact Test		,000	,000
N of Valid Cases	20		

Ces résultats sont en accord avec les calculs que l'on a réalisés à la main ci-dessus : le test

est significatif au seuil $\alpha = 5$ %.

```
Chi-Square Tests
 | ------------------------ | --------- | -- | --------------------- |
 |                          | Value     | df | Asymp. Sig. (2-sided) |
 | ------------------------ | --------- | -- | --------------------- |
 | Pearson Chi-Square       | 15,556(b) | 1  | ,000                  |
 | ------------------------ | --------- | -- | --------------------- |
 | Continuity Correction(a) | 11,429    | 1  | ,001                  |
 | ------------------------ | --------- | -- | --------------------- |
 | N of Valid Cases         | 20        |    |                       |
 | ------------------------ | --------- | -- | --------------------- |
a Computed only for a 2x2 table
b 3 cells (75,0%) have expected count less than 5.
The minimum expected count is 1,50.
```

On constate que SPSS propose un test du χ^2, avec ou sans la correction de Yates mentionnée au paragraphe 4.6.4. SPSS signale que 75 % des cellules ont des effectifs attendus de moins de 5 ce qui rend impossible l'utilisation du test du χ^2 dans notre situation : on ne pouvait qu'utiliser le test exact de Fisher.

.....

4.7.2. Deux variables ayant un nombre fini quelconque de modalités

On reprend l'idée du paragraphe 4.7.1 ci-dessus et on l'étend au cas où les deux variables étudiées ont un nombre fini quelconque, mais supérieur à deux, de modalités. Cette extension a été réalisée en premier par G. H. Freeman et J.H. Halton en 1951, voir [9], c'est pourquoi ce test est aussi parfois appelé test de Freeman-Halton ou de Fisher-Freeman-Halton.

Soit X et Y deux variables quantitatives discrètes ou qualitatives. On se donne un $n-$échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendant et identiquement distribué suivant la loi de (X, Y) ainsi qu'un échantillon $(\boldsymbol{x}, \boldsymbol{y})$ formé d'une réalisation de chaque (X_i, Y_i), $1 \leqslant i \leqslant n$. Enfin on note $\boldsymbol{x}$ l'échantillon des réalisations de X et $\boldsymbol{y}$ l'échantillon des réalisations de Y. On considère le tableau des effectifs $n_{i,j}$ suivant chacune des h modalités de X et des k modalités de Y apparaissant dans l'échantillon $(\boldsymbol{x}, \boldsymbol{y})$:

$X \backslash Y$	1	$\cdots$	j	$\cdots$	k	Total
1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,k}$	$n_{1,\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,k}$	$n_{i,\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
h	$n_{h,1}$	$\cdots$	$n_{h,j}$	$\cdots$	$n_{h,k}$	$n_{h,\bullet}$
Total	$n_{\bullet,1}$	$\cdots$	$n_{\bullet,j}$	$\cdots$	$n_{k,\bullet}$	n

Dans ce tableau les marges $(n_{1,\bullet}, \ldots, n_{h,\bullet}, n_{\bullet,1}, \cdots, n_{\bullet,k})$ n'apportent pas d'information sur l'éventuelle dépendance de X et de Y. En effet elles n'indiquent que la répartition des effectifs entre les h modalités de X, indépendemment de la valeur de Y, et la répartition des effectifs entre les k modalités de Y, indépendemment de la valeur de X. Ce sont les valeurs prisent par $n_{1,1}, n_{1,2}, \ldots, n_{1,k}, n_{2,1}, \ldots, n_{2,k}, \ldots, n_{i,j}, \ldots, n_{h,k-1}$ et $n_{h,k}$ qui servent pour étudier la dépendance de X et de Y.
L'idée du test exact de Fisher est de considérer l'ensemble Γ des tableaux ayant les mêmes marges $(n_{1,\bullet}, \ldots, n_{h,\bullet}, n_{\bullet,1}, \ldots, n_{\bullet,k})$ que le tableau des effectifs observés ci-dessus :

$$\Gamma^{(n_{1,\bullet},\ldots,n_{h,\bullet})}_{(n_{\bullet,1},\ldots,n_{\bullet,k})} = \left\{ \begin{array}{ccccc} n_{1,1} & \cdots & n_{1,j} & \cdots & n_{1,k} \\ \vdots & & \vdots & & \vdots \\ n_{i,1} & \cdots & n_{i,j} & \cdots & n_{i,k} \\ \vdots & & \vdots & & \vdots \\ n_{h,1} & \cdots & n_{h,j} & \cdots & n_{h,k} \end{array} \right\}, \text{ avec } \left\{ \begin{array}{rcl} n_{1,1} + \ldots + n_{1,k} & = & n_{1,\bullet} \\ \vdots & = & \vdots \\ n_{1,1} + \ldots + n_{h,1} & = & n_{h,\bullet} \\ n_{h,1} + \ldots + n_{h,k} & = & n_{\bullet,1} \\ \vdots & = & \vdots \\ n_{1,k} + \ldots + n_{h,k} & = & n_{\bullet,k} \end{array} \right\}.$$

Les marges $(n_{1,\bullet}, \ldots, n_{h,\bullet}, n_{\bullet,1}, \ldots, n_{\bullet,k})$ étant fixées, la connaissance de $hk - h - k + 1 = (h-1)(k-1)$ valeurs détermine celle de toutes les autres.
On montre que, sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes », la probabilité d'obtenir un tableau γ appartenant à Γ suit une loi hypergéométrique généralisée, voir [7] :

$$\begin{aligned} \mathbb{P}[\gamma] &= \frac{(n_{1,1} + \ldots + n_{1,k})! \times \cdots \times (n_{h,1} + \ldots + n_{h,k})! \times (n_{1,1} + \ldots + n_{h,1})! \times (n_{1,k} + \ldots + n_{h,k})!}{n_{1,1}! \times \cdots \times n_{h,k}!(n_{1,1} + \ldots + n_{h,k})!} \\ &= \frac{\prod_{i=1}^{h} n_{i,\bullet}! \prod_{j=1}^{k} n_{\bullet,j}!}{\left(\prod_{i=1}^{h}\left(\prod_{j=1}^{k} n_{i,j}!\right)\right)\left(\sum_{i=1}^{h}\sum_{j=1}^{k} n_{i,j}\right)!}, \end{aligned}$$

où $z!$ désigne la factorielle du nombre entier positif z qui vaut $z! = z \times (z-1) \times \cdots \times 1$ avec la convention $0! = 1$.

Procédure de test :
On teste l'hypothèse suivante :

$$\boxed{\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}}$$

contre

$$\boxed{\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}}$$

Le principe du test consiste à évaluer la probabilité de rencontrer une distribution aussi anormale ou plus anormale dans un tableau γ de Γ que celle que l'on a observée γ_{obs} :

$$\begin{aligned} p_2 &= \sum_{D(\gamma) \geqslant D(\gamma_{obs})} \mathbb{P}[\gamma] \\ &= \mathbb{P}[D(\gamma) \geqslant D(\gamma_{obs})]. \end{aligned}$$

où la fonction $\gamma \to D(\gamma)$ est définie pour tout $\gamma \in \Gamma$ par :

$$D(\gamma) = -2 \log \left(\frac{(2\pi)^{\frac{(h-1)(k-1)}{2}}}{N^{\frac{(hk-1)}{2}}} \prod_{i=1}^{h} n_{i,\bullet}^{\frac{h-1}{2}} \prod_{j=1}^{h} n_{\bullet,j}^{\frac{k-1}{2}} \mathbb{P}[\gamma] \right).$$

On définit alors la région critique du test comme la partie $\Gamma^\star$ de l'ensemble de référence Γ :

$$\Gamma^\star = \{y \in \Gamma \text{ tels que } D(y) \geqslant D(x)\}.$$

Le test bilatéral de niveau α conduit alors au rejet de l'hypothèse nulle $\mathcal{H}_0$ quand la probabilité totale p_2 calculée par la méthode ci-dessus est inférieure ou égale à α[1].

Il s'agit en fait d'un niveau de signification maximal α car la distribution hypergéométrique généralisée sur laquelle repose le test est discrète.

- La dénomination « exact » du test vient du fait que l'on ne fait appel à aucune approximation pour calculer la $p-$valeur du test.
- Le niveau de signification du test est d'au plus α, il se peut qu'il soit beaucoup plus faible, ce qui fait de ce test un test conservatif.
- Ce test peut également être utilisé pour comparer k pourcentages sur l populations.

[1] On remarque que la $p-$valeur calculée au paragraphe 4.7.1 pour le cas d'un test exact de Fisher sur un tableau de taille 2×2 n'est pas obtenue de la même manière qu'ici. Ceci est dû au fait que dans le cas du paragraphe 4.7.1, on pouvait déduire toutes les valeurs du tableaux en connaissant uniquement ses marges et une de valeurs des effectif $n_{1,1}$, $n_{1,2}$, $n_{2,1}$ ou $n_{2,2}$ et ainsi ranger « naturellement » les tableaux dans l'ordre croissant ce qui permettait alors de définir facilement ce que l'on entendait par une distribution aussi ou plus anormale que celle du tableau observée.

Chapitre 5

Tests de multinormalité

5.1. Tests de multinormalité

5.1.1. Utilisation de tests multiples unidimensionnels

On montre, voir le livre de J.-Y. Ouvrard [11], qu'un **vecteur aléatoire** $\boldsymbol{X}$ à n composantes $(X_1, \ldots, X_n)$ est **gaussien**, c'est-à-dire qu'il suit une loi normale de dimension n, parfois également appelée loi multinormale sur $\mathbb{R}^n$, **si et seulement si toutes** les combinaisons linéaires de ses composantes, voir le paragraphe 3.4.1 pour une définition de cette notion mathématique, sont des variables aléatoires gaussiennes sur $\mathbb{R}$, c'est-à-dire qu'elles suivent des lois normales unidimensionnelles.

Le résultat ci-dessus permet donc apparemment de ramener au cas de la normalité d'une **variable aléatoire réelle** Y l'étude de la normalité d'un **vecteur aléatoire** $\boldsymbol{X}$. Malheureusement, si l'on voulait se servir dans la pratique de ce résultat, il faudrait tester une infinité de combinaisons linéaires de composantes de $\boldsymbol{X}$! C'est bien sûr impossible non seulement pour des raisons de temps de calcul mais aussi pour des problèmes de risque de première espèce α. En effet une telle démarche, si l'on pouvait l'accomplir se solderait par un risque de première espèce global α_{global} de 100 % quelque soit le risque individuel α_{ind} non nul fixé au départ[1].

On a utilisé dans les exemples 3.3.1, 3.5.1 et 3.5.2 la propriété exposée ci-dessus pour vérifier sommairement la validité de l'hypothèse de multinormalité d'un vecteur aléatoire $\boldsymbol{X}$. Ne pouvant réaliser des tests pour toutes les combinaisons linéaires des composantes $(X_1, \ldots, X_n)$ des vecteurs que l'on a étudiés, on s'est contenté de concentrer notre étude sur certaines combinaisons linéaires :

[1] En effet on doit faire une **infinité** de tests et $\alpha_{global} = \lim_{n\to\infty} (1 - (1 - \alpha_{ind})^n) = 1$.

- Les composantes du vecteur $\boldsymbol{X}$: $X_1, \ldots, X_n$.
- Les sommes des composantes du vecteur $\boldsymbol{X}$: $X_1 + X_2$, $X_1 + X_3$, ..., $X_1 + X_n$, $X_2 + X_3$, ..., $X_i + X_j$, ..., $X_{n-1} + X_n$, où $1 \leqslant i < j \leqslant n$.

Si le vecteur $\boldsymbol{X}$ n'a que deux composantes (X_1, X_2) on peut rajouter la combinaison linéaire suivante :

- La différence des composantes du vecteur $\boldsymbol{X}$: $X_1 - X_2$.

Ainsi si l'on considère un vecteur à deux composantes (X_1, X_2) on réalise 4 tests de normalité individuels, celui de X_1, celui de X_2, celui de $X_1 + X_2$ et celui de $X_1 - X_2$. On adapte[2] donc la valeur du risque de première espèce individuel α_{ind} pour obtenir un risque de première espèce global de α_{global} de la manière suivante :

$$\begin{aligned} \alpha_{global} &= 1 - (1 - \alpha_{ind})^4, \\ \alpha_{ind} &= 1 - \sqrt[4]{(1 - \alpha_{global})}. \end{aligned}$$

Par exemple si on souhaite obtenir un $\alpha_{global} = 5$ % on trouve $\alpha_{ind} = 1,27$ %.

Plus généralement, supposons que n est supérieur ou égal 3 et considérons un vecteur à n composantes $(X_1, \ldots, X_n)$. On réalise alors $2^n - 1$ tests de normalité individuels, ceux de X_1, ..., X_n, ceux de $X_1 + X_2$, ..., $X_{n-1} + X_n$, ceux des sommes de trois composantes, ..., ceux des sommes de j composantes, ... et celui de la somme de toutes les composantes $X_1 + \cdots + X_n$. On adapte, voir la note de bas de page numéro 2, donc la valeur du risque de première espèce individuel α_{ind} pour obtenir un risque de première espèce global de α_{global} de la manière suivante :

$$\begin{aligned} \alpha_{global} &= 1 - (1 - \alpha_{ind})^{\frac{k(k+1)}{2}}, \\ \alpha_{ind} &= 1 - \sqrt[\frac{k(k+1)}{2}]{(1 - \alpha_{global})}. \end{aligned}$$

Par exemple si on souhaite obtenir un $\alpha_{global} = 5$ % on trouve :

[2]On utilise ici une procédure de Sidak [7], il ne s'agit que d'une possiblité parmi d'autres, on aurait aussi bien pu se servir de l'inégalité de Bonferroni, voir [7] et [6] et la recommandation de H.C. Thode dans [14] à ce sujet. On utilise ces procédures car le calcul du niveau exact du test est très complexe.

k	$\frac{k(k+1)}{2}$	α_{ind} %
3	7	0,6391
4	15	0,3201
5	31	0,1602
10	1023	0,0050
20	1048576	$0,4891 \times 10^{-5}$
50	1125899906842624	$0,4556 \times 10^{-14}$
100	1267650600228229401496703205376	$0,4046 \times 10^{-29}$

$\alpha_{ind} = 1,27$ %.
On voit que cette procédure n'est adaptée que pour des cas pratiques de petite dimension $n \leqslant 5$. Dans ce cas elle permet d'évaluer rapidement la possible multinormalité de la distribution sans avoir à recourir à des outils spécifiques qui sont plus complexes et donc plus long, voir le paragraphe 5.1.2 suivant.

Exemple 5.1.1.
Pour des applications de cette procédure voir les exemples 3.3.1, 3.5.1 et 3.5.2.

.....

5.1.2. Une test basée sur la modification du test de Shapiro-Wilk unidimensionnel

Malkovitch (1971), Malkovitch and Afifi (1973) ont présenté un test de multinormalité pour un vecteur aléatoire $\boldsymbol{X}$ de dimension n basé sur une modification du test de Shapiro-Wilk. Puisque le vecteur $\boldsymbol{X}$ est normal si et seulement si toutes les combinaisons linéaires des composantes $(X_1, \ldots, X_n)$ de $\boldsymbol{X}$ sont normales le test de multinormalité est accepté si

$$\inf_{\substack{\text{Sur les combinaisons linéaires } Y \\ \text{des composantes de } \boldsymbol{X}}} W_Y \geqslant \text{Valeur critique du test,}$$

où W_Y est la valeur de la réalisation de la statistique de Shapiro-Wilk calculée pour Y.

On détermine alors une combinaison linéaire susceptible de réaliser le minimum apparaissant dans le membre de droite de l'équation ci-dessus.

Les détails étant complexes, on n'exposera pas ici la procédure du test. Pour plus de détails voir l'ouvrage de H.C. Thode [14].

Malheureusement, ce test n'est pas disponible sous Minitab, ni sous la plupart des logiciels de statistique.

5.1.3. Asymétrie et aplatissement multivariés de Mardia

Mardia (1970) et (1974) a proposé un test de mulitinormalité basé sur l'asymétrie et l'aplatissement. Ce test est par exemple disponible sous S-Plus, R et la plupart des logiciels de statistique. On se donne donc un vecteur aléatoire $\boldsymbol{X}$ de dimension p.

On considère un $n-$échantillon $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ indépendant et identiquement distribué suivant la loi de $\boldsymbol{X}$. On se donne également $\boldsymbol{x}$ une réalisation du $n-$échantillon $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$.

5.1.4. Centrer-réduire.

Dans cette section on considère un vecteur aléatoire $\boldsymbol{X}$ qui admet une moyenne $\boldsymbol{\mu}$ et une matrice de variance-covariance $\boldsymbol{\Sigma}$.

On commence par calculer les valeurs centrées-réduites[3] :

$$\boldsymbol{Z_i} = \boldsymbol{\Sigma}^{\frac{-1}{2}} \left(\boldsymbol{X_i} - \boldsymbol{\mu}\right).$$

Si l'on suppose que le la loi de $\boldsymbol{X}$ est multinormale, $\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, les vecteurs aléatoires $\boldsymbol{Z_i}$ suivent alors une loi normale multidimensionnelle « centrée-réduite » $\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I_p}\right)$.

On calcule alors les rayons au carré[4] de la manière suivante :

$$\begin{aligned} R_i^2 &= \boldsymbol{Z_i}^T \boldsymbol{Z_i} \\ &= (\boldsymbol{X_i} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X_i} - \boldsymbol{\mu}), \end{aligned}$$

$(\boldsymbol{X_i} - \boldsymbol{\mu})^T$ est la transposée de $(\boldsymbol{X_i} - \boldsymbol{\mu})$.
Si l'on suppose que le la loi de $\boldsymbol{X}$ est multinormale, $\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, la variable aléatoire R_i^2 suit alors une loi du χ^2 à m degrés de liberté.

Or les paramètres de la loi, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, sont généralement inconnus. On doit donc les estimer pour pouvoir calculer les valeurs centrées réduites ainsi que les rayons au carré. On procède donc ainsi :

-i- On estime la moyenne $\boldsymbol{\mu}$ de $\boldsymbol{X}$ à l'aide de l'estimateur défini au paragraphe 3.3.2 ; on obtient la valeur $\widehat{\boldsymbol{\mu}}\left(\boldsymbol{x}\right)$. Puis on estime la matrice de variance-covariance $\boldsymbol{\Sigma}$ à l'aide de l'estimateur défini au paragraphe 3.3.2 ; on obtient la valeur $\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{x}\right)$.

[3] « Scaled residuals » en anglais. Cette transformation est la version multivariée du centrer-réduire que vous connaissez en dimension 1 : on soustrait la moyenne $\boldsymbol{\mu}$ puis on divise par la racine carrée de la variance notée $\boldsymbol{\Sigma}^{-\frac{1}{2}}$.

[4] « Squared radii » en anglais.

-ii- On estime alors les valeurs centrées réduites et les rayons au carré :

$$\begin{aligned}
\widehat{\boldsymbol{Z_i}}(\boldsymbol{x}) &= \widehat{\boldsymbol{\Sigma}}(\boldsymbol{x})^{-\frac{1}{2}}(\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}(\boldsymbol{x})). \\
\widehat{R_i^2}(\boldsymbol{x}) &= \widehat{\boldsymbol{Z_i}}(\boldsymbol{x})^T \widehat{\boldsymbol{Z_i}}(\boldsymbol{x}) \\
&= (\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}(\boldsymbol{x}))^T \widehat{\boldsymbol{\Sigma}}(\boldsymbol{x})^{-1}(\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}(\boldsymbol{x})).
\end{aligned}$$

De nombreuses procédures d'études de la multinormalité se basent sur les propriétés des rayons au carré et des rayons au carré généralisés définis par :

$$\begin{aligned}
R_{i,j} &= \boldsymbol{Z_i}^T \boldsymbol{Z_i} \\
&= (\boldsymbol{X_i} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X_i} - \boldsymbol{\mu}),
\end{aligned}$$

On obtient des estimations de ces variables ainsi :

$$\begin{aligned}
\widehat{R_{ij}}(\boldsymbol{x}) &= \widehat{\boldsymbol{Z_i}}(\boldsymbol{x})^T \widehat{\boldsymbol{Z_j}}(\boldsymbol{x}) \\
&= (\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}(\boldsymbol{x}))^T \widehat{\boldsymbol{\Sigma}}(\boldsymbol{x})^{-1}(\boldsymbol{x_j} - \widehat{\boldsymbol{\mu}}(\boldsymbol{x})).
\end{aligned}$$

On remarque que $R_{i,i} = R_i^2$ et que $\widehat{\boldsymbol{R_{i,i}}}(\boldsymbol{x}) = \widehat{\boldsymbol{R_i}}^2(\boldsymbol{x})$.

5.1.5. Asymétrie et aplatissement

La mesure d'asymétrie multivariée d'un vecteur aléatoire $\boldsymbol{X}$ est :

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij}^3.$$

On l'estime en utilisant les estimateurs de $R_{i,j}$ du paragraphe 5.1.4 :

$$\widehat{b_{1,p}}(\boldsymbol{x}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \widehat{R_{ij}}(\boldsymbol{x})^3.$$

La mesure d'aplatissement multivariée d'un vecteur aléatoire $\boldsymbol{X}$ est :

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} R_{ii}^2.$$

On l'estime en utilisant les estimateurs de $R_{i,j}$ du paragraphe 5.1.4 :

$$\widehat{b_{2,p}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \widehat{R_{ii}}(\boldsymbol{x})^2.$$

5.1.6. Étude sous l'hypothèse de multinormalité de $\boldsymbol{X}$.

Si l'on se place sous l'hypothèse nulle $\mathcal{H}_0$: « La loi de $\boldsymbol{X}$ est multinormale », $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, on a les résultats suivants :

-i- L'espérance de $\widehat{b_{1,p}}$ est :

$$\mathbb{E}\left[\widehat{b_{1,p}}\right] = \frac{p(p+2)((n+1)(p+1)-6)}{(n+1)(n+3)}.$$

-ii- On a un résultat sur la distribution asymptotique de $\widehat{b_{1,p}}$:

$$\frac{n\widehat{b_{1,p}}}{6} \text{ suit asymptotiquement une loi du } \chi^2_{\frac{p(p+1)(p+2)}{6}}.$$

-iii- L'espérance de $\widehat{b_{2,p}}$ est :

$$\mathbb{E}\left[\widehat{b_{2,p}}\right] = \frac{p(p+2)(n-1)}{(n+1)}.$$

-iv- La variance de $\widehat{b_{2,p}}$ est :

$$\mathrm{Var}\left[\widehat{b_{2,p}}\right] = \frac{8p(m+2)(n-3)(n-m-1)(n-m+1)}{(n+1)^2(n+3)(n+5)}.$$

-v- On a un résultat sur la distribution asymptotique de $\widehat{b_{2,p}}$:

$$\widehat{b_{2,p}} \text{ suit asymptotiquement une loi } \mathcal{N}\left(p(p+2), \frac{8p(p+2)}{n}\right).$$

Certaines des valeurs critiques sont tabulées, sinon l'on utilise l'approximation suivante pour $\widehat{b_{1,p}}$. On pose :

$$A = \frac{(p+1)(n+1)(n+3)}{((n+1)(p+1)-6)}\frac{\widehat{b_{1,p}}}{6}.$$

A suit alors une loi approximativement une loi du χ^2 à $\frac{p(p+1)(p+2)}{6}$ degrés de liberté.
Ce test est disponible sous SAS et R.

Il est recommandé par H.C. Thode dans [14] pour sa bonne puissance.

5.1.7. Un test basé sur la fonction caractéristique empirique

Il s'agit d'un test proposé par Henze et Zirkler (1990).

On mentionne ici ce test pour ces bonnes propriétés de puissance. En particulier celles de $T_{0,5}$.

La statistique du test est alors :

$$\begin{aligned} D_{n,\beta} &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\exp\left(-\frac{\beta^2}{2}\left\|\widehat{\boldsymbol{Z}_i}-\widehat{\boldsymbol{Z}_j}\right\|^2\right)+(1+2\beta^2)^{-\frac{m}{2}} \\ &\quad -\frac{2(1+\beta^2)^{-\frac{m}{2}}}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2}{2(1+\beta^2)}r_j^2\right). \end{aligned}$$

Ce test est disponible sous SAS.

Deuxième partie

Feuilles de travaux pratiques

Chapitre 6

Feuille de Travaux Pratiques n° 1

Quelques tests non paramétriques.

Les données des deux premiers exercices sont inspirées du livre de G. Pupion et P.-C. Pupion, éditions Economica.

Exercice I.1. Économie

On mesure un indice économique sur onze entreprises. On est amené à se poser la question suivante : « Peut-on considérer que la médiane associée à cet indice est nulle ? »

Entreprise	Indice
1	1
2	4
3	10
4	20
5	0,5
6	-3
7	-7
8	5
9	4
10	3
11	1

. . . . . .

Exercice I.2. Étude d'activité

On dispose de la variation du chiffre d'affaires de 20 entreprises dans un même secteur d'activité. Le chiffre d'affaires dans ce secteur d'activité est-il resté stable ?

Entreprise	1	2	3	4	5	6	7	8	9	10
x_i	-25	-2156	4525	2697	-379	404	-1123	-1733	-2658	-477
Entreprise	11	12	13	14	15	16	17	18	19	20
x_i	-3568	-12071	165	269	-4306	-983	-582	-1897	-1412	662

. . . . . .

Exercice I.3. Âge des arbres

On souhaite évaluer une nouvelle méthode permettant de déterminer l'âge d'un arbre sans avoir à l'abattre. Pour ce faire on sacrifie 11 arbres pour lesquels on a réalisé les deux types mesures : estimation de l'âge de l'arbre à l'aide de la méthode dont on souhaite tester l'efficacité puis calcul de l'âge exact de l'arbre après abattage. On a reporté les données dans le tableau ci-dessous :

Arbre	1	2	3	4	5	6	7	8	9	10	11
Âge estimé avant abattage	29	28	42	32	22	32	28	21	30	23	39
Âge réel après abattage	25	24	38	27	19	28	24	22	26	19	34

Peut-on se fier aux résultats de la nouvelle méthode proposée pour estimer l'âge d'un arbre ?

Exercice I.4. Hauteurs des arbres

On souhaite comparer la hauteur des arbres de deux types de hêtraies. Peut-on dire, à l'aide des mesures de taille exprimées en m et que l'on a reportées dans le tableau ci-dessous, qu'il y a une différence entre les tailles moyennes des arbres des deux hêtraies ?

Type 1	Type 2	Type 1	Type 2
23,4	22,5	24,4	22,9
24,6	23,7	24,9	24,0
25,0	24,4	26,2	24,5
26,3	25,3	26,8	26,0
26,8	26,2	26,9	26,4
27,0	26,7	27,6	26,9
27,7	27,4		28,5

On dispose désormais de mesures de taille, exprimées en m, provenant d'une troisième hêtraie.

Type 3	18,9	21,1	21,2	22,1	22,5	23,6	24,5	24,6	26,2	26,7

Y a-t-il des différences entres les tailles moyennes des arbres provenant des trois différentes hêtraies ?

. . . . . .

Exercice I.5. Rendements fouragers

On s'intéresse à l'ensemble des prairies d'une région donnée et on souhaite identifier l'importance, absolue ou relative, de la variabilité de la production fourragère, d'une part, d'une prairie à l'autre, et d'autre part, d'un endroit à l'autre, à l'intérieur des différentes prairies. Dans ce but, on a tout d'abord choisi au hasard trois prairies, dans l'ensemble du territoire, puis au sein de chacune de ces trois prairies, cinq petites parcelles, de deux mètres carrés. Dans l'optique d'un échantillonnage à deux degrés, les trois prairies constituent trois unités du premier degré, et les quinze petites parcelles quinze unités du deuxième degré.

Dans chacune des petites parcelles, on a mesuré les rendements en matière sèche à une date donnée. Les valeurs observées, exprimées en tonne par hectare, figurent dans le tableau ci-dessous.

	Prairie 1	Prairie 2	Prairie 3
Parcelle 1	2,06	1,59	1,92
Parcelle 2	2,99	2,63	1,85
Parcelle 3	1,98	1,98	2,14
Parcelle 4	2,95	2,25	1,33
Parcelle 5	2,70	2,09	1,83

Les rendements sont-ils homogènes ?

. . . . . .

Exercice I.6. Impact de promotions

Un dirigeant de magasin à succursales multiples envisage de faire trois types de promotions $P1$, $P2$ et $P3$ qui ont un coût sensiblement égal. Afin de déterminer celle qui sera finalement retenue, il fait tester les trois possibilités de promotion par un total de 10 magasins : 3 pour $P1$, 3 pour $P2$ et 4 pour $P3$. Le relevé de δ, le taux d'accroissement du chiffre d'affaires, exprimé en %, de chacun de ces magasins a été reporté dans le tableau ci-dessous.

Promotion 1	2,1	3,5	4,0	3,1	2,3	
Promotion 2	1,8	3,6	4,3	2,7	5,1	
Promotion 3	2,2	2,5	3,1	3,8	6,0	3,5

En utilisant la statistique de Jonckheere-Terpstra, déterminer si les promotions ont la même influence sur δ le taux d'accroissement du chiffre d'affaires.

. . . . . .

Exercice I.7. Comparaison de résultats

On dispose de trente échantillons dont on souhaite déterminer la teneur en un composé chimique donné. Chacun d'entre eux est analysé avec trois méthodes différentes d'analyse chimique. Les résultats obtenus ont été reproduits dans le tableau ci-dessous.

	Méthode				Méthode		
Échantillon	1	2	3	Échantillon	1	2	3
1	133	129	138	16	153	150	152
2	131	132	138	17	125	123	122
3	119	121	121	18	124	120	124
4	124	124	121	19	127	125	124
5	123	124	124	20	136	132	130
6	122	122	123	21	131	130	133
7	127	131	135	22	136	136	133
8	116	116	115	23	123	120	123
9	116	118	122	24	123	117	116
10	104	101	101	25	122	118	121
11	119	117	115	26	101	104	107
12	126	120	121	27	96	97	98
13	96	93	93	28	108	106	108
14	100	97	99	29	124	122	119
15	103	99	102	30	137	136	134

Observe-t-on une différence entre les résultats des différentes méthodes d'analyse chimique ?

. . . . . .

Exercice I.8. Compotes de pommes

Lors d'une évaluation sensorielle, 31 personnes ont jugé 6 compotes de pommes sur la base de critères relatifs à l'odeur, l'aspect, la texture et la saveur. À la fin chacun attribue une note allant de 0 (je n'aime pas du tout) à 10 (j'aime beaucoup), avec une précision de un dixième. On considérera ces notes comme issues de réalisations de variables quantitatives continues. Le tableau ci-dessous reprend un extrait des $31 \times 6 = 186$ données sur lesquelles sont réalisées les analyses.

Les résultats ont été reportés dans les tableaux suivants. Peut-on mettre en évidence l'influence d'un des facteurs *Juge* ou *Produit* sur la note finale ?

Juge	Ordre	Produit	Note	Juge	Ordre	Produit	Note	Juge	Ordre	Produit	Note	Juge	Ordre	Produit	Note
1	1	andros	4	1	2	scoup	4	1	3	st mamet	1	1	4	delisse	5
1	5	poti	2	1	6	carrefour	7	2	1	scoup	6	2	2	delisse	3
2	3	andros	5	2	4	carrefour	3	2	5	st mamet	7	2	6	poti	0
3	1	st mamet	7	3	2	andros	9	3	3	poti	8	3	4	scoup	0
3	5	carrefour	0	3	6	delisse	0	4	1	poti	3	4	2	st mamet	7
4	3	carrefour	7	4	4	andros	2	4	5	delisse	2	4	6	scoup	0
5	1	delisse	9	5	2	carrefour	2	5	3	scoup	1	5	4	poti	1
5	5	andros	0	5	6	st mamet	3	6	1	carrefour	3	6	2	poti	6
6	3	delisse	4	6	4	st mamet	1	6	5	scoup	0	6	6	andros	1
8	1	poti	7	8	2	delisse	9	8	3	scoup	2	8	4	carrefour	9
8	5	st mamet	7	8	6	andros	8	10	1	delisse	6	10	2	carrefour	7
10	3	poti	2	10	4	andros	6	10	5	scoup	8	10	6	st mamet	3
11	1	andros	8	11	2	st mamet	0	11	3	carrefour	0	11	4	scoup	0
11	5	delisse	0	11	6	poti	3	13	1	andros	6	13	2	poti	0
13	3	st mamet	8	13	4	carrefour	2	13	5	scoup	0	13	6	delisse	6
15	1	st mamet	5	15	2	andros	3	15	3	scoup	1	15	4	poti	0
15	5	delisse	1	15	6	carrefour	6	16	1	carrefour	6	16	2	delisse	2
16	3	poti	1	16	4	scoup	3	16	5	andros	1	16	6	st mamet	0
17	1	scoup	5	17	2	st mamet	1	17	3	delisse	7	17	4	andros	8
17	5	carrefour	4	17	6	poti	0	18	1	delisse	5	18	2	scoup	0
18	3	carrefour	1	18	4	st mamet	5	18	5	poti	0	18	6	andros	3
20	1	carrefour	4	20	2	delisse	3	20	3	scoup	3	20	4	poti	2
20	5	st mamet	2	20	6	andros	4	22	1	scoup	4	22	2	carrefour	1
22	3	st mamet	3	22	4	delisse	1	22	5	andros	1	22	6	poti	0
23	1	poti	4	23	2	andros	7	23	3	delisse	7	23	4	st mamet	5
23	5	carrefour	3	23	6	scoup	4	24	1	st mamet	3	24	2	scoup	6
24	3	andros	0	24	4	carrefour	4	24	5	poti	0	24	6	delisse	5

Juge	Ordre	Produit	Note	Juge	Ordre	Produit	Note	Juge	Ordre	Produit	Note	Juge	Ordre	Produit	Note
26	1	delisse	5	26	2	andros	2	26	3	carrefour	7	26	4	poti	1
26	5	scoup	0	26	6	st mamet	4	27	1	poti	7	27	2	st mamet	5
27	3	andros	6	27	4	scoup	3	27	5	delisse	7	27	6	carrefour	7
28	1	scoup	3	28	2	carrefour	3	28	3	st mamet	2	28	4	delisse	3
28	5	poti	0	28	6	andros	3	29	1	st mamet	8	29	2	scoup	6
29	3	poti	2	29	4	carrefour	6	29	5	andros	4	29	6	delisse	7
30	1	carrefour	7	30	2	delisse	5	30	3	scoup	3	30	4	andros	5
30	5	st mamet	4	30	6	poti	2	31	1	andros	4	31	2	delisse	9
31	3	poti	1	31	4	carrefour	1	31	5	scoup	0	31	6	st mamet	8
36	1	st mamet	3	36	2	scoup	1	36	3	carrefour	2	36	4	poti	0
36	5	delisse	1	36	6	andros	1	39	1	scoup	7	39	2	poti	2
39	3	st mamet	3	39	4	carrefour	5	39	5	andros	3	39	6	delisse	6
41	1	delisse	8	41	2	andros	4	41	3	carrefour	5	41	4	st mamet	4
41	5	poti	10	41	6	scoup	2	42	1	carrefour	5	42	2	delisse	5
42	3	poti	7	42	4	andros	5	42	5	scoup	3	42	6	st mamet	3
44	1	andros	6	44	2	scoup	6	44	3	delisse	3	44	4	st mamet	7
44	5	poti	2	44	6	carrefour	2	46	1	carrefour	4	46	2	poti	6
46	3	st mamet	0	46	4	delisse	4	46	5	scoup	0	46	6	andros	4
47	1	poti	8	47	2	delisse	3	47	3	carrefour	1	47	4	andros	1
47	5	st mamet	4	47	6	scoup	0								

Chapitre 7

Feuille de Travaux Pratiques n° 2

Valeurs non représentatives.

Exercice II.1. Vitesse du vent

Les 31 données suivantes représentent la vitesse quotidienne moyenne, en *mph*, du vent au large de l'aéroport MacArthur de Long Island au mois de juillet 1985. (NOAA, 1985).

7,7	11,1	7,8	9,5	5,9
8,5	8,8	11,5	5,6	10,7
6,9	8,9	10,2	6,2	7,7
11,1	9,0	8,7	10,4	5,2
17,1	11,2	10,7	12,5	3,8
13,3	6,2	8,8	8,1	7,4
8,9				

1. Représenter graphiquement l'échantillon afin de détecter de potentielles valeurs non représentatives. Une hypothèse de normalité est-elle vraisemblable ?
2. Réaliser un test de Grubbs basé sur la statistique T.
3. Utiliser la statistique de Tietjen Moore. On justifiera le choix de $k = 2$.
4. Utiliser la procédure RST de Rosner. On justifiera le choix de $k = 2$.
5. Les résultats obtenus aux questions 2. et 4. sont-ils différents ? Comment expliquez-vous cette situation ?

Exercice II.2. Hauteurs de plantes

L'échantillon suivant, dont l'effectif est 15, représente les différences de hauteur, en huitième de pouce, entre des plants de « Zea May » qui ont été soit fertilisés par eux-mêmes, soit entre eux. (Fisher, 1971)

50	-67	8
16	6	23
28	41	14
29	56	24
75	60	-48

1. Réprésenter graphiquement l'échantillon afin de détecter de potentielles valeurs non représentatives. Une hypothèse de normalité est-elle vraisemblable ?
2. Utiliser la statistique $r'_{1,0}$ de Dixon pour tester la non représentativité de $x_{(1)}$.
3. Utiliser la statistique $r'_{2,0}$ de Dixon pour tester la non représentativité de $x_{(1)}$.
4. D'après Dixon, quelle est la statistique $r'_{j,k}$ de Dixon qu'il faudrait utiliser pour tester la non représentativité de $x_{(1)}$? Faire ce test.
5. Les résultats obtenus aux questions 2., 3. et 4. sont-ils différents ? Comment expliquez-vous cette situation ?

Exercice II.3. Leucémie

On a reporté dans le tableau ci-dessous, la période de latence de leucémie aigüe, en mois, suite à une chimiothérapie pour 20 patients. (Kapadia, Krause, Ellis, Pan & Wald, American Cancer Society, 1980)

16	72	54	52
62	12	21	44
56	32	60	60
168	66	50	11
132	48	120	72

1. Représenter graphiquement l'échantillon afin de détecter de potentielles valeurs non représentatives. Une hypothèse de normalité est-elle vraisemblable ?
2. Faire un test de Grubbs pour la valeur non représentative $x_{(n)}$.
3. D'après Dixon, quelle est la statistique $r_{j,k}$ de Dixon qu'il faudrait utiliser pour tester la non représentativité de $x_{(n)}$? Faire ce test.
4. Utiliser un test de Grubbs pour $k = 3$ valeurs non représentatives dans une direction donnée. Justifier le choix de ce test et la valeur de k.
5. Utiliser une procedure séquentielle de Prescott avec $k = 3$.
6. Les résultats obtenus aux questions 2., 3., 4. et 5. sont-ils différents ? Comment expliquez-vous cette situation ?

Chapitre 8

Feuille de Travaux Pratiques n° 3

Corrélations non paramétriques

Exercice III.1. D'après Frontier, Davoult, Gentilhomme, Lagadeuc *Statistique pour les sciences de la vie et de l'environnement*, Dunod, 2001.

21 poissons d'une même espèce ont été pesés (en g), et leur concentration en polychlorobiphényles, notée PCB, mesurée (en $\mu g/g$ de tissu). Voici les résultats :

Masse	PCB	Masse	PCB	Masse	PCB	Masse	PCB	Masse	PCB
144	0,57	114	1,32	78	0,37	78	0,51	455	1,55
123	0,61	161	0,13	82	0,81	130	0,75	214	0,82
93	0,33	92	0,83	310	1,91	733	1,48	159	0,77
157	0,61	86	0,65	319	0,66	1030	1,11	212	2,31
95	5,58								

1. À l'aide de Minitab, construire le diagramme de dispersion des valeurs. Observer sur le diagramme que les points se groupent en deux ensembles :
   - un nuage de points correspondant à des individus de faible poids et non contaminés
   - un nuage de six points évoquant une relation décroissante, d'allure hyperbolique, entre les deux variables.
2. À l'aide de Minitab, construire le diagramme de dispersion des rangs. Repérer ces six points sur ce deuxième diagramme.
3. Y a-t-il lieu de calculer et tester, dans ces conditions, un coefficient de corrélation linéaire simple[1] sur l'ensemble des valeurs de poids et de PCB ? Pourquoi ?

[1] Ce coefficient est aussi appelé le coefficient de corrélation de Bravais-Pearson

4. Calculer le coefficient de rangs de Spearman relatif à l'ensemble des données puis relatif aux six individus distingués auparavant et ce de deux manières différentes :
   - Pour la première méthode, vous utiliserez la formule du cours et la calculatrice de Minitab.
   - Pour la seconde, vous utiliserez la statistique du test de Spearman calculée par Minitab. Attention, Minitab ne donne pas la p-value. Vous devez donc vous servir des tables associées à ce test.
5. Calculer le coefficient de rangs de Kendall relatif à l'ensemble des données puis relatif aux six individus distingués auparavant et ce de deux manières différentes :
   - Pour la première méthode, vous utiliserez la formule du cours et la calculatrice de Minitab.
   - Pour la seconde, vous utiliserez la statistique du test de Kendall calculée par Minitab. Attention, Minitab ne donne pas la p-value. Vous devez donc vous servir des tables associées à ce test.

Test de Mann-Whitney pour des échantillons indépendants

Exercice III.2. D'après Frontier, Davoult, Gentilhomme, Lagadeuc *Statistique pour les sciences de la vie et de l'environnement*, Dunod, 2001.

Des larves issues de deux populations d'une espèce d'insectes éloignés géographiquement sont récoltées à un même stade de développement. Leur taille (en μm) sont les suivantes :

Population 1	264	285	275	254	296
Population 2	290	317	307	296	291

1. Peut-on déterminer si les deux populations larvaires sont caractérisées par des tailles différentes ? Pour cela, peut-on faire un test paramétrique que vous avez généralement rencontré en deuxième année de Licence ? Si oui, alors quel est le nom de ce test ? Quelles sont les conditions d'application de ce test ? Que concluez-vous avec le test paramétrique ? Doit-on calculer la puissance du test ? Si oui, calculer-là.
2. Malgré la très faible taille des échantillons, peut-on déterminer, par le test de Mann-Whitney, si les deux populations larvaires sont caractérisées par des tailles différentes ?
3. La conclusion obtenue avec le test de Mann-Whitney diffère-t-elle de celle obtenue à la question 1. ? Discuter.

Test de Wilcoxon pour des échantillons appariés

Exercice III.3. D'après Mercier, Morin, Viel, Jolly, Daures, Chastang *Biostatistique et probabilités*, Ellipses, 1996.

20 singes ont été groupés en dix couples de telle sorte que les deux singes d'un même couple soient de poids similaire. Un singe a été choisi aléatoirement dans chaque couple pour recevoir le régime A constitué de carottes, alors que le second reçoit le régime B constitué de pommes. Les gains observés par jour sont les suivants

Couple	1	2	3	4	5	6	7	8	9	10
Régime A	21	21	19	16	26	19	18	29	27	19
Régime B	30	25	25	16	29	18	18	19	24	22

1. Peut-on dire que les deux régimes sont équivalents ? Pour cela, vous effectuerez un test à l'aide de Minitab que vous avez généralement rencontré en deuxième année de Licence. Quelle est la conclusion de ce test ?
2. On peut envisager de faire un autre type de test. De quel test s'agit-il ?
3. Réaliser ce test à l'aide de Minitab. Quelle conclusion obtenez-vous ?
4. La conclusion obtenue avec le dernier test diffère-t-elle de celle obtenue à la question 1. ? Discuter la similitude ou la différence des deux conclusions ?

Bibliographie

[1] A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1) :131–152, 1992.

[2] A. Boomsma. Comparing approximations of confidence intervals for the product-moment correlation coefficient. *Statistica Neerlandica*, 31 :179–186, 1977.

[3] J. Bouyer. *Méthodes Statistiques*. Editions INSERM, 1996.

[4] P. Chapouille. *Planification et analyse des expériences*. Masson, Paris, 1973.

[5] R. Christensen. *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer Texts in Statistics. Springer-Verlag, 1991.

[6] P. Dagnélie. *Statistique Théorique et Appliquée*, volume 2. De Boeck & Larcier, Bruxelles, 1998.

[7] P. Dagnélie. *Statistique Théorique et Appliquée*, volume 1. De Boeck & Larcier, Bruxelles, 1998.

[8] B. Falissard. *Comprendre et utiliser les statistiques dans les sciences de la vie*. Abrégés. Masson, Paris, 3ème edition, 2005.

[9] G. H. Freeman and J.H. Halton. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38 :141–149, 1952.

[10] S. Kotz, C. B. Read, and N. Balakrishnan, editors. *Encyclopædia Of Statistical Sciences*. Wiley-Interscience, 2nd edition, 1996.

[11] J.-Y. Ouvrard. *Probabilités*, volume 2. Cassini, Paris, 2000.

[12] G. Parreins. *Techniques Statistiques : moyens rationnels de choix et de décision*. Dunod technique, Paris, 1974.

[13] G. Pupion and P.-C. Pupion. *Tests non paramétriques*. Statistique mathématique et probabilité. Economica, Paris, 1998.

[14] H.C. Thode. *Testing for Normality*. Number 164 in Statistics : textbooks and monographs. Marcel Dekker, New-York, 2002.

[15] E. Weisstein. Correlation coefficient-bivariate normal distribution from mathworld–a wolfram web resource. Lien internet : `http ://mathworld.wolfram.com/-CorrelationCoefficientBivariateNormalDistribution.html`.

Table des matières

I Notes de cours **3**

1 Quelques tests non paramétriques **5**

1.1 Les tests non paramétriques sur un échantillon 5

1.1.1 Test du signe 5

1.1.2 Test des rangs signés de Wilcoxon 7

1.1.2.1 Cas où il n'y a pas d'ex æquo. 8

1.1.2.2 Cas où il y a des ex æquo. 8

1.2 Les tests non paramétriques sur deux échantillons 10

1.2.1 Les échantillons sont indépendants : Test de Mann-Whitney 10

1.2.1.1 Cas où il n'y a pas d'ex aequo. 10

1.2.1.2 Cas où il y a des ex æquo. 11

1.2.2 Les échantillons sont indépendants : Test de la médiane de Mood . 12

1.2.3 Les échantillons sont dépendants : Test de Wilcoxon 13

1.2.3.1 Cas où il n'y a pas d'ex aequo. 13

1.2.3.2 Cas où il y a des ex æquo. 14

1.3 Les tests non paramétriques sur k échantillons : 1 facteur 14

1.3.1 Les échantillons sont indépendants : Test de Kruskal-Wallis 14

1.3.1.1 Cas où il n'y a pas d'ex æquo. 15

1.3.1.2 Cas où il y a des ex æquo. 15

1.3.2 Les échantillons sont indépendants : Test de Jonckheere-Terpstra . . 17

1.3.2.1 Cas où il n'y a pas d'ex æquo. 17

1.3.2.2 Cas où il y a des ex æquo. 18

1.3.3 Les échantillons ne sont pas indépendants : Test de Friedman 18

1.3.3.1 Cas où il n'y a pas d'ex æquo. 19

1.3.3.2 Cas où il y a des ex æquo. 19

1.4 Les tests non paramétriques sur nk échantillons : 2 facteurs 20

1.4.1 Les échantillons sont indépendants : Test de Friedman 20

1.4.1.1 Cas où il n'y a pas d'ex æquo. 21

1.4.1.2 Cas où il y a des ex æquo. 22

2 Valeurs non représentatives **25**

2.1 Valeurs extrêmes, valeurs non représentatives 25

2.2 Que vérifier ? 26

2.3 Que faire avec une valeur potentiellement non-représentative ? 27
2.3.1 Notations 27
2.3.2 Test de Grubbs pour une seule valeur non représentative 28
2.3.3 Test de Dixon pour une seule valeur non représentative 29
2.3.4 Test basé sur l'étendue 31
2.4 Test simultané de k valeurs non représentatives 32
2.4.1 Test de Grubbs pour k valeurs non représentatives dans une direction donnée. 32
2.4.2 Test de Grubbs pour deux valeurs non représentatives, une de chaque côté 33
2.4.3 Test de Tietjen-Moore pour k valeurs non représentatives de l'un ou des deux côtés 34
2.5 Procédures pour détecter un nombre de valeurs non représentatives non fixé à l'avance 36
2.5.1 La boîte à moustaches 36
2.5.2 Test basé sur le coefficient d'asymétrie 37
2.5.3 Test basé sur le coefficient d'aplatissement 38
2.6 Procédures séquentielles de détections de valeurs non représentatives . . . 38
2.6.1 Procédure séquentielle de Prescott 38
2.6.2 Procédure RST de Rosner 39
2.7 Conclusion 40

3 Mesures de liaison paramétriques. 43
3.1 Coefficient de corrélation simple 43
3.2 Le cas bidimensionnel 45
3.2.1 Loi normale bidimensionnelle 45
3.2.2 Estimation des paramètres 46
3.2.3 Procédure de test 51
3.2.4 Remarques sur les liaisons 53
3.3 Le cas général ($n \geqslant 2$) 54
3.3.1 Loi multinormale 54
3.3.2 Estimation 54
3.3.3 Test de l'hypothèse $\rho_{ij} = 0$ 57
3.3.4 Test de l'hypothèse $\rho_{ij} = \rho_0$, $\rho_0 \neq 0$ 57
3.4 Corrélation multiple 60
3.4.1 Définition 60
3.4.2 Estimation 61
3.4.3 Asymptotique 62
3.4.4 Test de l'hypothèse $R(X_1, \boldsymbol{X_2}) = 0$ 62
3.4.5 Test de l'hypothèse $R(X_1, \boldsymbol{X_2}) = R_0$, $R_0 \neq 0$ 63
3.5 Corrélation partielle 64
3.5.1 Définition 64
3.5.2 Estimation 65

3.5.3 Asymptotique 65
3.5.4 Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$ 66
3.5.5 Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$ 66
3.5.6 Cas de trois variables réelles 67

4 Mesures de liaison non paramétriques 83
4.1 Mesures non paramétriques 83
4.2 La statistique $\rho_{\mathrm{S},n}$ de Spearman 83
4.2.1 Cadre d'application 83
4.2.2 Le coefficient de corrélation $\rho_{\mathrm{S}}(X,Y)$ de Spearman 84
4.2.3 Estimation de $\rho_{\mathrm{S}}(X,Y)$ 85
4.2.4 Procédure de test 86
4.2.5 Cas des ex æquo 88
4.2.6 Statistique corrigée $\rho^{\star}_{\mathrm{S},n}$ 88
4.2.7 Départition des ex æquo 89
4.3 Corrélation partielle de Spearman 90
4.3.1 Coefficient de corrélation partielle de Spearman $\rho_{\mathrm{S}}(X,Y|Z)$ 90
4.3.2 Estimation de $\rho_{\mathrm{S}}(X,Y|Z)$ 90
4.3.3 Méthode d'utilisation 90
4.4 La statistique τ_n de Kendall 91
4.4.1 Cadre d'application 91
4.4.2 Le coefficient de corrélation $\tau(X,Y)$ de Kendall 91
4.4.3 Estimation de $\tau(X,Y)$ 92
4.4.4 Procédure de test 94
4.4.5 Cas des ex æquo 97
4.4.6 Statistique corrigée $\tau^{\star}_n$ 97
4.4.7 Départition des ex æquo 98
4.5 Corrélation partielle de Kendall 98
4.5.1 Coefficient de corrélation partiel de Kendall $\tau(X,Y|Z)$ 98
4.5.2 Notations 99
4.5.3 Estimateur de $\tau(X,Y|Z)$ 101
4.5.4 Méthode d'utilisation 101
4.6 Test du χ^2 d'indépendance 102
4.6.1 Les classes 102
4.6.2 Statistique du test 103
4.6.3 Méthode d'utilisation 103
4.6.4 Correction de Yates 104
4.6.5 Étude des résidus 105
4.7 Test exact de Fisher 105
4.7.1 Deux variables à deux modalités 105
4.7.2 Deux variables ayant un nombre fini quelconque de modalités . . . 110

5 Tests de multinormalité **113**
5.1 Tests de multinormalité 113
5.1.1 Utilisation de tests multiples unidimensionnels 113
5.1.2 Une test basée sur la modification du test de Shapiro-Wilk unidimensionnel 115
5.1.3 Asymétrie et aplatissement multivariés de Mardia 116
5.1.4 Centrer-réduire. 116
5.1.5 Asymétrie et aplatissement 117
5.1.6 Étude sous l'hypothèse de multinormalité de $\boldsymbol{X}$. 118
5.1.7 Un test basé sur la fonction caractéristique empirique 119

II Feuilles de travaux pratiques **121**

6 Feuille de Travaux Pratiques n° 1 **123**

7 Feuille de Travaux Pratiques n° 2 **131**

8 Feuille de Travaux Pratiques n° 3 **133**