Cours de Statistique pour Licence troisième année de Biologie

Le chapitre un a été rédigé par Frédéric Bertrand et Myriam Maumy.
La version originale des chapitres deux et trois a été rédigée par Photis Nobelis
et modifiée par Frédéric Bertrand, Jean-Luc Dortet, Claire Dupuis et Myriam Maumy.

Table des matières

1 Procédures non paramétriques sur un ou deux échantillons **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
- 1.2 Les tests non paramétriques sur deux échantillons . . . 9
  - 1.2.1 Les échantillons sont indépendants - Test de Mann-Whitney . . . 9
  - 1.2.2 Les échantillons sont dépendants - Test de Wilcoxon . . . 11

2 Analyse de la variance à un facteur - Test de comparaison de plusieurs moyennes théoriques **13**
- 2.1 Modélisation statistique . . . 13
- 2.2 Tableau de l'Analyse de la Variance - Test (cas équilibré) . . . 15
- 2.3 Vérification des trois conditions . . . 18
  - 2.3.1 Indépendance. . . . 18
  - 2.3.2 Normalité. . . . 18
  - 2.3.3 Homogénéité. . . . 20
- 2.4 Comparaisons multiples . . . 21
  - 2.4.1 Le test de Tukey . . . 21
  - 2.4.2 Le test de Dunnett . . . 22
- 2.5 Risque de deuxième espèce et risque a posteriori . . . 23
- 2.6 Transformations . . . 23
- 2.7 Facteurs aléatoires . . . 24
- 2.8 Analyse de la Variance non paramétrique - Test de Kruskal-Wallis . . . 24
  - 2.8.1 Cas où il n'y a pas d'ex-æquo . . . 25
  - 2.8.2 Cas où il y a des ex æquo . . . 27
- 2.9 Quelques précisions sur les comparaisons multiples . . . 27

3 Analyse de régression linéaire : Corrélation linéaire - Régression linéaire simple **31**
- 3.1 Introduction . . . 31
- 3.2 Le coefficient de corrélation linéaire . . . 32
- 3.3 Tests d'hypothèse . . . 35
  - 3.3.1 Test de l'hypothèse nulle $(\mathcal{H}_0) : \varrho(X, Y) = 0$ . . . 35
  - 3.3.2 Test de l'hypothèse nulle $(\mathcal{H}_0) : \varrho(X, Y) = \varrho_0(X, Y)$ . . . 36
- 3.4 Intervalle de confiance de $\varrho(X, Y)$ . . . 37
- 3.5 Le coefficient de détermination et le rapport de corrélation . . . 37
- 3.6 La régression linéaire simple . . . 39
- 3.7 La méthode des moindres carrés ordinaire . . . 40
- 3.8 La validation du modèle . . . 41
- 3.9 Vérification des conditions . . . 45
  - 3.9.1 La normalité . . . 45
  - 3.9.2 Étude graphique des résidus. . . . 45
  - 3.9.3 L'homogénéité. . . . 46
- 3.10 Étude des paramètres a et b . . . 46
  - 3.10.1 Intervalles de confiance . . . 46
  - 3.10.2 Tests d'hypothèses . . . 47

Chapitre 1

Procédures non paramétriques sur un ou deux échantillons

1.1. Les tests non paramétriques sur un échantillon

Dans cette section nous nous intéressons à deux **tests non paramétriques** (**nonparametric tests**) :
– le test du signe
– le test des rangs signés de Wilcoxon.
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** (**sample**) indépendant et identiquement distribué constitué de n variables aléatoires notées $X_1, \ldots, X_n$ d'une loi continue F dont la **médiane** (**median**) est notée m_e et la **moyenne** (**mean**) μ. Le **test du signe** (**sign test**) permet de tester l'hypothèse (**hypothesis**) suivante :

Hypothèses :

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

contre

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

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 :

$$(\mathcal{H}_1^{'}) : \mathbb{P}[X_i > 0] < \frac{1}{2}$$

soit

$$(\mathcal{H}_1^{''}) : \mathbb{P}[X_i > 0] > \frac{1}{2}.$$

Remarque 1.1.2. Plus généralement ce test permet de tester l'**hypothèse nulle** (**null hypothesis**)

$$(\mathcal{H}_0) : m_e = m_0$$

contre l'**hypothèse alternative** (**alternative hypothesis**)

$$(\mathcal{H}_1) : m_e \neq m_0$$

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

$$(\mathcal{H}_1^{'}) : m_e < m_0$$

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

$$(\mathcal{H}_1^{''}) : m_e > m_0$$

lorsque m_0 est un nombre réel. Pour cela il suffit de considérer l'échantillon $Z_1, \ldots, Z_n$ avec $Z_i = X_i - m_0$ et nous sommes ramenés au test précédent.

Statistique : S_n désigne le nombre de variables aléatoires 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 a les trois propriétés suivantes :*

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

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

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

où Φ est la fonction de répartition de la loi normale centrée réduite et k est un nombre entier compris entre 0 et n.

Décision 1.1.1. *Pour un seuil donné α ($= 5\% = 0,05$ en général), nous cherchons le plus grand nombre entier k_α tel que $\mathbb{P}_{(\mathcal{H}_0)}[S_n \leqslant k_\alpha] \leqslant \alpha/2$. Alors nous décidons :*

$$\begin{cases} \text{si } S_{n,obs} \notin]k_\alpha, n-k_\alpha[& (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } S_{n,obs} \in]k_\alpha, n-k_\alpha[& (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Remarque 1.1.3. Le niveau de signification réel du test est alors égal à $2\mathbb{P}[S_n \leqslant k_\alpha]$ qui est généralement différent de α.

Exemple 1.1.1. Il est admis que des prématurés nés avec un poids de 2,15 kg arrivent à un poids de 2,80 kg en un mois s'ils sont nourris avec du lait maternel. Douze nourrissons pesant approximativement 2,15 kg à la naissance ont été nourris avec un lait sensé remplacer le lait maternel. Les gains de poids en kilogrammes, notés x_i, ont été de

$$0,55, \quad 0,62, \quad 0,54, \quad 0,58, \quad 0,66, \quad 0,64, \quad 0,60, \quad 0,62, \quad 0,59, \quad 0,67, \quad 0,62, \quad 0,61.$$

Est-il possible de conclure, au seuil de $\alpha = 5\%$, à une différence significative entre les effets du lait maternel et ceux du lait de remplacement relativement à la prise de poids des nourrissons ?

Compte tenu du faible effectif de l'échantillon, nous décidons de tester l'hypothèse nulle :

$$(\mathcal{H}_0) : m_e = m_0 = 2,80\ kg$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : m_e \neq m_0 = 2,80\ kg$$

Nous allons d'abord transformer les données observées $x_1, \ldots, x_{12}$ en un échantillon $z_1, \ldots, z_{12}$. Pour cela, nous calculons les z_i avec la formule suivante : $z_i = 2,15 + x_i - 2,80$. Nous avons donc la suite de données suivantes :

$$-0,10, \quad -0,03, \quad -0,11, \quad -0,07, \quad +0,01, \quad -0,01, \quad -0,05, \quad -0,03, \quad -0,06, \quad +0,02, \quad -0,03, \quad -0,04.$$

Nous en déduisons que $S_{12,obs} = 2$. Il ne reste plus qu'à déterminer le plus grand nombre entier k_α tel que

$$\mathbb{P}_{(\mathcal{H}_0)}[S_{12} \leqslant k_\alpha] \leqslant 0,025.$$

Pour trouver ce nombre entier, nous allons utiliser le logiciel **MINITAB** et en particulier, dans le menu **Calc**, **les lois de probabilité**. Ensuite, nous sélectionnons le sous menu **Binomiale** et nous remplissons la fenêtre en cochant **Probabilité cumulée**, en remplissant **Nombre d'essais** par 12 et **Probabilité de succès** par $0,5$. Il ne reste plus qu'à remplir la case de la **Colonne d'entrée** qui n'est rien d'autre qu'une colonne que nous avons prélablement remplie par les entiers de 0 à 12. Nous trouvons que ce nombre entier k_α est égal à 2. Par conséquent, comme $S_{12} \notin]2, 10[$, le test est significatif au seuil de $\alpha = 5\%$. Nous rejetons donc l'hypothèse nulle $(\mathcal{H}_0)$ et nous décidons que l'hypothèse alternative $(\mathcal{H}_1)$ est vraie. Il est donc possible de conclure, avec un risque d'erreur de première espèce de $\alpha = 5\%$, à une différence significative entre les effets du lait maternel et ceux du lait de remplacement relativement à la prise de poids des nourrissons.

Une autre façon de procéder aurait été de faire faire au logiciel **MINITAB** le **test du signe pour la médiane** sur les données z_i. À l'issue de cette procédure, le logiciel **MINITAB** nous donne

```
Test des signes de la médiane = 0,00000 par rapport à  non =  0,00000

               N  Au-dessous   Egal  Au-dessus         P    Médiane
z_i           12          10      0          2    0,0386   -0,03500
```

Comme la plupart des logiciels, le logiciel **MINITAB** nous fournit la $P-$**valeur** ($P-$**value**) du test.

Définition 1.1.1. *La $P-$valeur est la probabilité, calculée sous l'hypothèse nulle ($\mathcal{H}_0$), que la statistique du test prenne la valeur obtenue ou toute valeur plus extrême du côté de l'hypothèse alternative ($\mathcal{H}_1$). En d'autres termes, la $P-$valeur est la probabilité de « voir ce que j'ai vu ou pire du côté de l'hypothèse alternative ($\mathcal{H}_1$) ».*

La règle de décision pour un seuil α donné est alors :

$$\begin{cases} \text{si } P-\text{valeur} \leqslant \alpha & (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } P-\text{valeur} > \alpha & (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Il ne reste plus qu'à interpréter cette $P-$valeur pour conclure sur la décision du test que nous allons prendre. Comme la $P-$valeur est égale à 0,0386 qui est inférieure à 0,05, c'est-à-dire à α, nous rejetons l'hypothèse nulle ($\mathcal{H}_0$) et nous décidons que l'hypothèse alternative ($\mathcal{H}_1$) est vraie.

Le niveau de signification réel du test est alors égal à 0,03858.

1.1.2. Test des rangs signés de Wilcoxon

Soit un échantillon indépendant et identiquement distribué constitué de n variables aléatoires notées $X_1, \ldots, X_n$ d'une loi continue F dont la valeur médiane est notée m_e et la moyenne μ. Le test des rangs signés de Wilcoxon (**Wilcoxon signed rank test**) permet de tester l'hypothèse suivante :

Hypothèses :

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

contre

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

Remarque 1.1.4. Si nous savons que la loi continue F est symétrique, alors le test des rangs signés de Wilcoxon permet de tester l'hypothèse suivante :

$$(\mathcal{H}_0) : \mu = 0$$

contre

$$(\mathcal{H}_1) : \mu \neq 0,$$

ce qui permet de s'intéresser à la moyenne μ de la loi continue F.

Remarque 1.1.5. La formulation de ce test est la formulation d'un test bilatéral. Nous pourrions envisager d'étudier les deux tests unilatéraux correspondants.

Cas où il n'y a pas d'ex æquo.

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

Remarque 1.1.6. La lettre a est là pour rappeler que nous travaillons sur les valeurs absolues des x_i.

Statistique : Nous déterminons alors la somme $W_{n,obs}^+$ des rangs r_i^a des seules observations strictement positives. La statistique W_n^+ des rangs signés de Wilcoxon est la variable aléatoire qui prend pour valeur la somme $W_{n,obs}^+$. Par conséquent, la statistique W_n^+ des rangs signés de Wilcoxon se définit par :

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

Propriétés 1.1.2. *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. *$\text{Var}[W_n^+] = n(n+1)(2n+1)/24$.*

3. Elle est tabulée pour de faibles valeurs de n. Pour $n \geqslant 15$, nous utiliserons l'approximation normale avec correction de continuité :

$$\mathbb{P}\left[W_n^+ \leqslant w\right] = \Phi\left(\frac{2w+1-\dfrac{n(n+1)}{2}}{\sqrt{\dfrac{n(n+1)(2n+1)}{6}}}\right)$$

où Φ est la **fonction de répartition** *(***distribution function***) de la loi normale centrée réduite et w est un nombre entier compris entre 0 et $n(n+1)/2$.*

Décision 1.1.2. – **Premier cas :** *La taille n est inférieure à 15. Pour un seuil donné α ($= 5\% = 0,05$ en général), nous cherchons le plus grand nombre entier w_α tel que $\mathbb{P}_{(\mathcal{H}_0)}\left[W_n^+ \leqslant w_\alpha\right] \leqslant \alpha/2$. Alors nous décidons :*

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

– **Deuxième cas :** *La taille n est supérieure à 15. La statistique W_n^+ suit approximativement une loi normale et nous utilisons alors la statistique suivante :*

$$Z_n = \frac{2W_n^+ + 1 - \dfrac{n(n+1)}{2}}{\sqrt{\dfrac{n(n+1)(2n+1)}{6}}}.$$

Pour un seuil donné α ($= 5\% = 0,05$ en général), la table de la loi d'une variable aléatoire Z normale centrée réduite nous fournit une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}\left[-c \leqslant Z_n \leqslant c\right] \leqslant 1-\alpha$. Alors nous décidons :

$$\begin{cases} \text{si } Z_{n,obs} < -c \quad \text{ou} \quad \text{si } Z_{n,obs} > c & (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } -c \leqslant Z_{n,obs} \leqslant c & (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Remarque 1.1.7. Pour tester l'hypothèse nulle « $(\mathcal{H}_0) : \mu = \mu_0$ », nous introduisons l'échantillon $Z_1, \ldots, Z_n$ avec $Z_i = X_i - \mu_0$.

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. Deux procédures sont alors employées. (Par extension nous pourrons utiliser ces procédures lorsque la loi F est discrète.)

- *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 la variable 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 $|x_1|, \ldots, |x_n|$ sont 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 peuvent 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 $(\mathcal{H}_0)$: « La loi continue F est symétrique en 0 » et pour $n > 15$, nous admettrons que la variable aléatoire

$$\frac{W_n^{+\star} - m^\star}{\sigma^\star}$$

où

$$m^\star = \frac{1}{4}\left(n(n+1) - d_0(d_0+1)\right) \quad \text{et} \quad \left(\sigma^\star\right)^2 = \frac{1}{24}\left(n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)\right) - \frac{1}{48}\sum_{j=1}^{h}\left(d_j^3 - d_j\right).$$

suit approximativement une loi normale centrée réduite $\mathcal{N}(0\ ;\ 1)$.

Remarque 1.1.8. Lorsque 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^+.

Remarque 1.1.9. Le symbole $\star$ est là pour rappeler que nous sommes dans le cas où il y a des ex æquo.

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 X aléatoire, continue, sur deux populations, ou sur une population divisée en deux sous-populations. Nous obtenons ainsi deux séries d'observations notées $(x_1, \ldots, x_{n_1})$ pour la première et $(y_1, \ldots, y_{n_2})$ pour la seconde. Nous notons $\mathcal{L}_i$ la loi de la variable aléatoire X sur la (sous-)population d'ordre i. Le **test de Mann-Whitney** (**Mann-Whitney test**) permet de tester l'hypothèse suivante :

Hypothèses :

$$(\mathcal{H}_0) : \text{Les deux lois } \mathcal{L}_i \text{ sont identiques ou encore de façon équivalente} \quad \mathcal{L}_1 = \mathcal{L}_2$$

contre

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

Cas où il n'y a pas d'ex aequo.

Statistique : Pour obtenir la statistique du test notée U_{n_1,n_2} en général, nous devons procéder à des calculs successifs :

1. 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 les sommes des rangs (**rank sums**) pour chacun des deux échantillons, notées R_{n_1} et R_{n_2}.
4. Nous en déduisons les quantités U_{n_1} et U_{n_2} qui se calculent ainsi :

$$U_{n_1} = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_{n_1} \quad \text{et} \quad U_{n_2} = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_{n_2} = n_1 n_2 - U_{n_1}.$$

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. $\mathbb{E}\left[U_{n_1,n_2}\right] = (n_1 n_2)/2$.
2. $\mathrm{Var}\left[U_{n_1,n_2}\right] = (n_1 n_2)(n_1 + n_2 + 1)/12$.
3. *La variable aléatoire U_{n_1,n_2} est tabulée pour de faibles valeurs de n_1 et de n_2. Pour $n_1 \geqslant 20$ et $n_2 \geqslant 20$, nous utilisons l'approximation normale :*

$$\mathbb{P}_{(\mathcal{H}_0)}\left[U_{n_1,n_2} \leqslant u\right] = \Phi\left(\frac{u - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite et u est un nombre entier.

Décision 1.2.1. – **Premier cas :** *Les tailles n_1 ou n_2 sont inférieures à 20. Pour un seuil donné α ($= 5\% = 0,05$ en général), les tables de Mann-Whitney nous fournissent une valeur critique c. Alors nous décidons :*

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

– **Second cas :** *Les tailles n_1 et n_2 sont supérieures ou égales à 20. La statistique U_{n_1,n_2} suit approximativement une loi normale et nous utilisons alors la statistique suivante :*

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

Pour un seuil donné α ($= 5\% = 0,05$ en général), la table de la loi d'une variable aléatoire Z normale centrée réduite nous fournit une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[-c \leqslant Z_{n_1,n_2} \leqslant c] = 1-\alpha$. Alors nous décidons :

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

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. Deux procédures sont alors employées. (Par extension nous pourrons utiliser ces procédures lorsque les lois $\mathcal{L}_1$ et $\mathcal{L}_2$ sont discrètes.)

- *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 $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ sont ordonnées puis regroupées en h classes d'ex æquo C_j, $1 \leqslant j \leqslant h$. Certaines classes C_j peuvent 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 la variable X_i son rang moyen $R_i^\star$ dans ce classement et en sommant les rangs de tous les X_i, nous obtenons la statistique :

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

  Sous l'hypothèse nulle $(\mathcal{H}_0)$: « Les deux lois $\mathcal{L}_i$ sont identiques » et pour $n_1 > 15$ et $n_2 > 15$, nous admettrons que la variable aléatoire

  $$\frac{U^\star_{n_1,n_2} - m^\star}{\sigma^\star}$$

  où

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

  suit approximativement une loi normale centrée réduite $\mathcal{N}(0\ ;\ 1)$.

Remarque 1.2.1. Lorsque 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 dépendants - Test de Wilcoxon

Nous considérons deux variables aléatoires continues X et Y 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 $(x_1, y_1), \ldots, (x_n, y_n)$. Le **test de Wilcoxon** (**Wilcoxon test**) permet de tester l'hypothèse suivante :

Hypothèses :

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

contre

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

Remarque 1.2.2. Ce test suppose que la loi de la différence entre les deux variables étudiées X et Y est symétrique par rapport à 0.

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 calculs successifs :

1. 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.
2. Nous attribuons à chaque $|d_i|$ le rang correspondant.
3. Nous restituons ensuite à chaque rang le signe de la différence correspondante.
4. 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 :** *La taille n est inférieure à 15. Pour un seuil donné α ($= 5\% = 0,05$ en général), nous cherchons le plus grand nombre entier k_α tel que $\mathbb{P}_{(\mathcal{H}_0)}[S^+ \leqslant k_\alpha] \leqslant \alpha/2$. Alors nous décidons :*

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

– **Second cas :** *La taille n est supérieure à 15. Nous utilisons l'approximation normale avec correction de continuité :*

$$\mathbb{P}_{(\mathcal{H}_0)}\left[S^+ \leqslant k\right] = \Phi\left(\frac{2k + 1 - \dfrac{n(n+1)}{2}}{\sqrt{\dfrac{n(n+1)(2n+1)}{6}}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite et k un nombre entier compris entre 0 et n.

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.

Chapitre 2

Analyse de la variance à un facteur - Test de comparaison de plusieurs moyennes théoriques

2.1. Modélisation statistique

Nous étudions un test statistique permettant de comparer globalement les moyennes de plusieurs variables gaussiennes de même variance et de même nature. C'est l'une des procédures les plus utilisées dans les applications de la Statistique.

Exemple 2.1.1. Le service Recherche et Développement d'un laboratoire pharmaceutique a réalisé une étude sur la stabilité dans le temps de l'hydrophilie d'éponges artificielles. Douze éponges ont été choisies pour être conservées dans les mêmes conditions. Quatre durées ont été considérées :
- 3 mois,
- 6 mois,
- 12 mois,
- 24 mois.

Trois éponges ont été « affectées au hasard » à chaque durée. Les résultats, en unités d'hydrophilie, sont donnés dans le tableau suivant :

3 mois	6 mois	12 mois	24 mois
43	36	28	32
40	40	24	29
41	39	33	32

Cette écriture du tableau est dite « désempilée ». Nous pouvons l'écrire sous forme standard (« empilée »), c'est-à-dire avec deux colonnes, une pour la durée et une pour l'hydrophilie, et douze lignes, une pour chaque unité observée.

Éponge	Durée	Hydrophilie
1	3 mois	43
2	3 mois	40
3	3 mois	41
4	6 mois	36
5	6 mois	40
6	6 mois	39
7	12 mois	28
8	12 mois	24
9	12 mois	33
10	24 mois	32
11	24 mois	29
12	24 mois	32

Remarque 2.1.1. Dans la plupart des logiciels, et en particulier le logiciel **MINITAB**, c'est sous cette forme que sont saisies et traitées les données. Dans les deux tableaux, nous avons omis les unités de l'hydrophilie et ceci pour abréger l'écriture. Mais en principe cela doit être indiqué entre parenthèses à côté d'hydrophilie.

Remarque 2.1.2. Il va de soi que lorsque vous rentrerez des données sous le logiciel **MINITAB** vous n'indiquerez pas le mot « mois » à côté des nombres (3, 6, 12, 24). Il est juste là pour vous faciliter la compréhension du tableau mais il faudra plutôt le mettre en haut à coté du mot « Durée ».

Remarque 2.1.3. Nous avons en fait quatre échantillons chacun de taille trois ! Les populations de référence sont toutes abstraites : elles sont constituées de l'ensemble des éponges fabriquées par ce processus industriel et conservées durant l'une des périodes fixées pour l'expérience.

Sur **chaque unité**, nous observons **deux variables** :

1. la durée qui est totalement contrôlée. Elle est considérée comme qualitative avec quatre modalités bien déterminées. Nous l'appelons **le facteur (factor)**. Ici le facteur « Durée » est à **effets fixes (fixed effects)**.
2. l'hydrophilie qui est une mesure. Elle est parfois appelée **la réponse (response)**.

Notations 2.1.1. *La variable mesurée dans un tel schéma expérimental sera notée Y. Pour les observations nous utilisons deux indices :*
- *le premier indice indique le numéro de population (« Durée »),*
- *le second indice indique le numéro de l'observation dans l'échantillon.*

Pour le premier indice, nous utilisons i (ou encore i', i'', i_1, i_2). **Pour le second indice**, nous utilisons j (ou encore j', j'', j_1, j_2).

Ainsi les observations sont en général notées par :

$$y_{ij}, \quad i = 1, \ldots, I \quad j = 1, \ldots, J(i).$$

Lorsque *les échantillons sont de même taille* J, nous disons que l'expérience est **équilibrée (balanced)**. C'est le cas dans l'**Exemple 2.1.1.** avec

$$J = 3 \quad \text{et} \quad I = 4.$$

Si *les tailles des échantillons sont différentes*, alors elles sont notées par :

$$n_i, \quad i = 1, \ldots, I.$$

Mais ce plan expérimental est à éviter parce que les différences qu'il est alors possible de détecter sont supérieures à celles du schéma équilibré.

En se plaçant dans le **cas équilibré** nous notons les **moyennes (means)** de chaque échantillon par :

$$\overline{y}_i = \frac{1}{J} \sum_{j=1}^{J} y_{ij}, \quad i = 1, \ldots, I, \tag{2.1.1}$$

et les **variances (variances)** de chaque échantillon par :

$$s_i^2(y) = \frac{1}{J} \sum_{j=1}^{J} (y_{ij} - \overline{y}_i)^2, \quad i = 1, \ldots, I. \tag{2.1.2}$$

Remarque 2.1.4. Cette dernière formule exprime la variance non corrigée. Très souvent, dans les ouvrages ou les logiciels, c'est la variance corrigée qui est utilisée : au lieu d'être divisée par J, la somme est divisée par $J - 1$. **Cette remarque s'applique naturellement au logiciel MINITAB.**

Retour à l'Exemple 2.1.1. : Après calculs, nous avons :

$$\overline{y}_1 = 41,333 \quad \overline{y}_2 = 38,333 \quad \overline{y}_3 = 28,333 \quad \overline{y}_4 = 31,000$$

et

$$s_1^2(y) = 1,556 \quad s_2^2(y) = 2,889 \quad s_3^2(y) = 13,556 \quad s_4^2(y) = 2,000.$$

Le nombre total d'observations est égal à :

$$n = IJ = 12.$$

Conditions 2.1.1. *Nous supposons que les observations* $\{y_{ij}\}$ *sont des réalisations des variables* $\{Y_{ij}\}$ *qui satisfont aux trois conditions suivantes :*

1. *Elles sont* **indépendantes (independent)**.
2. *Elles ont* **même variance** σ^2 *inconnue. C'est la condition d'* **homogénéité (homogeneity)** *ou d'* **homoscédasticité (homoscedasticity)**.
3. *Elles sont de* **loi gaussienne (normal distribution)**.

Nous pouvons donc écrire le modèle :

$$\mathcal{L}(Y_{ij}) = \mathcal{N}(\mu_i \ ; \ \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

Ainsi nous constatons que, si les lois $\mathcal{L}(Y_{ij})$ sont différentes, elles ne peuvent différer que par leur moyenne théorique. Il y a donc un simple décalage entre elles.

Remarque 2.1.5. Dans certains livres de statistique, les auteurs écrivent le modèle statistique de la façon suivante :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où

$$\sum_{i=1}^{I} \alpha_i = 0$$

et

$$\mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0 \ ; \ \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

Nous avons donc la correspondance suivante :

$$\mu_i = \mu + \alpha_i \quad i = 1, \ldots, I.$$

Les deux modèles sont donc statistiquement équivalents.

Test de comparaison 2.1.1. Nous nous proposons de tester l'hypothèse :

$$(\mathcal{H}_0) : \mu_1 = \mu_2 = \cdots = \mu_I$$

contre

$$(\mathcal{H}_1) : \text{Les moyennes } \mu_i \text{ ne sont pas toutes égales.}$$

La méthode statistique qui permet d'effectuer ce test est appelée l'**Analyse de la Variance à un Facteur (one way Analysis of Variance)**.

En effet la comparaison des moyennes théoriques s'effectue à partir de la dispersion des moyennes observées comparée à la dispersion des données dans leur ensemble. Elle a été introduite par R. A. Fisher (1918, 1925, 1935).

2.2. Tableau de l'Analyse de la Variance - Test (cas équilibré)

Le test est fondé sur deux propriétés des moyennes et des variances.

Propriété 2.2.1. *La moyenne de toutes les observations est la moyenne des moyennes de chaque échantillon. Ceci s'écrit :*

$$\overline{y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{I} y_{ij} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} = \frac{1}{I} \sum_{i=1}^{I} \overline{y}_i. \tag{2.2.1}$$

Retour à l'Exemple 2.1.1. Pour cet exemple, nous constatons cette propriété. En effet, nous avons :

$$\overline{y} = \frac{1}{12} \times 417 = \frac{1}{4}(41,333 + 38,333 + 28,333 + 31,000) = \frac{1}{4} \times 139 = 34,750,$$

puisque $n = 12 = I \times J = 4 \times 3$.

Propriété 2.2.2. *La variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances. Ceci s'écrit :*

$$s^2(y) = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2 = \frac{1}{I}\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2 + \frac{1}{I}\sum_{i=1}^{I} s_i^2(y). \tag{2.2.2}$$

Retour à l'Exemple 2.1.1. Pour cet exemple, un calcul simple nous donne :

$$s^2(y) = 32,854.$$

D'autre part, nous constatons que la variance des moyennes est égale à :

$$\frac{1}{I}\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2 = \frac{1}{4}\left((41,333 - 34,750)^2 + (38,333 - 34,750)^2 + (28,333 - 34,750)^2 + (31,000 - 34,750)^2\right) = 27,854,$$

que la moyenne des variances est égale à :

$$\frac{1}{I}\sum_{i=1}^{I} s_i^2(y) = \frac{1}{4}(1,556 + 2,889 + 13,556 + 2,000) = 5,000.$$

En faisant la somme des deux derniers résultats, nous retrouvons bien la valeur de $32,854$ que nous avons obtenue par le calcul simple. Donc la relation (2.2.2) est bien vérifiée.

Remarque 2.2.1. En multipliant les deux membres par n de l'équation (2.2.2), nous obtenons :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2 = J\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2 + \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij} - \overline{y}_i)^2\right)$$

ou encore ce qui s'écrit :

$$SC_{Tot} = SC_F + SC_R. \tag{2.2.3}$$

Retour à l'Exemple 2.1.1 Dans cet exemple, nous avons d'une part

$$SC_{Tot} = 394,248$$

et d'autre part

$$SC_F = 334,248 \quad \text{et} \quad SC_R = 60,000.$$

Donc lorsque nous faisons la somme des deux derniers résulats nous retrouvons bien la valeur du premier résultat. Donc la relation (2.2.3) est bien vérifiée.

Définition 2.2.1. *Nous appelons* **variation totale (total variation)** *le terme :*

$$SC_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2. \tag{2.2.4}$$

Il indique la dispersion des données autour de la moyenne générale.

Définition 2.2.2. *Nous appelons* **variation due au facteur (variation between)** *le terme :*

$$SC_F = J\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2. \tag{2.2.5}$$

Il indique la dispersion des moyennes autour de la moyenne générale.

Définition 2.2.3. *Nous appelons* **variation résiduelle (variation within)** *le terme :*

$$SC_R = \sum_{i=1}^{I} \left(\sum_{j=1}^{J} (y_{ij} - \overline{y}_i)^2 \right). \tag{2.2.6}$$

Il indique la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

Principe du test : Si l'hypothèse nulle ($\mathcal{H}_0$) est vraie alors la quantité SC_F doit être petite par rapport à la quantité SC_R. Par contre, si l'hypothèse alternative ($\mathcal{H}_1$) est vraie alors la quantité SC_F doit être grande par rapport à la quantité SC_R. Pour comparer ces quantités, R. A. Fisher, après les avoir « corrigées »par leurs degrés de liberté (*ddl*), a considéré leur rapport.

Propriété 2.2.3. *Nous appelons* **variance due au facteur** *le terme*

$$s_F^2 = \frac{SC_F}{I-1} \tag{2.2.7}$$

et **variance résiduelle** *le terme*

$$s_R^2 = \frac{SC_R}{n-I}. \tag{2.2.8}$$

Si les **trois conditions 2.1.1.** *sont satisfaites et si l'hypothèse nulle ($\mathcal{H}_0$) est vraie alors*

$$F_{obs} = \frac{s_F^2}{s_R^2} \tag{2.2.9}$$

est une réalisation d'une variable aléatoire F qui suit une loi de Fisher à $I-1$ degrés de liberté au numérateur et $n-I$ degrés de liberté au dénominateur. Cette loi est notée $\mathcal{F}_{I-1,n-I}$.

Décision 2.2.1. *Pour un seuil donné α (=5%=0,05 en général), les tables de Fisher nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[F \leqslant c] = 1-\alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } c \leqslant F_{obs} & (\mathcal{H}_1) \text{ est vraie}, \\ \text{si } F_{obs} < c & (\mathcal{H}_0) \text{ est vraie}. \end{cases}$$

L'ensemble de la procédure est résumé par un tableau, appelé **tableau de l'Analyse de la Variance (analysis of variance table)**, du type suivant :

Variation	SC	ddl	s^2	F_{obs}	F_c
Due au facteur	SC_F	$I-1$	s_F^2	$\frac{s_F^2}{s_R^2}$	c
Résiduelle	SC_R	$n-I$	s_R^2		
Totale	SC_{Tot}	$n-1$			

Retour à l'Exemple 2.1.1. Pour les données de cet exemple, le tableau de l'Analyse de la Variance s'écrit :

Variation	SC	ddl	s^2	F_{obs}	F_c
Due au facteur	334,248	3	111,416	14,856	4,066
Résiduelle	60,000	8	7,500		
Totale	394,248	11			

Pour un seuil $\alpha = 5\% = 0,05$, les tables de Fisher nous fournissent la valeur critique $c = 4,066$. Nous décidons donc que l'hypothèse alternative ($\mathcal{H}_1$) est vraie : il y a donc des différences entre les moyennes théoriques d'hydrophilie selon la durée. **Nous en concluons que l'hydrophilie n'est pas stable.**

Remarque 2.2.2. Nous avons décidé que les moyennes théoriques sont différentes dans leur ensemble, mais nous ne savons pas exactement les différences qui sont significatives et celles qui ne le sont pas. Nous les analyserons par la suite avec des tests de comparaisons multiples (cf paragraphe 4).

Remarque 2.2.3. Le risque d'erreur de notre décision est ici le seuil, c'est-à-dire $\alpha = 5\% = 0,05$. Le risque de deuxième espèce et le risque a posteriori peuvent être évalués, mais avec une démarche complexe. Nous verrons comment par la suite (cf paragraphe 5).

2.3. Vérification des trois conditions

Nous étudions les possibilités d'évaluer la validité des **trois conditions 2.1.1.** que nous avons supposées satisfaites.

2.3.1. Indépendance.

Il n'existe pas, dans un contexte général, **de test statistique simple permettant d'étudier l'indépendance.** Ce sont les conditions de l'expérience qui nous permettront d'affirmer que nous sommes dans le cas de l'indépendance.

2.3.2. Normalité.

Nous ne pouvons pas, en général, la tester pour chaque échantillon. En effet le nombre d'observations est souvent très limité.

Retour à l'Exemple 2.1.1. Ici, nous avons trois observations pour chaque échantillon.

Cependant remarquons que si les conditions sont satisfaites et si nous notons :

$$\mathcal{E}_{ij} = Y_{ij} - \mu_i,$$

alors

$$\mathcal{L}(\mathcal{E}_{ij}) = \mathcal{N}(0\ ;\ \sigma^2),$$

alors c'est la même loi pour l'ensemble des unités. Les moyennes μ_i étant inconnues, nous les estimons par les estimateurs bien connus de la moyenne : les $\overline{Y}_i$ où $\overline{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$. Nous obtenons alors les estimations $\overline{y}_i$ Les quantités obtenues s'appellent les **résidus (residuals)** et sont notées $\widehat{e}_{ij}$. Les résidus s'expriment par :

$$\widehat{e}_{ij} = y_{ij} - \overline{y}_i, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J. \tag{2.3.1}$$

Les résidus peuvent s'interpréter comme des estimations des erreurs de mesures. Nous pouvons alors tester la normalité, avec le **test de Shapiro-Francia** (ici nous utilisons l'approximation de Weisberg and Bingham (1975) pour obtenir une valeur approchée, mais plus facile à calculer, de la statistique du test), sur l'ensemble des résidus. Nous rappelons la procédure avec **l'Exemple 2.1.1**.

Retour à l'Exemple 2.1.1. Pour effectuer le test de nomalité, nous construisons le tableau suivant :

Éponge	Durée	Hydrophilie	Résidus	Résidus classés	Rangs moyens	Fréquences cumul. cor.	Scores normaux
1	3 mois	43	1,667	-4,333	1,000	0,051	-1,635
2	3 mois	40	-1,333	-2,333	2,000	0,133	-1,114
3	3 mois	41	-0,333	-2,000	3,000	0,214	-0,792
4	6 mois	36	-2,333	-1,333	4,000	0,296	-0,536
5	6 mois	40	1,667	-0,333	5,500	0,418	-0,206
6	6 mois	39	0,667	-0,333	5,500	0,418	-0,206
7	12 mois	28	-0,333	0,667	7,000	0,541	0,103
8	12 mois	24	-4,333	1,000	8,500	0,663	0,421
9	12 mois	33	4,667	1,000	8,500	0,663	0,421
10	24 mois	32	1,000	1,667	10,500	0,827	0,941
11	24 mois	29	-2,000	1,667	10,500	0,827	0,941
12	24 mois	32	1,000	4,667	12,000	0,949	1,635

- Les **rangs moyens** notés r_{ij} sont les moyennes des rangs pour les valeurs ex æquo.

- Les **fréquences cumulées corrigées** sont données par l'expression :

$$f_{c,ij} = \frac{r_{ij} - 0,375}{n + 0,250}. \tag{2.3.2}$$

- Les **scores normaux** notés z_{ij} sont les réalisations d'une loi normale centrée réduite $\mathcal{N}(0\ ;\ 1)$ correspondant aux fréquences cumulées corrigées des résidus classés.

Remarque 2.3.1. Les **scores normaux** ne se calculent pas à la main. Pour les calculer, il faut soit avoir recours à « une table » de la loi normale (à vous de trouver celle qui est adéquate dans ce cas-là) soit se servir d'une des fonctions du logiciel **MINITAB.**

Nous notons $\widehat{\mathcal{E}}_{ij}$ la variable aléatoire dont le résidu $\widehat{e}_{ij}$ est la réalisation.

Hypothèses :

$$(\mathcal{H}_0) : \mathcal{L}(\widehat{\mathcal{E}}_{ij}) = \mathcal{N}$$

contre

$$(\mathcal{H}_1) : \mathcal{L}(\widehat{\mathcal{E}}_{ij}) \neq \mathcal{N}.$$

Statistique : En notant $\widehat{e}_{(ij)}$ les résidus classés, nous considérons l'expression :

$$r(\widehat{e}\ ;\ z) = \frac{\dfrac{1}{n}\displaystyle\sum_{i,j} \widehat{e}_{(ij)} z_{ij}}{s(\widehat{e})\, s(z)}. \tag{2.3.3}$$

Remarque 2.3.2. Par souci de simplification de notation, nous utiliserons par la suite la notation suivante :

$$r(\widehat{e}\ ;\ z) = r_{obs}.$$

Remarque 2.3.3. Nous retrouverons le nombre $r(\widehat{e}\ ;\ z)$, défini par l'expression (2.3.3) dans le chapitre suivant : il s'agit du coefficient de corrélation linéaire.

Propriété 2.3.1. *Sous l'hypothèse nulle $(\mathcal{H}_0)$ le nombre $r(\widehat{e}\ ;\ z)$ défini par* (2.3.3) *est la réalisation d'une variable aléatoire R qui suit une loi dont l'expression est très difficile à établir. En pratique, les valeurs critiques ont été calculées par simulation en fonction de n et pour trois seuils différents.*

Décision 2.3.1. *Pour un seuil donné α $(= 5\% = 0,05$ en général), les tables de Shapiro-Francia nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[R \leqslant c] = \alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } r_{obs} \leqslant c & (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } c < r_{obs} & (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Retour à l'Exemple 2.1.1. Pour un seuil $\alpha = 5\% = 0,05$, les tables de Shapiro-Francia nous fournissent, avec $n = 12$, la valeur critique $c = 0,9261$. Mais nous avons $r_{obs} = 0,985$ si nous utilisons la précision maximale de **MINITAB** ou $r_{obs} = 0,992$ si nous le calculons à partir des valeurs reportées dans le tableau précédent. Comme $c < r_{obs}$, l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, c'est-à-dire que **nous décidons que l'hypothèse de normalité est satisfaite.**

2.3.3. Homogénéité.

Plusieurs tests permettent de tester l'égalité de plusieurs variances. Parmi ceux-ci, le test le plus utilisé est le **test de Bartlett** dont le protocole est le suivant :

Hypothèses :

$$(\mathcal{H}_0) : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2$$

contre

$$(\mathcal{H}_1) : \text{Les variances } \sigma_i^2 \text{ ne sont pas toutes égales.}$$

Statistique : Nous considérons l'expression :

$$B_{obs} = \frac{1}{C_1}\left[(n-I)\ln(s_R^2) - \sum_{i=1}^{I}(n_i-1)\ln(s_{c,i}^2)\right] \tag{2.3.4}$$

où
- la quantité C_1 est définie par :

$$C_1 = 1 + \frac{1}{3(I-1)}\left(\left(\sum_{i=1}^{I}\frac{1}{n_i-1}\right) - \frac{1}{n-I}\right), \tag{2.3.5}$$

- s_R^2 est la variance résiduelle,
- $s_{c,i}^2$ la variance corrigée des observations de l'échantillon d'ordre i, $(i = 1, \ldots, I)$.

Propriété 2.3.2. *Sous l'hypothèse nulle $(\mathcal{H}_0)$ le nombre B_{obs} défini par (2.3.4) est la réalisation d'une variable aléatoire B qui suit asymptotiquement une loi du khi-deux à $I-1$ degrés de liberté.* **En pratique**, *nous pouvons l'appliquer lorsque les effectifs n_i des I échantillons sont tous au moins égaux à 3. Ce test dépend de la normalité des observations.*

Remarque 2.3.4. Notons que B_{obs} est une valeur qui est toujours positive par construction.

Décision 2.3.2. *Pour un seuil donné α $(= 5\% = 0,05$ en général), les tables du khi-deux nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[B \leqslant c] = 1-\alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } c \leqslant B_{obs} & (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } B_{obs} < c & (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Retour à l'Exemple 2.1.1. Pour les données de cet exemple, nous avons :

$$C_1 = 1 + \frac{1}{3(4-1)}\left(\left(\sum_{i=1}^{4}\frac{1}{3-1}\right) - \frac{1}{12-4}\right) = 1,208.$$

Par conséquent, en se souvenant que **les n_i sont tous égaux dans cet exemple**, nous avons :

$$\begin{aligned} B_{obs} &= \frac{1}{1,208}\left[(12-4)\ln(7,500) - (3-1)\left(\ln\left(\frac{3}{2}\times 1,556\right) + \ln\left(\frac{3}{2}\times 2,889\right) + \ln\left(\frac{3}{2}\times 13,556\right) + \ln\left(\frac{3}{2}\times 2,000\right)\right)\right] \\ &= 2,706. \end{aligned}$$

Pour un seuil $\alpha = 5\% = 0,05$ la valeur critique d'un khi-deux à 3 degrés de liberté, est $c = 7,815$. Comme $B_{obs} < c$, nous décidons que l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, c'est-à-dire que l'hypothèse d'homogénéité des variances est vérifiée.

2.4. Comparaisons multiples

Lorsque pour la comparaison des moyennes théoriques la décision est « l'hypothèse alternative $(\mathcal{H}_1)$ est vraie », pour analyser les différences nous procédons à des tests qui comparent les moyennes entre elles. Ce sont les tests de comparaisons multiples, adaptations du test de Student. Un des tests les plus connus est : **le test de Tukey**[1].

2.4.1. Le test de Tukey

Les moyennes observées $\overline{y}_i$ sont rangées par ordre croissant. Nous les notons alors $\overline{y}_{(1)},\ \overline{y}_{(2)}, \ldots,\ \overline{y}_{(I)}$, et les moyennes théoriques associées $\mu_{(1)},\ \mu_{(2)}, \ldots,\ \mu_{(I)}$. **La procédure du test de Tukey est la suivante :**

Pour chaque $i < i'$, nous considérons les
Hypothèses :

$$(\mathcal{H}_0)\ :\ \mu_{(i)} = \mu_{(i')}$$

contre

$$(\mathcal{H}_1)\ :\ \mu_{(i')} > \mu_{(i)}.$$

Statistique : Nous considérons le rapport :

$$t_{i',i,obs} = \frac{\overline{y}_{(i')} - \overline{y}_{(i)}}{\sqrt{\frac{s_R^2}{2}\left(\frac{1}{n_{i'}} + \frac{1}{n_i}\right)}}. \tag{2.4.1}$$

Propriété 2.4.1. *Le rapport $t_{i',i,obs}$ défini par* (2.4.1) *est la réalisation d'une variable aléatoire T qui, si l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, suit une loi appelée* **étendue studentisée (studentized range)** *et que nous notons $\widetilde{\mathcal{T}}_{n-I,I}$.*

Décision 2.4.1. *Pour un seuil donné α $(= 5\% = 0,05$ en général), les tables de l'étendue studentisée nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[T \leqslant c] = 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } c \leqslant t_{i',i,obs} & (\mathcal{H}_1) \text{ est vraie}, \\ \text{si } t_{i',i,obs} < c & (\mathcal{H}_0) \text{ est vraie}. \end{cases}$$

Remarque 2.4.1. La valeur critique c ne dépend que des indices $n - I$, degrés de liberté de la somme des carrés résiduelle, et de I, nombre des moyennes comparées. De plus, les moyennes théoriques, dont les moyennes observées sont comprises entre deux moyennes observées, dont les moyennes théoriques correspondantes sont déclarées égales, sont déclarées égales avec ces dernières.

Retour à l'Exemple 2.1.1. Nous avons les moyennes dans l'ordre :

$$\overline{y}_{(1)} = 28,333 \quad \overline{y}_{(2)} = 31,000 \quad \overline{y}_{(3)} = 38,333 \quad \overline{y}_{(4)} = 41,333.$$

Le tableau de l'Analyse de la Variance à un facteur nous donne la variance résiduelle égale à :

$$s_R^2 = 7,500,$$

$$n - I = 8 \quad \text{et} \quad I = 4.$$

Dans ce cas nous avons :

$$n_1 = n_2 = n_3 = n_4 = 3.$$

Nous remarquons que le dénominateur est le même pour toutes les comparaisons :

$$\sqrt{\frac{7,500}{2}\left(\frac{1}{3} + \frac{1}{3}\right)} = 1,581.$$

[1] En fait, le nom exact de ce test est le test de Tukey-Kramer mais pour plus de simplicité, et comme le logiciel **MINITAB** le nomme aussi le test de Tukey, nous utiliserons la même appelation que le logiciel **MINITAB**. Le test de Tukey est initialement prévu que dans le cas où le plan expérimental est équilibré.

Les tables de l'étendue studentisée nous fournissent la valeur critique $c = 4,5288$. Nous pouvons ainsi dresser le tableau de toutes les comparaisons :

Comparaisons	$\overline{y}_{(i')} - \overline{y}_{(i)}$	Dénominateurs	$t_{i',i}$	Décisions
$\mu_{3 \text{ mois}} = \mu_{12 \text{ mois}}$	$13,000$	$1,581$	$8,223$	$\mu_{3 \text{ mois}} > \mu_{12 \text{ mois}}$
$\mu_{3 \text{ mois}} = \mu_{24 \text{ mois}}$	$10,333$	$1,581$	$6,536$	$\mu_{3 \text{ mois}} > \mu_{24 \text{ mois}}$
$\mu_{3 \text{ mois}} = \mu_{6 \text{ mois}}$	$3,000$	$1,581$	$1,898$	$\mu_{3 \text{ mois}} = \mu_{6 \text{ mois}}$
$\mu_{6 \text{ mois}} = \mu_{12 \text{ mois}}$	$10,000$	$1,581$	$6,325$	$\mu_{6 \text{ mois}} > \mu_{12 \text{ mois}}$
$\mu_{6 \text{ mois}} = \mu_{24 \text{ mois}}$	$7,333$	$1,581$	$4,638$	$\mu_{6 \text{ mois}} > \mu_{24 \text{ mois}}$
$\mu_{24 \text{ mois}} = \mu_{12 \text{ mois}}$	$2,667$	$1,581$	$1,687$	$\mu_{24 \text{ mois}} = \mu_{12 \text{ mois}}$

Remarque 2.4.2. Il est à noter que le logiciel **MINITAB** adopte une autre procédure. Les moyennes observées ne sont pas rangées par ordre croissant. Les statistiques calculées par le logiciel **MINITAB** sont $\frac{t_{i',i,obs}}{\sqrt{2}}$ et peuvent donc être positives ou négatives, alors que dans la procédure que nous proposons toutes les quantitées calculées seront toujours positives. Ce sont les $P-$valeurs qui permettent de prendre les décisions finales du test.

La synthèse des différentes décisions est généralement présentée sous la forme d'un tableau dans lequel les espérances considérées comme égales sont classées dans un même type, notées par la même lettre (A, ou B, ...).

Retour à l'Exemple 2.1.1. Dans cet exemple nous en déduisons le tableau :

Moyennes	Résultats des tests	
$\mu_{3 \text{ mois}}$	A	
$\mu_{6 \text{ mois}}$	A	
$\mu_{24 \text{ mois}}$		B
$\mu_{12 \text{ mois}}$		B

2.4.2. Le test de Dunnett

Dans le cas où l'une des populations est considérée comme **référence**, si l'analyse de la variance a mis en évidence un effet du facteur étudié, **le test de Dunnett** permet la comparaison des effets entre les différentes modalités du facteur avec la modalité « référence », qui est représentée ici par l'indice 0. Le principe est le même que celui du **test de Tukey** que nous venons d'étudier ci-dessus.

Hypothèses :

$$(\mathcal{H}_0) \ : \ \mu_0 = \mu_i$$

contre

$$(\mathcal{H}_1) \ : \ \mu_0 \neq \mu_i.$$

Statistique : Nous considérons le rapport :

$$t_{0,i,obs} = \frac{\overline{y}_0 - \overline{y}_i}{\sqrt{s_R^2 \left(\frac{1}{n_0} + \frac{1}{n_i} \right)}}. \tag{2.4.2}$$

Propriété 2.4.2. *Le rapport $t_{0,i,obs}$ défini par* (2.4.2) *est la réalisation d'une variable aléatoire T qui, lorsque l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, suit une loi dite de* **Dunnett**, *est notée $\mathcal{D}_{n-I,I}$.*

Décision 2.4.2. *Pour un seuil donné α $(= 5\% = 0,05$ en général), les tables de Dunnett nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[-c \leqslant T \leqslant c] = 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } t_{0,i,obs} \leqslant -c, \text{ ou si } c \leqslant t_{0,i,obs} & (\mathcal{H}_1) \text{ est vraie}, \\ \text{si } -c < t_{0,i,obs} < c & (\mathcal{H}_0) \text{ est vraie}. \end{cases}$$

Remarque 2.4.3. Notons que nous avons également des tests de ce type unilatéraux.

2.5. Risque de deuxième espèce et risque a posteriori

Lorsque nous décidons que l'hypothèse nulle ($\mathcal{H}_0$) est vraie pour le Test 2.1.1 (voir Décision 2.2.1) de la l'Analyse de la Variance, il est utile de connaître le risque de mauvaise décision appelé **risque de deuxième espèce** ou encore risque β. Celui-ci est égal à la probabilité de décider l'hypothèse nulle ($\mathcal{H}_0$) vraie alors qu'en réalité c'est l'hypothèse alternative ($\mathcal{H}_1$) qui est vraie. Fixons dans l'hypothèse alternative ($\mathcal{H}_1$) des moyennes théoriques différentes $\mu_1,\ \mu_2, \ldots,\ \mu_I$.

Pour **calculer le risque** β nous utilisons des abaques et l'expression :

$$\phi = \sqrt{\frac{J}{I\sigma^2}\sum_{i=1}^{I}(\mu_i - \overline{\mu})^2} \tag{2.5.1}$$

où

$$\overline{\mu} = \frac{1}{I}\sum_{i=1}^{I}\mu_i.$$

En général il est impossible de calculer la quantité ϕ car il faudrait connaître la variance σ^2. Nous la remplaçons alors par la meilleure estimation dont nous disposons à savoir par l'estimateur s_R^2. Nous avons alors une évaluation de ϕ. Les **abaques** pour $\nu_1 = I - 1,\ \nu_2 = n - I$ et ϕ nous permettent d'évaluer $1 - \beta$, qui est la **puissance** (**power**) du test.

Il arrive que l'utilisateur n'a pas d'idée a priori sur les moyennes théoriques μ_i. Dans ce cas il peut fonder son calcul du risque β sur les moyennes observées, ce qui revient à supposer que par un hasard extraordinaire nous avons observé les vraies moyennes. Nous avons alors le **risque a posteriori**. Pour ce faire nous calculons, chaque fois que cela a un sens (lorsque $s_F^2 - s_R^2 \geqslant 0$)

$$\phi = \sqrt{\frac{(I-1)(s_F^2 - s_R^2)}{Is_R^2}}. \tag{2.5.2}$$

Nous utilisons les mêmes abaques.

Ces mêmes calculs et ces mêmes abaques, permettent de déterminer approximativement le nombres d'observations nécesaires pour atteindre un risque β donné. **En général la valeur de 0,20 est considérée comme satisfaisante.**

2.6. Transformations

Lorsque la normalité ou l'homogénéité ne sont pas vérifiées, nous pouvons tenter d'utiliser des transformations afin de vérifier ces deux conditions. Les transformations habituelles sont :

$$x_{ij} = \log(y_{ij})$$

ou

$$x_{ij} = (y_{ij})^{\lambda}.$$

La difficulté consiste à déterminer, dans le deuxième cas la « bonne » puissance λ. Lorsque les donnés de l'expérimentateur sont des proportions nous pouvons utiliser la transformation décrite dans la propriété suivante :

Propriété 2.6.1. *Si pour une variable Y nous avons $\mathcal{L}(Y) = \mathcal{B}(n\ ;\ p)$ alors nous obtenons le résultat :*

$$\lim_{n\longrightarrow+\infty}\mathcal{L}\left(\sqrt{n}\left(\left(\arcsin\left(\sqrt{\frac{Y}{n}}\right) - \arcsin(\sqrt{2p})\right)\right)\right) = \mathcal{N}(0\ ;\ 1). \tag{2.6.1}$$

Ainsi nous pouvons grâce à cette transformation faire une analyse de la variance sur des proportions, mais à condition qu'elles aient été calculées sur le même nombre d'observations.

Mais une transformation n'est acceptable et utilisable que si elle admet une interprétation concrète.
Ainsi :

- la transformation

$$x = a + by$$

  est un changement d'origine et d'échelle.

- La transformation
$$x = a + b\ln(y)$$
peut être utilisée pour changer des effets multiplicatifs en effets additifs.
- La transformation
$$x = \exp(a + by)$$
permet de décrire des phénomènes qui ont un comportement précis dans le temps.

Si malgré toutes les tentatives de transformations, les conditions (c'est-à-dire la normalité des résidus et l'homogénéité des résidus) ne sont pas satisfaites, alors il faut utiliser le **test non paramétrique de Kruskal-Wallis** (cf paragraphe 2.8).

2.7. Facteurs aléatoires

Dans certaines expériences il arrive que les modalités du facteur ne soient pas déterminées de manière précise.

Exemple 2.7.1. Elles peuvent correspondre à des individus (expérimentateurs), à des sites (laboratoires), etc.

Dans ce cas nous sommes obligés de recourir à un modèle du type :

$$\mathcal{L}(Y_{ij}) = \mu + A_i + \mathcal{E}_{ij} \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

où
- μ est la moyenne générale,
- $\mathcal{L}(A_i) = \mathcal{N}(0\ ;\ \sigma_A^2)$ correspond à un facteur aléatoire,
- $\mathcal{L}(\mathcal{E}_{ij}) = \mathcal{N}(0\ ;\ \sigma^2)$ correspond au terme d'erreur.

Le test sur l'effet du facteur s'exprime alors de la façon suivante :

Hypothèses :
$$(\mathcal{H}_0) : \sigma_A^2 = 0$$
contre
$$(\mathcal{H}_1) : \sigma_A^2 \neq 0.$$

La procédure de décision est identique au cas non aléatoire. Cependant les comparaisons multiples n'ont pas de sens dans ce cas et ne sont donc pas effectuées. De plus, les conditions sont plus difficiles à vérifier et le risque β plus compliqué à évaluer. Pour de plus amples renseignements sur ce sujet, nous renvoyons au cours de Frédéric Bertrand, téléchargeable à l'adresse suivante :
`http ://www-irma.u-strasbg.fr/~fbertran/enseignement/Master1_2006/Master1ModelesANOVA.pdf`

2.8. Analyse de la Variance non paramétrique - Test de Kruskal-Wallis

Lorsque les conditions d'une analyse de la variance paramétrique (lois gaussiennes homogènes) ne sont pas satisfaites, nous utilisons une procédure de test plus générale. C'est une procédure dite **non paramétrique** (**non parametric**). Elle s'applique dans tous les cas où les observations peuvent être classées, mais surtout elle est adaptée pour des lois continues non gaussiennes.

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

Hypothèses :
$$(\mathcal{H}_0) : \mathcal{L}_1 = \ldots = \mathcal{L}_I$$
contre
$$(\mathcal{H}_1) : \text{Les lois } \mathcal{L}_i \text{ ne sont pas égales.}$$

Nous observons un n_i-échantillon de valeurs indépendantes dans la population d'ordre i, et ceci pour $i = 1, \ldots, I$. Nous classons l'ensemble des observations en un **seul** échantillon, R_{ij} désignant le rang de l'observation y_{ij}. Nous posons :

$$R_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}, \quad i = 1, \ldots, I. \tag{2.8.1}$$

Ce sont les moyennes des rangs, dans le classement général, de toutes les observations de chaque échantillon. Il est facile de remarquer que, si $n = n_1 + \ldots + n_I$, nous avons :

$$\overline{R} = \frac{1}{n} \sum_{i=1}^{I} n_i R_{i\bullet} = \frac{n+1}{2}. \tag{2.8.2}$$

Cette quantité s'appelle la moyenne générale.

2.8.1. Cas où il n'y a pas d'ex-æquo

Dans ce paragraphe, les observations seront toutes distinctes.

Nous considérons comme

Statistique : L'écart, au carré, des rangs moyens à leur moyenne générale, divisé par leur variance, qui s'écrit de la façon suivante :

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{I} n_i \left(R_{i\bullet} - \frac{n+1}{2} \right)^2. \tag{2.8.3}$$

C'est la **statistique de Kruskal-Wallis** qui est tout-à-fait analogue à celle de l'analyse de la variance à un facteur, mais avec une variance connue.

Si nous posons

$$R_i = n_i R_{i\bullet},$$

somme des rangs des observations de l'échantillon d'ordre i, alors la statistique K définie par (2.8.3) peut s'écrire plus simplement :

$$K = \left(\frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} \right) - 3(n+1). \tag{2.8.4}$$

Remarque 2.8.1. C'est cette dernière expression de la statistique K qu'il faudra utiliser la plupart du temps pour faire les calculs « à la main ».

Propriété 2.8.1. *Lorsque l'hypothèse nulle $(\mathcal{H}_0)$ est vraie et lorsque $n \longrightarrow +\infty$ nous avons :*

$$\mathcal{L}_{(\mathcal{H}_0)}(K) = \chi^2_{I-1},$$

loi du khi-deux à $(I-1)$ degrés de liberté.

Il existe des tables exactes de la statistique K de Kruskal-Wallis, sous l'hypothèse nulle $(\mathcal{H}_0)$, mais nous utiliserons, la plupart du temps, la loi asymptotique ci-dessus.

Décision 2.8.1. *Pour un seuil donné α ($= 5\% = 0,05$ en général), les tables du khi-deux nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[K < c] \geqslant 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } c \leqslant K_{obs} & (\mathcal{H}_1) \text{ est vraie}, \\ \text{si } K_{obs} < c & (\mathcal{H}_0) \text{ est vraie}. \end{cases}$$

Exemple 2.8.1. Nous étudions la durée de l'activité de trois types différents de subtances. Nous avons prélevé au hasard un échantillon de chaque type, et nous avons noté la durée de l'activité. Voici les résultats :

Type	A	B	C
Durées de l'activité	73	84	82
	64	80	79
	67	81	71
	62	77	75
	70		

Il est connu par ailleurs qu'une variable aléatoire qui exprime une durée est rarement gaussienne. **En général**, sa loi est asymétrique et peut être parfois ajustée par :
- une loi de type log-normale,
- une loi de type exponentielle,
- une loi de type gamma,
- une loi de type Weibull
- ou encore, dans des cas extrêmes, par une loi de type Gumbel.

C'est pourquoi nous appliquons la procédure non paramétrique. Comme il s'agit de trois échantillons (supposés indépendants) nous utilisons le **test de Kruskal-Wallis**.

Nous notons Y la variable « durée de l'activité » et $\mathcal{L}_A,\ \mathcal{L}_B$ et $\mathcal{L}_C$ sa loi sur les populations $A,\ B$ et C respectivement.

Nous testons :

$$(\mathcal{H}_0) : \mathcal{L}_A = \mathcal{L}_B = \mathcal{L}_C$$

contre

$$(\mathcal{H}_1) : \mathcal{L}_A,\ \mathcal{L}_B,\ \mathcal{L}_C \text{ ne sont pas égales.}$$

Nous classons les observations en un seul échantillon et nous déterminons leur rang. Les résultats sont donnés dans le tableau suivant :

Échantillon A	62	64	67	70		73							
Échantillon B								77		80	81		84
Échantillon C					71		75		79			82	
Rangs	1	2	3	4	5	6	7	8	9	10	11	12	13

Pour calculer la statistique K du **test de Kruskal-Wallis** nous complétons le tableau suivant :

	R_i	R_i/n_i	R_i^2
Échantillon A	16	3,20	256
Échantillon B	42	10,50	1764
Échantillon C	33	8,25	1089

En utilisant l'expression (2.8.4), nous pouvons alors calculer la statistique K du **test de Kruskal-Wallis** :

$$\begin{aligned} K_{obs} &= \left(\frac{12}{13(13+1)}\left(\frac{256}{5}+\frac{1764}{4}+\frac{1089}{4}\right)\right) - 3(13+1) \\ &= 8,403. \end{aligned}$$

Pour un seuil $\alpha = 5\% = 0,05$, les tables du khi-deux à $I-1 = 3-1 = 2$ degrés de liberté nous fournissent la valeur critique 5,991. Comme $c < K_{obs}$, nous décidons que l'hypothèse alternative $(\mathcal{H}_1)$ est vraie, c'est-à-dire que **nous décidons que les lois $\mathcal{L}_A$, $\mathcal{L}_B$ et $\mathcal{L}_C$ ne sont pas égales.**

2.8.2. Cas où il y a des ex æquo

Lorsqu'il y des observations ex æquo, nous utilisons les rangs moyens. La statistique K du **test de Kruskal-Wallis** doit être alors **corrigée**. Elle devient :

$$K^* = \frac{\left(\frac{12}{n(n+1)}\sum_{i=1}^{I}\frac{R_i^2}{n_i}\right) - 3(n+1)}{1 - \frac{1}{n^3-n}\sum_{h=1}^{n}(d_h^3 - d_h)}, \tag{2.8.5}$$

où $(d_1, \ldots, d_n)$ est la configuration observée, c'est-à-dire chaque d_h est le nombre d'observations égales à l'observation d'ordre h.

Pour la décision, nous procédons de la même manière que dans le cas où il n'y a pas d'ex æquo.

2.9. Quelques précisions sur les comparaisons multiples

Dans le cas où la décision a été : « l'hypothèse alternative $(\mathcal{H}_1)$ est vraie », il y a une possibilité de comparaisons multiples. Désignons par :

$$c(\alpha, n-I)$$

la valeur critique positive pour un test bilatéral au seuil α avec la loi de Student à $(n-I)$ degrés de liberté.

Nous pouvons alors comparer deux à deux les populations en décidant que la population d'ordre i est différente de la population d'ordre i' si

$$\left|\frac{R_i}{n_i} - \frac{R_{i'}}{n_{i'}}\right| > c(\alpha, n-I)\sqrt{S^2 \times \left(\frac{n-1-K}{n-I}\right)}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}, \tag{2.9.1}$$

où

- dans le cas sans ex æquo

$$S^2 = \frac{n(n+1)}{12}$$

- dans le cas avec ex æquo

$$S^2 = \frac{1}{n-1}\left(\sum_{i,j} R_{ij}^2 - n\frac{(n+1)^2}{4}\right).$$

Retour à l'Exemple 2.8.1. Nous organisons les comparaisons multiples de la manière suivante. Nous devons calculer le terme :

$$c(5\%, 13-3)\sqrt{\frac{13(13+1)}{12} \times \left(\frac{13-1-8,403}{13-3}\right)} = 5,204, \quad \text{où} \quad c(5\%, 10) = 2,2281.$$

Ce terme doit être multiplié par :

$$\sqrt{\frac{1}{5} + \frac{1}{4}} = 0,671$$

lorsque l'on compare $\mathcal{L}_A$ et $\mathcal{L}_B$ les lois de Y sur les populations A (l'échantillon A a un effectif égal à 5) et B (l'échantillon B a un effectif égal à 4) ou $\mathcal{L}_A$ et $\mathcal{L}_C$ les lois de Y sur les populations A (l'échantillon A a un effectif égal à 5) et C (l'échantillon C a un effectif égal à 4) ou par

$$\sqrt{\frac{1}{4} + \frac{1}{4}} = 0,707$$

lorsque l'on compare $\mathcal{L}_B$ et $\mathcal{L}_C$ les lois de Y sur les populations B (l'échantillon B a un effectif égal à 4) et C (l'échantillon C a un effectif égal à 4).

Ceci nous donne ainsi les valeurs critiques que nous utilisons dans les comparaisons multiples :

$$3,492$$

et

$$3,679.$$

Nous effectuons ces comparaisons dans le tableau suivant :

Comparaisons	$\left\lvert \frac{R_i}{n_i} - \frac{R_{i'}}{n_{i'}} \right\rvert$	c	Décision
$\mathcal{L}_A$ avec $\mathcal{L}_B$	7,300	3,492	$\mathcal{L}_A \neq \mathcal{L}_B$
$\mathcal{L}_B$ avec $\mathcal{L}_C$	2,250	3,679	$\mathcal{L}_B = \mathcal{L}_C$
$\mathcal{L}_A$ avec $\mathcal{L}_C$	5,050	3,492	$\mathcal{L}_A \neq \mathcal{L}_C$

Comme pour le cas paramétrique, nous résumons les résultats dans le tableau :

Populations	Classes d'égalité	
B	$*$	
C	$*$	
A		$*$

Ceci complète l'**Exemple 2.8.1.**.

Exemple 2.9.1. Le nombre de caries, pondérées par leur gravité, a été observé sur quatre groupes d'animaux :
• un groupe témoin A,
• et trois groupes B, C, D d'animaux subissant trois traitements différents.
Pouvons-nous tester l'hypothèse selon laquelle les traitements n'ont pas d'influence ?

Pour simplifier les calculs les résultats sont fournis classés par ordre croissant dans chaque groupe :

Groupe A	30	32	34	36	36	37	42	42	46	50
Groupe B	24	34	34	35	37	40	42	42	44	50
Groupe C	26	28	30	32	33	36	38	40	42	44
Groupe D	30	32	32	34	36	40	42	46	46	48

Nous admettons que le nombre de caries pondérées par leur gravité ne suit pas une loi normale (encore qu'un test de normalité indique le contraire, mais avec un risque inconnu). C'est pourquoi nous appliquons **le test de Kruskal-Wallis**.

Nous testons l'hypothèse :

$$(\mathcal{H}_0) : \text{Les traitements sont identiques}$$

contre

$$(\mathcal{H}_1) : \text{Les traitements sont différents.}$$

Nous classons les observations en un seul échantillon, puis nous déterminons les rangs moyens et la configuration.

Groupe A				30	32		34		36 36
Groupe B	24						34 34	35	
Groupe C		26	28	30	32	33			36
Groupe D				30	32 32		34		36
Rangs moyens	1	2	3	5	8,5	11	13,5	16	18,5
Configuration	1	1	1	3	4	1	4	1	4

Groupe A	37			42 42		46		50
Groupe B	37		40	42 42	44			50
Groupe C		38	40	42	44			
Groupe D			40	42		46 46	48	
Rangs moyens	21,5	23	25	29,5	33,5	36	38	39,5
Configuration	2	1	3	6	2	3	1	2

Il est facile alors de constater que les sommes des rangs moyens des groupes sont respectivement :

$$220 \quad 222 \quad 159 \quad 219.$$

Ce qui nous permet de calculer la statistique K du **test de Kruskal-Wallis** :

$$\begin{aligned} K_{obs} &= \frac{12}{40(40+1)}\left(\frac{220^2}{10}+\frac{222^2}{10}+\frac{159^2}{10}+\frac{219^2}{10}\right)-3(40+1) \\ &= 2,068. \end{aligned}$$

Mais la présence d'ex æquo nécessite la correction de cette statistique K_{obs} par :

$$1-\frac{1}{40^3-40}\left(3(2^3-2)+3(3^3-3)+3(4^3-4)+(6^3-6)\right)=0,992.$$

Ce qui donne la statistique K^* du **test de Kruskal-Wallis** corrigée :

$$K^*_{obs}=\frac{2,068}{0,992}=2,084 \quad \text{ou} \quad =2,051,$$

en fonction des deux cas à savoir si vous avez gardé en mémoire les valeurs numériques ou si vous faites les calculs au fur et à mesure avec les valeurs numériques données auparavant. Nous admettons que nous pouvons approcher la loi $\mathcal{L}_{\mathcal{H}_0}(K^*)$ par une loi du χ^2_{4-1}. Pour un seuil $\alpha=5\%=0,05$, les tables du khi-deux nous fournissent la valeur critique $c=7,815$. Comme $K^*_{obs}<c$, **nous décidons ainsi que l'hypothèse nulle** $(\mathcal{H}_0)$ **est vraie, c'est-à-dire qu'il n'y a pas de différence entre les traitements.**

Chapitre 3

Analyse de régression linéaire : Corrélation linéaire - Régression linéaire simple

3.1. Introduction

Contrairement au chapitre précédent, dans celui-ci nous supposons que sur chaque unité de notre échantillon nous observons les réalisations de deux variables aléatoires quantitatives notées sous la forme d'un couple (X, Y).

Exemple 3.1.1. Nous étudions la teneur du sang en cholestérol en fonction de l'âge, pour une population donnée. Le tableau suivant donne les résultats pour un échantillon de 11 femmes habitant le nord de la France.

Individu	Âge X (ans)	Cholestérol Y (cg/L)
1	46	181
2	52	228
3	39	182
4	65	249
5	54	259
6	33	201
7	49	121
8	76	339
9	71	224
10	41	112
11	58	189

Remarquons qu'a priori, les 11 personnes étant choisies au hasard dans la population, il est impossible de prévoir exactement leur âge. Notons également qu'en général pour chaque valeur de la variable X, il correspond qu'une seule valeur de la variable Y. Pour noter cette situation, que nous appelons **plan expérimental 1**, nous pouvons utiliser :

$$(x_i, y_i), \quad i = 1, \ldots, n.$$

Dans cet exemple, notons que $n = 11$.

Exemple 3.1.2. La concentration en hormone de croissance a été mesurée sur 12 rats. Ces mesures ont été effectuées à un temps donné, noté X et mesuré en minutes, après une injection intraveineuse. La concentration est notée Y et mesurée en g/mL de plasma. Au cours de l'expérience, chaque rat ne subit qu'un seul prélèvement. Voici les résultats :

Temps X (mn)	Hormone Y (g/mL)		
2	462,5	533,0	456,0
4	396,0	324,0	302,0
8	159,0	214,0	176,0
12	126,0	120,0	108,0

La forme standard empilée du tableau de données est la suivante :

Temps X (mn)	Hormone Y (g/mL)
2	462,5
2	533,0
2	456,0
4	396,0
4	324,0
4	302,0
8	159,0
8	214,0
8	176,0
12	126,0
12	120,0
12	108,0

Dans cette expérience non seulement pour chaque valeur de la variable X il correspond plusieurs valeurs de la variable Y (trois dans notre cas), mais la variable X est totalement contrôlée par l'expérimentateur, qui en fixe les valeurs. Les observations pour ce type d'expérimentation, que nous appélerons **plan expérimental 2**, peuvent être notées :

$$(x_i,\ y_{ij}), \quad j = 1, \ldots, J \text{ (ou } n_i),\ i = 1, \ldots, I.$$

Dans ce dernier exemple, notons que $J = 3$, $I = 4$ et par conséquent $n = IJ = 12$.

Ce plan expérimental est, bien sûr, à rapprocher d'une analyse de la variance à un facteur, mais ici **le facteur est quantitatif**. Il est facile de constater, en répétant les valeurs que peut prendre la variable X, qu'il est possible d'utiliser pour ce deuxième cas les notations du premier plan expérimental.

Objectifs. Lors de telles démarches, l'expérimentateur poursuit l'un des deux objectifs suivants :

1. Exprimer l'intensité de la liaison éventuelle entre les deux variables X et Y. Il s'agit alors d'utiliser les procédures de **corrélation**. Les deux variables jouent un rôle symétrique dans les calculs (**plan expérimental 1**).

2. Analyser le type de dépendance de Y en fonction de X, avec l'aide d'une fonction mathématique comme une droite, une parabole, etc. Il s'agit alors d'utiliser les procédures de **régression**. Les deux variables ne jouent pas le même rôle (**plan expérimental 2**). La variable X est appelée variable **indépendante** ou encore **contrôlée**, sous-entendu par l'expérimentateur. La variable Y est appelée variable **dépendante** ou encore **réponse**. Le modèle mathématique, s'il est bien ajusté aux observations, permet de comprendre le comportement du système et d'en prévoir l'évolution.

Remarque 3.1.1. Il est très important, avant l'expérimentation, de bien préciser les objectifs et ensuite choisir le plan expérimental approprié.

3.2. Le coefficient de corrélation linéaire

Considérons un couple de variables aléatoires (X, Y). Nous supposons que ce couple est distribué selon une loi **normale** bidimensionnelle. Cette loi dépend de 5 paramètres. Les 4 premiers, notés μ_X, μ_Y et σ_X, σ_Y sont, respectivement, les moyennes théoriques et les écart-types théoriques des variables aléatoires X et Y.

Remarque 3.2.1. Il ne suffit pas de tester la normalité de chacune des variables séparément, pour en déduire la normalité du couple. Celle-ci est très difficile à tester et la procédure qui permet de le faire ne sera pas décrite dans ces notes de cours. Dans toute la suite, la normalité d'un couple de variables aléatoires sera toujours admise.

Définition 3.2.1. *Le cinquième paramètre noté $\varrho = \varrho(X,Y)$ est appelé le* **coefficient de corrélation linéaire** *théorique des variables aléatoires X et Y. En fait ce coefficient se déduit d'un autre paramètre, appelé la* **covariance** *théorique des variables aléatoires X et Y, notée $\mathrm{Cov}(X,Y)$, qui est égale à la moyenne théorique du produit μ_{XY} moins le produit des moyennes théoriques μ_X et μ_Y. Nous avons :*

$$\varrho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \ \sigma_Y}. \tag{3.2.1}$$

Propriétés 3.2.1.
- *$\varrho(X,Y)$ est bien un coefficient, c'est-à-dire $\varrho(X,Y)$ ne dépend pas des unités des variables étudiées.*
- *$\varrho(X,Y)$ est compris entre -1 et 1. Si $\varrho(X,Y) > 0$ (respectivement $\varrho(X,Y) < 0$), les variables aléatoires X et Y varient dans le même sens (respectivement dans le sens contraire).*
- *Si $\varrho(X,Y) = -1$ ou 1, alors il existe deux nombres fixes a et b, tels que*

$$Y = a + bX.$$

- *Si $\varrho(X,Y) = 0$, les variables aléatoires X et Y sont dites non corrélées, et dans le cas* **gaussien** *les variables aléatoires X et Y sont* **indépendantes**.

Remarque 3.2.2. Dans le cas du couple de variables aléatoires (X,Y) suivant une loi normale bidimensionnelle, le coefficient $\varrho(X,Y)$ décrit la totalité de la liaison entre les deux variables aléatoires X et Y.

Nous proposons à présent d'estimer ce coefficient. Considérons un n-échantillon

$$\{(x_i, y_i), i = 1, \ldots, n\}.$$

Nous utilisons donc ici la notation la plus simple, celle du plan expérimental 1.

Nous estimons en premier lieu la covariance théorique notée $\mathrm{Cov}(X,Y)$ par la covariance observée notée $\mathrm{Cov}(x,y)$:

$$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}. \tag{3.2.2}$$

Définition 3.2.2. *Nous appelons* **coefficient de corrélation linéaire observé**, *le nombre défini par :*

$$r(x,y) = \frac{\mathrm{Cov}(x,y)}{s(x)\ s(y)} = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\displaystyle\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}}. \tag{3.2.3}$$

Ce nombre r est une réalisation de la variable $\varrho = \varrho(X,Y)$.

Interprétation 3.2.1. *Comme nous le verrons par la suite, nous interprétons $r^2(x,\ y)$ comme la part de la variation de la variable Y expliquée par (ou due à) une expression linéaire en la variable X, c'est-à-dire $a + bX$.*

Retour à l'exemple 3.1.1. Nous calculons le coefficient de corrélation linéaire observé $r(x,y)$. Nous avons :

$$\overline{x} = \frac{584}{11} = 53,091\ ans$$

et

$$\overline{y} = \frac{2285}{11} = 207,727\ cg/L.$$

Pour les variances non corrigées nous avons :

$$s^2(x) = \left(\frac{32834}{11}\right) - 53,091^2 = 166,264\ ans^2 \quad \text{et} \quad s(x) = \sqrt{166,264} = 12,894\ ans,$$

et

$$s^2(y) = \left(\frac{515355}{11}\right) - 207,727^2 = 3699,835\ cg^2/L^2 \quad \text{et} \quad s(y) = \sqrt{3699,835} = 60,826\ cg/L.$$

Enfin, la somme des produits est égale à :

$$\sum_{i=1}^{11} x_i y_i = 127235\ ans \cdot cg/L,$$

et la covariance observée à :

$$\text{Cov}(x,y) = \left(\frac{127235}{11}\right) - (53,091 \times 207,727) = 538,388\ ans \cdot cg/L.$$

Nous pouvons à présent calculer le coefficient de corrélation linéaire observé :

$$r(x,y) = \frac{538,388}{12,894 \times 60,826} = 0,686.$$

Comme $r^2(x,y) = 0,686^2 = 0,471$, nous interprétons la valeur du coefficient de corrélation linéaire observé en affirmant qu'environ 47% de la variation du taux de cholestérol peut être expliquée par une fonction linéaire de l'âge. La valeur $0,686$ correspond à une **estimation ponctuelle** de $\rho(X,Y)$, coefficient de corrélation linéaire inconnu entre l'âge (X) et le taux de cholestérol (Y) sur l'ensemble de la population étudiée.

Retour à l'exemple 3.1.2. Nous calculons le coefficient de corrélation linéaire observé $r(x,y)$. Nous avons :

$$\overline{x} = 6,500\ mn$$

et

$$\overline{y} = 281,375\ g/mL.$$

Pour les écart-types non corrigés, nous avons :

$$s(x) = 3,841\ mn$$

et

$$s(y) = 145,160\ g/mL.$$

De plus la covariance observée est égale à :

$$\text{Cov}(x,y) = \left(\frac{15631}{12}\right) - (6,500 \times 281,375) = -526,355\ mn \cdot g/mL.$$

Nous pouvons à présent calculer le coefficient de corrélation linéaire observé :

$$r(x,y) = \frac{-526,355}{3,841 \times 145,160} = -0,944$$

et

$$r^2(x,y) = 0,891.$$

Ainsi environ 89% de la variation de la concentration de l'hormone (Y) peut être représentée par une fonction linéaire du temps (X). Remarquons que le fait que la concentration est décroissante par rapport au temps, s'exprime par un coefficient de corrélation linéaire observé négatif.

Remarque 3.2.3. Toutes les méthodes statistiques ne permettent de mettre en évidence, le cas échéant, qu'une liaison entre les variables aléatoires X et Y et en aucun cas une relation de cause à effet. Cette dernière n'est qu'une interprétation donnée par l'expérimentateur.

3.3. Tests d'hypothèse

3.3.1. Test de l'hypothèse nulle $(\mathcal{H}_0) : \varrho(X, Y) = 0$

Nous avons vu préalablement qu'une corrélation linéaire nulle entre deux variables aléatoires est équivalente, dans le cas gaussien à l'indépendance de ces deux variables. C'est pourquoi nous présentons le test d'hypothèse bilatéral :

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

contre

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

La loi de la variable aléatoire $\varrho(X, Y)$, sous l'hypothèse nulle $(\mathcal{H}_0)$ est très difficile à calculer. Nous allons utiliser une transformation qui nous ramène à une loi connue.

Propriété 3.3.1. *Si l'hypothèse nulle $(\mathcal{H}_0)$ est vraie et dans le cadre de la normalité du couple (X, Y), la variable aléatoire définie par*

$$t = \sqrt{n-2}\frac{\varrho(X, Y)}{\sqrt{1 - \varrho^2(X, Y)}}$$

suit une loi de Student $\mathcal{T}_{n-2}$ à $n-2$ degrés de liberté.

Décision 3.3.1. *Pour un seuil fixé $\alpha(= 5\% = 0,05$ en général), les tables de la loi de Student, à $n-2$ degrés de liberté, nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[-c \leqslant \mathcal{T}_{n-2} \leqslant c] = 1 - \alpha$. Si nous utilisons un logiciel, par exemple le logiciel* **MINITAB***, celui-ci nous fournit une $P-$valeur. Alors nous décidons*

$$\begin{cases} (\mathcal{H}_0) \text{ est vraie} & \text{si } -c < t_{obs} < c, & \text{ou si } P > \alpha \\ (\mathcal{H}_1) \text{ est vraie} & \text{si } t_{obs} \leqslant -c \text{ ou } c \leqslant t_{obs}, & \text{ou si } P \leqslant \alpha. \end{cases}$$

Retour à l'exemple 3.1.1. Nous testons :

Hypothèses :

$(\mathcal{H}_0)$: Le taux de cholestérol est indépendant linéairement de l'âge

contre

$(\mathcal{H}_1)$: Le taux de cholestérol dépend linéairement de l'âge.

Nous supposons que le couple (Âge, Cholestérol) suit une loi normale bidimensionnelle et nous calculons la statistique de test :

$$t_{obs} = \sqrt{11-2}\frac{0,686}{\sqrt{1 - 0,686^2}} = 2,832.$$

Les tables de la loi de Student à 9 degrés de liberté nous fournissent, pour $\alpha = 5\% = 0,05$, la valeur critique $c = 2,262$. Nous décidons donc que l'hypothèse alternative $(\mathcal{H}_1)$, c'est-à-dire « le taux de cholestérol dépend linéairement de l'âge » est vraie. Le risque de cette décision est aussi le seuil du test, c'est-à-dire $\alpha = 5\% = 0,05$.

Retour à l'exemple 3.1.2. Dans cet exemple il ne faut pas effectuer le test précédent. En effet la variable X est totalement fixée par l'expérimentateur. La variable X n'est donc pas aléatoire, comme dans le plan expérimental 1. Ainsi nous n'avons pas de loi normale bidimensionnelle pour le couple (X, Y). Mais nous verrons par la suite comment tester l'indépendance dans ce cas.

3.3.2. Test de l'hypothèse nulle $(\mathcal{H}_0) : \varrho(X,Y) = \varrho_0(X,Y)$

Le cas d'un coefficient de corrélation linéaire non nul est plus délicat à traiter. Soit un nombre donné $\varrho_0 \in]-1,0[\cup]0,1[$. Nous considérons le test d'hypothèse suivant :

Hypothèses :

$$(\mathcal{H}_0) : \varrho(X,Y) = \varrho_0(X,Y)$$

contre

$$(\mathcal{H}_1) : \varrho(X,Y) \neq \varrho_0(X,Y).$$

Propriété 3.3.2. *Posons :*

$$Z = \frac{1}{2}\ln\left(\frac{1+\varrho(X,Y)}{1-\varrho(X,Y)}\right) = \operatorname{argtanh}(\varrho(X,Y))$$

et

$$z_0 = \operatorname{argtanh}(\varrho_0(X,Y)).$$

R. A. Fisher a montré que, si l'hypothèse nulle $(\mathcal{H}_0)$ est vraie et dans le cadre de la normalité, alors :

$$\lim_{n\to+\infty} \mathcal{L}\left(\sqrt{n-3}(Z-z_0)\right) = \mathcal{N}(0\ ;\ 1).$$

En pratique dès que la taille $n \geqslant 30$, nous pouvons utiliser cette loi asymptotique pour réaliser le test.

Décision 3.3.2. *Pour un seuil fixé $\alpha(= 5\% = 0,05$ en général), les tables de la loi normale centrée et réduite nous fournissent une valeur critique $c = 1,96$ telle que $\mathbb{P}_{(\mathcal{H}_0)}[-c \leqslant \mathcal{N}(0\ ;\ 1) \leqslant c] = 1-\alpha$. Alors nous décidons que*

$$\begin{cases} (\mathcal{H}_1) \text{ est vraie si} & \sqrt{n-3}(z-z_0) \leqslant -c \text{ ou } c \leqslant \sqrt{n-3}(z-z_0), \\ (\mathcal{H}_0) \text{ est vraie si} & -c < \sqrt{n-3}(z-z_0) < c. \end{cases}$$

Remarque 3.3.1. Mais ce test n'est pas utilisé fréquemment, compte tenu des difficultés d'interprétation d'une valeur autre que 0 pour le coefficient de corrélation.

Retour à l'exemple 3.1.1. Posons $\varrho_0 = 0,750$ et faisons le test suivant :

Hypothèses :

$(\mathcal{H}_0)$: Le taux de cholestérol et l'âge ont un coefficient de corrélation linéaire égal à 0,75

contre

$(\mathcal{H}_1)$: Le taux de cholestérol et l'âge ont un coefficient de corrélation linéaire différent de 0,75.

Des résultats numériques précédents nous obtenons :

$$z = \frac{1}{2}\ln\left(\frac{1+0,686}{1-0,686}\right) = 0,841$$

et

$$z_0 = \frac{1}{2}\ln\left(\frac{1+0,750}{1-0,750}\right) = 0,973.$$

D'où

$$\sqrt{11-3}(z-z_0) = -0,373.$$

Cette valeur est comprise entre $-1,96$ et $1,96$. Nous décidons que l'hypothèse nulle $(\mathcal{H}_0)$ « *Le taux de cholestérol et l'âge ont un coefficient de corrélation linéaire égal à 0,750* » est vraie. Notons que le risque associé à cette décision n'a pas été calculé. De plus, **cette décision n'est pas fondée dans la mesure où la taille $n = 11 < 30$. Nous avons effectué ce test uniquement pour présenter un exemple d'application**.

3.4. Intervalle de confiance de $\varrho(X,Y)$

La même approximation permet de construire un intervalle de confiance de $\varrho(X,Y)$ au seuil $\alpha = 5\% = 0,05$:

$$\left]\tanh\left(z - \frac{1,96}{\sqrt{n-3}}\right); \tanh\left(z + \frac{1,96}{\sqrt{n-3}}\right)\right[,$$

où
- z désigne la réalisation de Z sur l'échantillon,
- $\tanh(u) = \dfrac{e^u - e^{-u}}{e^u + e^{-u}}$.

Retour à l'exemple 3.1.1. Notons que $n = 11$ est très faible pour utiliser une loi asymptotique, mais nous construisons néanmoins un intervalle de confiance de $\varrho(X,Y)$ pour montrer le déroulement des calculs. Nous avons :

$$z = \frac{1}{2}\ln\left(\frac{1+0,686}{1-0,686}\right) = 0,841.$$

Les bornes de l'intervalle de confiance dépendent des quantités suivantes sur lesquelles il faudra prendre tanh après :

$$z_1 = 0,841 - \frac{1,96}{\sqrt{11-3}} = 0,841 - 0,693 = 0,148$$

et

$$z_2 = 0,841 + \frac{1,96}{\sqrt{11-3}} = 0,841 + 0,693 = 1,534.$$

Les formules correspondant à la fonction inverse de

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

nous donnent :

$$r_1 = \tanh(z_1) = \frac{e^{0,148} - e^{-0,148}}{e^{0,148} + e^{-0,148}} = 0,147$$

et

$$r_2 = \tanh(z_2) = \frac{e^{1,534} - e^{-1,534}}{e^{1,534} + e^{-1,534}} = 0,911.$$

En conclusion, nous affirmons que $[0,147; 0,911]$ est parmi les 95% environ d'intervalles que nous pouvons construire avec cette méthode qui contiennent la vraie valeur inconnue du coefficent de corrélation linéaire $\varrho(X,Y)$.

Remarque 3.4.1. Remarquons que l'intervalle de confiance est très large. Ceci est dû, entre autres, au fait que la taille de l'échantillon, $n = 11$, est très faible.

3.5. Le coefficient de détermination et le rapport de corrélation

Nous nous plaçons dans le cadre du **plan expérimental 2** défini à l'exemple 3.1.2. Nous supposons que la variable X est entièrement contrôlée par l'expérimentateur et nous la noterons donc x, dans la mesure où elle n'a pas un comportement aléatoire. Nous observons ainsi une variable aléatoire Y qui dépend d'une variable déterministe x. Nous fixons I valeurs de x, notées $\{x_1, \ldots, x_I\}$, et pour chacune d'elles nous observons J ou n_i valeurs de Y, notées $\{y_{ij}, j = 1, \ldots, J\}$. Nous obtenons ainsi

$$n = IJ$$

ou encore

$$n = \sum_{i=1}^{I} n_i$$

observations. Nous rappelons l'équation de l'Analyse de la Variance dans le cas équilibré :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{I} J(\overline{y}_i - \overline{y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y}_i)^2.$$

Nous avions abrégé cette écriture (cf. Chapitre 2, équation (2.2.3)) en :

$$SC_{Tot} = SC_F + SC_R,$$

avec les notations

$$SC_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2 \qquad \text{variation totale,}$$

$$SC_F = \sum_{i=1}^{I} J(\overline{y}_i - \overline{y})^2 \qquad \text{variation des moyennes,}$$

$$s^2(y \mid x = x_i) = \frac{1}{J}\sum_{j=1}(y_{ij} - \overline{y}_i)^2 \qquad \text{variance des } y \text{ pour } x = x_i,$$

$$SC_R = \sum_{i=1}^{I} Js^2(y \mid x = x_i) \qquad \text{variation résiduelle.}$$

Nous posons :

Définition 3.5.1. *Nous appelons* **coefficient de détermination** *théorique la variable* $\varrho^2(X, Y)$.

Définition 3.5.2. *Nous appelons* **coefficient de détermination observé** *le nombre* $r^2(x, y)$*, que le logiciel* **MINITAB** *note également* R^2.

Les propriétés d'un coefficient de corrélation linéaire montrent que le coefficient de détermination est compris entre 0 et 1. S'il est égal à 1, alors la variable Y dépend linéairement de x. S'il est égal à 0 alors les variables x et Y sont non corrélées linéairement. Un autre coefficient permettant l'étude de la dépendance de Y par rapport à x, avec ce plan expérimental, est le suivant :

Définition 3.5.3. *Nous appelons* **rapport de corrélation** *de* Y *en* x *la quantité :*

$$\eta^2(y \mid x) = \frac{\displaystyle\sum_{i=1}^{I} n_i(\overline{y}_i - \overline{y})^2}{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2} = \frac{SC_F}{SC_{Tot}}.$$

Remarques 3.5.1.

- Remarquons que ce coefficient mesure l'intensité de la liaison entre les variables Y et x ou encore c'est la part de la variation de la variable Y « expliquée » par la variation de la variable x (et pas par une fonction linéaire de la variable x).
- L'équation de l'Analyse de la Variance implique que :

$$0 \leqslant \eta^2(y \mid x) \leqslant 1.$$

- Si $\eta^2(y \mid x) = 0$ ou, ce qui est équivalent, $s^2(\overline{y} \mid x) = 0$, c'est-à-dire

$$\overline{y}_i = \overline{y}, \qquad \text{pour tout} \qquad i = 1, \ldots, I,$$

  alors pour tous les $x = x_i$, la moyenne $\overline{y}_i$ reste constante. Ceci veut dire que la variable y ne dépend pas, en moyenne, de x.
- Si $\eta^2(y \mid x) = 1$ ou, ce qui est équivalent, $\overline{s^2(y \mid x)} = 0$, c'est-à-dire

$$s^2(\overline{y}_i) = 0, \qquad \text{pour tout} \qquad i = 1, \ldots, I,$$

  alors pour chaque $x = x_i$, c'est la même valeur de Y, $\overline{y}_i$, qui est observée, et ceci pour tous les $i = 1, \ldots, I$. Dans ce cas la variable Y dépend entièrement de la variable x.

- Il très rare d'obtenir en pratique des valeurs 0 ou 1 pour ce coefficient, mais c'est la proximité à ces valeurs qui sera interprétée comme un indicateur de l'existence d'une dépendance fonctionnelle de la variable Y par rapport à la variable x. Nous avons également la relation suivante :

$$r^2(x,y) \leqslant \eta^2(y \mid x).$$

Interprétation 3.5.1. *Pour décrire le type de dépendance de la variable Y par rapport à la variable x, la démarche descriptive suivante est proposée :*

- *si $r^2(x,y)$ et $\eta^2(y \mid x)$ sont petits, inférieurs à 0,1 par exemple, alors il sera admis que la variable Y ne dépend pas de la variable x.*
- *si $r^2(x,y)$ et $\eta^2(y \mid x)$ sont tous les deux grands, supérieurs à 0,9 par exemple, alors il sera admis que la variable Y dépend fonctionnellement de la variable x et que cette fonction est une droite.*
- *si $r^2(x,y)$ est petit et $\eta^2(y \mid x)$ est moyen, alors il sera admis que la variable Y dépend partiellement de la variable x, mais cette liaison partielle n'a pas la forme d'une droite.*
- *si $r^2(x,y)$ et $\eta^2(y \mid x)$ sont moyens, alors il sera admis que la variable Y dépend partiellement de la variable x et cette liaison partielle peut être décrite par une droite.*
- *si $r^2(x,y)$ est moyen et $\eta^2(y \mid x)$ est grand, alors il sera admis que la variable Y dépend fonctionnellement de la variable x, mais que cette fonction ne peut être décrite que très approximativement par une droite.*

Rappelons cependant que, contrairement à r^2 qui peut toujours être interprété, un rapport de corrélation η^2 n'a de sens que si nous avons appliqué **le plan expérimental 2**.

Retour à l'exemple 3.1.1. Notons que nous ne pouvons pas calculer ici le rapport de corrélation $\eta^2(y \mid x)$. En effet nous avons appliqué le **plan expérimental 1** et pour chaque valeur x_i nous avons observé qu'une seule valeur y_i.

Retour à l'exemple 3.1.2. Le rapport de corrélation $\eta^2(y \mid x)$ peut être calculé dans cet exemple. Nous avons le **plan expérimental 2** et pour chaque valeurs x_i plusieurs valeurs de y. Pour le calculer, nous utilisons le tableau de l'analyse de la variance sur les y_{ij}, en considérant que la variable x est qualitative :

Variation	SC	ddl	s^2	F_{obs}
Due au Facteur	242621,73	3	80873,91	63,21
Résiduelle	10235,83	8	1279,48	
Totale	252857,56	11		

Nous avons :

$$\eta^2(y \mid x) = \frac{242621,73}{252857,56} = 0,960.$$

Appliquons les recommandations formulées au paragraphe 3.5.1 :

– $\eta^2(y \mid x) = 0,960 > 0,900$, la concentration de l'hormone de croissance dépend fonctionnellement du temps écoulé après l'injection.

– $r^2(x,y) = 0,891 < 0,900$ nous ne pouvons approcher cette fonction que très approximativement par une droite..

3.6. La régression linéaire simple

La régression linéaire simple est une méthode statistique permettant d'étudier une dépendance linéaire d'une variable (aléatoire) quantitative Y par rapport à une autre variable x, contrôlée par l'expérimentateur. Nous supposons avoir n couples :

$$(x_1, y_1), \ldots, (x_n, y_n)$$

correspondant à n valeurs observées des variables :

$$(x_1, Y_1), \ldots, (x_n, Y_n),$$

c'est-à-dire nous avons **le plan expérimental 1**.

Définition 3.6.1. *Par* **régression linéaire simple**, *nous entendons le modèle suivant :*

$$Y_i = a + bx_i + \mathcal{E}_i, \qquad (i = 1, \ldots, n),$$

où la variable Y est appelée **variable à expliquer** *(ou encore* **variable dépendante***) et la variable x est appelée* **variable explicative** *(ou encore* **la variable indépendante***). Les nombres a et b sont deux paramètres fixes mais inconnus. Les $\mathcal{E}_i$ sont les termes d'erreurs qui, ici dans toute la suite, sont supposés être des variables aléatoires suivant des lois normales centrées de même variance inconnue σ^2.*

Remarque 3.6.1. Nous avons la même définition avec le plan expérimental 2 :

$$Y_{ij} = a + bx_i + \mathcal{E}_{ij}, \qquad (j = 1, \ldots, J \text{ (ou } n_i) \ ; \ i = 1, \ldots, I).$$

Nous supposons toujours qu'il existe au moins deux valeurs x_i qui sont distinctes, que les variables aléatoires Y_i sont indépendantes, mais n'ont pas toutes la même loi.

Remarque 3.6.2. Considérons le modèle d'analyse de la variance à un facteur :

$$Y_{ij} = \mu_i + \mathcal{E}_{ij}, \qquad (j = 1, \ldots, J \text{ (ou } n_i) \ ; \ i = 1, \ldots, I),$$

où les μ_i sont des paramètres fixes et inconnus, et $\mathcal{E}_{ij}$ des variables aléatoires centrées. La régression linéaire simple stipule simplement que les μ_i dépendent des valeurs connues x_i et d'une manière linéaire des paramètres inconnus a et b :

$$\mu_i = a + bx_i \qquad (i = 1, \ldots, I).$$

3.7. La méthode des moindres carrés ordinaire

Nous estimons les deux paramètres inconnus a et b.

Définition 3.7.1. *Nous appelons méthode des moindres carrés ordinaire appliquée à l'ensemble de points $\{(x_i,\ y_i), i = 1, \ldots, n\}$, la méthode qui consiste en la recherche de nombres $\widehat{a}$ et $\widehat{b}$ satisfaisant la relation :*

$$\sum_{i=1}^{n} (y_i - \widehat{a} - \widehat{b}x_i)^2 = \min_{a,b \in \mathbb{R}} \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Remarque 3.7.1. Nous cherchons donc la droite en x qui passe globalement le plus près possible des y.

Définition 3.7.2. *La droite $y = \widehat{a} + \widehat{b}x$ ainsi obtenue s'appelle* **la droite de régression**. *Dans le cadre du plan expérimental 2, la courbe définie par des segments de droite passant successivement par les points $\{(x_i, \overline{y}_i), i = 1, \ldots, n_i\}$ est appelée* **courbe de régression**.

Le calcul de ces estimations est relativement simple. En annulant les dérivées partielles de

$$Q(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

par rapport à a et b, nous obtenons un système d'équations linéaires en les inconnues a et b dont la solution est le couple suivant :

$$\widehat{b} = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\mathrm{Cov}(x,\ y)}{s^2(x)} = r(x,\ y)\frac{s(y)}{s(x)},$$

et

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x}.$$

Définition 3.7.3. *Dans toute la suite nous notons, pour $i = 1, \ldots, n$:*

$$\widehat{y}_i = \widehat{a} + \widehat{b}x_i$$

et

$$\widehat{e}_i = y_i - \widehat{y}_i.$$

Les premières valeurs sont des estimations de Y données par le modèle et les secondes sont des estimations des erreurs, appelées **résidus***, que nous pouvons en déduire.*

Remarque 3.7.2. Nous avons deux propriétés de la droite de régression linéaire :
- la somme des résidus est nulle,

$$\sum_{i=1}^{n} \widehat{e}_i = 0 \ ;$$

- la droite de régression passe par le point $(\overline{x}, \overline{y})$.

Remarque 3.7.3. Nous estimons l'équation de la droite de régression. Les coefficients estimés de la droite de régression linéaire simple sont donnés par :

$$\widehat{b} = 0,686 \times \frac{60,827}{12,894} = 3,236 \ cg/(L \cdot ans)$$

et

$$\widehat{a} = 207,727 - 3,236 \times 53,091 = 35,925 \ cg/L.$$

Ainsi, si le modèle de la régression linéaire simple peut être validé, nous l'estimons par l'expression :

$$\text{Cholestérol} = 35,925 + 3,236 \times \text{Âge}.$$

Remarque 3.7.4. Les coefficients estimés de la droite de régression linéaire simple sont donnés par :

$$\widehat{b} = -0,945 \times \frac{145,160}{3,841} = -35,721 \ g/(mL \cdot mn),$$

et

$$\widehat{a} = 281,375 + 35,721 \times 6,500 = 513,561 \ g/mL.$$

Ainsi, si le modèle de la régression linéaire simple peut être validé, nous l'estimons par l'expression :

$$\text{Concentration} = 513,561 - 35,721 \times \text{Minutes}.$$

3.8. La validation du modèle

Nous nous proposons de valider le modèle linéaire simple. Comme nous le constaterons dans la suite, ceci ne peut se faire que dans le cadre du plan expérimental 2 :

$$\{(x_i, \ y_{ij}), j = 1, \ldots, J \text{ (ou } n_i) \ ; \ i = 1, \ldots, I\},$$

où les y_{ij} sont des réalisations indépendantes de lois, $\mathcal{N}(\mu_i \ ; \ \sigma^2)$. Nous nous proposons de tester :

$$(\mathcal{H}_0) \ : \ \forall i = 1, \ldots, I, \quad \mu_i = a + bx_i$$

contre

$$(\mathcal{H}_1) \ : \ \text{au moins un des } \mu_i \neq a + bx_i.$$

Pour réaliser ce test nous adoptons une démarche analogue à celle de l'analyse de la variance. Nous considérons les sommes de carrés suivantes :

$$SC_{M|RG} = \sum_i (\overline{y}_i - \widehat{a} - \widehat{b}x_i)^2$$

et

$$SC_R = \sum (y_{ij} - \overline{y}_i)^2.$$

Dans la seconde nous reconnaissons la somme **résiduelle** des carrés qui, en dehors de toute considération de modèle linéaire simple, peut être utilisée pour estimer la variance inconnue σ^2 si on la divise par le nombre de degrés de liberté associé à cette somme de carrés, à savoir $n - I$. Par conséquent, nous pouvons établir la définition suivante :

Proposition 3.8.1. *Une estimation de la variance inconnue σ^2 est le carré moyen résiduel noté s_R^2.*

La première, que nous nommons sommes des carrés **du modèle autour de la régression**, sera utilisée pour mesurer l'écart du modèle linéaire simple à la courbe de régression. Un calcul relativement fastidieux permet de montrer que la somme des ces deux sommes donne :

$$SC_{R|M} = \sum (y_{ij} - \widehat{a} - \widehat{b}x_i)^2,$$

que nous interpréterons comme la variation résiduelle autour du modèle linéaire simple. La théorie statistique nous permet, comme pour l'analyse de la variance, de construire le tableau suivant :

Source de variation	Somme des carrés SC	ddl	Carrés moyens s^2	Fisher F_{obs}
Du modèle autour de la rég.	$SC_{M\|RG} = \sum(\overline{y}_i - \widehat{a} - \widehat{b}x_i)^2$ $= (\eta^2(y \mid x) - r^2(x,y))ns^2(y)$	$I-2$	$s^2_{M\|RG} = \dfrac{SC_{M\|RG}}{I-2}$	
Résiduelle	$SC_R = \sum(y_{ij} - \overline{y}_i)^2$ $= (1 - \eta^2(y \mid x))ns^2(y)$	$n-I$	$s^2_R = \dfrac{SC_R}{n-I}$	$F_{M,obs} = \dfrac{s^2_{M\|RG}}{s^2_R}$
Résiduelle autour du mod.	$SC_{R\|M} = \sum(y_{ij} - \widehat{a} - \widehat{b}x_i)^2$ $= (1 - r^2(x,y))ns^2(y)$	$n-2$		

Décision 3.8.1. *Pour un seuil donné $\alpha(= 0,05 = 5\%$ en général), les tables de Fisher nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[\mathcal{F}_{I-2,n-I} \leqslant c] = 1 - \alpha$. Alors nous décidons*

$$\begin{cases} (\mathcal{H}_1) \text{ est vraie si} & F_{M,obs} \geqslant c, \\ (\mathcal{H}_0) \text{ est vraie si} & F_{M,obs} < c. \end{cases}$$

Remarque 3.8.1. Remarquons que ce test ne peut pas être mis en œuvre si nous n'avons qu'une seule valeur y pour chaque x_i, c'est-à-dire si nous sommes dans le cas du plan expérimental 1. En effet dans ce cas $SC_R = 0$. En fait, les logiciels présentent un tableau un peu plus complexe qui est le suivant :

Tableau de l'analyse de la régression linéaire simple

Source de variation	Somme des carrés SC	ddl	Carrés moyens s^2	Fisher F_{obs}
De la régression linéaire (modèle)	$SC_M = \sum(\widehat{a} + \widehat{b}x_i - \overline{y})^2$ $= r^2(x,y)ns^2(y)$	$2-1$	$s^2_M = \dfrac{SC_M}{2-1}$	
Du modèle autour de la rég.	$SC_{M\|RG} = \sum(\overline{y}_i - \widehat{a} - \widehat{b}x_i)^2$ $= (\eta^2(y \mid x) - r^2(x,y))ns^2(y)$	$I-2$	$s^2_{M\|RG} = \dfrac{SC_{M\|RG}}{I-2}$	
Résiduelle	$SC_R = \sum(y_{ij} - \overline{y}_i)^2$ $(1 - \eta^2(y \mid x))ns^2(y)$	$n-I$	$s^2_R = \dfrac{SC_R}{n-I}$	$F_{M,obs} = \dfrac{s^2_{M\|RG}}{s^2_R}$
Résiduelle autour du mod.	$SC_{R\|M} = \sum(y_{ij} - \widehat{a} - \widehat{b}x_i)^2$ $(1 - r^2(x,y))ns^2(y)$	$n-2$	$s^2_{R\|M} = \dfrac{SC_{R\|M}}{n-2}$	$F_{R\|M,obs} = \dfrac{s^2_M}{s^2_{R\|M}}$
Totale	$SC_T = \sum(y_{ij} - \overline{y})^2$ $= ns^2(y)$	$n-1$		

Retour à l'Exemple 3.1.2. Nous avons déjà calculé :

$$\eta^2(y|x) = 0,960$$

et

$$r^2(x,\ y) = 0,891.$$

Comme $n = 12$, nous en déduisons le tableau suivant :

Variation	SC	ddl	s^2	F_{obs}
Du modèle autour de la régression	17226,000	4-2= 2	8613,000	
Résiduelle	10236,000	12-4= 8	1279,500	6,732

Les tables de Fisher nous fournissent, pour 2 et 8 degrés de liberté, la valeur critique $c = 4,46$. Nous en déduisons que : « l'hypothèse alternative ($\mathcal{H}_1$) est vraie », c'est-à-dire, nous n'acceptons pas le modèle de la droite de régression. Donc la concentration ne varie pas comme une droite du temps. Que pouvons-nous faire alors ?

Si nous effectuons le changement de variable $u_{ij} = \ln(y_{ij})$, nous obtenons :

$$\overline{u} = 5,494,$$

$$s(u) = 0,554,$$

et

$$\frac{1}{12}\sum_{ij} x_i u_{ij} = 33,623.$$

Comme nous avons déjà calculé la moyenne de x

$$\overline{x} = 6,500$$

et l'écart-type non corrigé de x

$$s(x) = 3,841,$$

nous en déduisons la covariance observée :

$$\mathrm{Cov}(x,u) = 33,623 - 6,500 \times 5,494 = -2,088.$$

Par conséquent, nous pouvons calculer le coefficient de corrélation linéaire observé :

$$r(x,u) = \frac{-2,088}{3,841 \times 0,554} = -0,981$$

ainsi que le carré de ce coefficient :

$$r^2(x,u) = 0,963.$$

Nous pouvons à présent calculer les estimations des coefficients a et b de la droite de régression, et donner l'équation de la droite de régression linéaire simple :

$$\ln(\text{Concentration}) = 6,412 - 0,141 \times \text{Minutes}.$$

De plus une **analyse de la variance** sur les u_{ij}, sans tenir compte du fait que la variable x est quantitative, nous donne :

Tableau de l'analyse de variance (réalisé à l'aide d'une ANOVA)

Variation	SC	ddl	s^2	F_{obs}
Due au facteur	3,5703	3	1,1901	84,4043
Résiduelle	0,1125	8	0,0141	
Totale	3,6827	11		

Remarque 3.8.2. Si vous faites les calculs du tableau de l'analyse de variance (ci-dessus) avec le logiciel **MINITAB**, vous trouverez comme $F_{obs} = 84,67$ et vous lisez comme valeur pour F_{obs}, dans le tableau, une valeur égale à $84,4043 = \frac{1,1901}{0,0141}$. Ceci est dû aux erreurs d'approximation dans le tableau de l'analyse de variance que nous avons recopié à partir du logiciel **MINITAB**.

Ce tableau nous permet de calculer :

$$\eta^2(u|x) = \frac{3,570}{3,682} = 0,970.$$

Rappelons que nous avions obtenu :

$$\eta^2(y|x) = 0,960$$

et

$$r^2(x,y) = 0,891.$$

Nous remarquons que $\eta^2(u|x) = \eta^2(\ln(Y)|x)$ n'est pas égal à $\eta^2(y|x)$ alors que l'on aurait $\eta^2(y|\ln(x)) = \eta^2(y|x)$ dès que le rapport $\eta^2(y|\ln(x))$ est défini.

Il semblerait donc que le modèle $\ln(y) = a + b\ x$ décrive mieux le phénomène. Nous allons tester la validité de cet ajustement. Nous posons :

$$(\mathcal{H}_0)\ :\ \mu(U_{ij}) = a + b\ x_i$$

contre

$$(\mathcal{H}_1)\ :\ \text{au moins un des } \mu(U_{ij}) \neq a + b\ x_i.$$

Les valeurs de $\eta^2(u|x)$ et de $r^2(x,u)$ nous permettent de construire le tableau suivant :

Variation	SC	ddl	s^2	F_{obs}
Du modèle autour de la régression	0,0363	2	0,0181	1,2837
Résiduelle	0,1125	8	0,0141	

La valeur $F_{M,obs} = 1,2837$ comparée à la même valeur critique $c = 4,46$ nous permet de décider que l'hypothèse nulle $(\mathcal{H}_0)$ est vraie. Le logarithme de la concentration s'exprime comme une fonction linéaire du temps en minutes.

Le deuxième test qui est présenté dans le tableau grand tableau précédent, tableau de l'analyse de la régression, n'est fondé que si, sous les conditions d'indépendance, de normalité et d'homogénéité, le modèle de la droite de régression a été accepté. Nous allons l'étudier. Remarquons qu'au numérateur intervient :

$$SC_M = \sum(\widehat{a} + \widehat{b}x_i - \overline{y})^2.$$

Ce terme mesure visiblement l'écart du modèle (estimé) à la moyenne générale des y ; une autre manière de le dire est l'écart de la droite du modèle (estimé) à la position horizontale. Donc le test permettra de répondre à la question : « la droite est-elle horizontale ? »ou, ce qui revient au même (si le modèle est validé), « Y dépend-il de x ? ». Au dénominateur intervient :

$$SC_{R|MD} = \sum(y_{ij} - \widehat{a} - \widehat{b}x_i)^2.$$

Ce terme, divisé par ses degrés de liberté, est dans ce cas, toujours sous les conditions précédentes et aprés validation du modèle, la meilleure estimation de la variance résiduelle, c'est-à-dire la meilleure estimation de σ^2.

Ainsi il nous apparaît clairement que le test concerne les hypothèses :

$$(\mathcal{H}_0) \ : \ Y \text{ ne dépend pas de } x$$

contre

$$(\mathcal{H}_1) \ : \ Y \text{dépend de } x.$$

Ceci peut se dire :

$$(\mathcal{H}_0) \ : \ b = 0$$

contre

$$(\mathcal{H}_1) \ : \ b \neq 0.$$

La statistique de test s'écrit :

$$F_{R|M} = \frac{s^2_M}{s^2_{R|M}} = (n-2)\frac{r^2(x,\ y)}{1 - r^2(x,\ y)}.$$

C'est exactement le carré de la statistique que nous avons utilisé pour tester :

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

contre

$$(\mathcal{H}_1) \ : \ \varrho(X, Y) \neq 0,$$

et le carré de la valeur critique bilatérale pour une loi de Student est la même que celle d'une loi de Fisher avec au numérateur 1 degré de liberté et au dénominateur le même que celui de la loi de Student.

3.9. Vérification des conditions

3.9.1. La normalité

Nous considérons les résidus de l'analyse de régression (cf. Définition 3.7.3.) :

$$\widehat{e_i} = y_i - \widehat{y_i} = y_i - \widehat{a} - \widehat{b}x_i, \quad i = 1, \ldots, n,$$

qui sont des estimations des termes d'erreurs. En général les logiciels les donnent comme un des résultats des calculs de l'analyse de la régression. Nous savons que l'une des conditions de l'application de cette dernière est la normalité des erreurs. Nous pouvons la tester à l'aide du test de Shapiro-Francia sur les résidus. Il est à noter que les logiciels donnent aussi des transformés des résidus, dont l'étude, plus précise, dépasse le cadre de ce cours.

3.9.2. Étude graphique des résidus.

Le graphique des points $\{(i, \widehat{e_i}), i = 1, \ldots, n\}$ est en général tracé. Il est très informatif quant à l'adéquation du modèle. En effet les points doivent être répartis “au hasard” de part et d'autre de 0 ; dés qu'une régularité ou une tendance apparaît, il faut en déduire que vraisemblablement le modèle de la droite ne convient pas.

3.9.3. L'homogénéité.

Elle ne peut, bien évidemment, être testée que dans le cas d'observations avec répétitions, c'est-à-dire dans le cas du plan expérimental 2. Nous utilisons alors le test de Bartlett, comme dans une analyse de la variance à un facteur (cf chapitre 2, 2.3.3.).

Retour à l'Exemple 3.1.1. Les résultats des calculs nous donnent une statistique de Shapiro-Francia $R = 0,9781$, ce qui nous permet de décider que la condition de normalité est satisfaite. Par contre comme nous sommes dans le cas d'un plan expérimental 1, nous ne pouvons pas tester l'égalité des variances.

Retour à l'Exemple 3.1.2. Les calculs sur les données $\{x_i,\ u_{ij}\}$, nous donnent une statistique de Shapiro-Francia $R = 0,9831$, ce qui nous permet de décider que la condition de normalité est satisfaite. Le test de Bartlett sur l'égalité des variances, qui est possible dans ce cas nous une statistique $B = 1,039$ et une $P-$valeur égale à 0,792 ; ceci nous permet de décider que les variances sont égales.

3.10. Étude des paramètres a et b

Dans tout ce qui suit, nous supposons que le modèle suivant est validé, c'est-à-dire que les conditions suivantes sont satisfaites :

$$Y_i = a + bx_i + \mathcal{E}_i, \qquad i = 1, \ldots, n,$$

où les $\mathcal{E}_i$ sont n variables indépendantes de même loi $\mathcal{N}(0\ ;\ \sigma^2)$. Nous utilisons l'écriture du plan expérimental 1 pour simplifier.

Propriétés 3.10.1. *La meilleure estimation de σ^2 est donné par :*

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{n}{n-2}s^2(y)\left(1 - r^2(x\ ,\ y)\right) = s^2_{R|M}.$$

Nous pouvons alors déduire des lois utilisables en pratique.

Propriétés 3.10.2. *Nous avons les résultats suivants :*

$$\begin{aligned}
1)\quad \mathcal{L}\left(\frac{\widehat{A} - a}{S_A}\right) &= \mathcal{T}_{n-2},\\
2)\quad \mathcal{L}\left(\frac{\widehat{B} - b}{S_B}\right) &= \mathcal{T}_{n-2},\\
3)\quad \mathcal{L}\left(\frac{\widehat{Y}(x) - a - bx}{S_{\widehat{Y}(x)}}\right) &= \mathcal{T}_{n-2},
\end{aligned}$$

avec

$$S_B^2 = \frac{S^2_{R|M}}{ns^2(x)}, \qquad S_A^2 = \frac{S^2_{R|M}}{n}\left(1 + \frac{\overline{x}^2}{s^2(x)}\right), \qquad S^2_{\widehat{Y}(x)} = \frac{S^2_{R|M}}{n}\left(1 + \frac{(x-\overline{x})^2}{s^2(x)}\right).$$

3.10.1. Intervalles de confiance

Propriétés 3.10.3. *Désignons par c_{n-2} la valeur critique bilatérale pour une loi de Student $\mathcal{T}_{n-2}$ à $n-2$ degrés de liberté pour un seuil fixé α. Alors un intervalle de confiance pour le paramètre a au seuil α est donné par :*

$$\left[\widehat{a} - c_{n-2}s_A\ ;\ \widehat{a} + c_{n-2}s_A\right].$$

Un intervalle de confiance pour le paramètre b au seuil α est donné par :

$$\left[\widehat{b} - c_{n-2}s_B\ ;\ \widehat{b} + c_{n-2}s_B\right].$$

Soit x fixé. Alors un intervalle de confiance pour $y(x) = a + bx$ au seuil α est donné par :

$$\left[\widehat{y}(x) - c_{n-2}s_{\widehat{y}(x)};\ \widehat{y}(x) + c_{n-2}s_{\widehat{y}(x)}\right].$$

Nous avons posé :

$$s_B^2 = \frac{s_{R|M}^2}{ns^2(x)}, \quad s_A^2 = \frac{s_{R|M}^2}{n}\left(1 + \frac{\overline{x}^2}{s^2(x)}\right), \quad s_{\widehat{Y}(x)}^2 = \frac{s_{R|M}^2}{n}\left(1 + \frac{(x-\overline{x})^2}{s^2(x)}\right).$$

Remarque 3.10.1. Notons que pour obtenir des intervalles de faible amplitude, nous devons avoir une taille d'échantillon n élevée et une variance $s^2(x)$ la plus grande possible. Comme c'est l'expérimentateur qui fixe les x_i, c'est à lui de faire en sorte que ces conditions soient réalisées.

Retour à l'Exemple 3.1.1. Nous supposons avoir validé le modèle linéaire simple pour le cholestérol en fonction de l'âge. Nous avons déjà calculé $\overline{x} = 53,091$, $s^2(x) = 166,255$ et $s_{R|M}^2 = 2392,22$ (pour ce dernier nombre nous utilisons le fait que $r^2(x,y) = 0,471$ et que $s^2(y) = 3699,948$). Nous avons ainsi :

$$s_B^2 = \frac{2392,22}{11 \times 166,255} = 1,308$$

et

$$s_B = 1,143 \ ;$$

$$s_A^2 = \frac{2392,22}{11}\left(1 + \frac{53,091^2}{166,255}\right) = 3904,495$$

et

$$s_A = 62,486.$$

La valeur critique bilatérale pour une loi de Student $\mathcal{T}_9$ est $c_9 = 2,2622$. Nous en déduisons les intervalles de confiance au seuil de 5 %, pour b :

$$[3,236 - 2,2622 \times 1,143 \ ; \ 3,236 + 2,2622 \times 1,143] = [0,650 \ ; \ 5,822] \ ;$$

et pour a :

$$[35,925 - 2,2622 \times 62,486 \ ; \ 35,925 + 2,2622 \times 62,486] = [-105,431 \ ; \ 177,281] \ ;$$

Nous constatons que les intervalles sont relativement larges ; ceci est dû à la faible taille de l'échantillon.

Retour à l'Exemple 3.1.2. Nous souhaitons estimer la concentration en hormone de croissance 10 minutes après l'injection. Nous avons validé le modèle linéaire $\ln(y) = a + bx$. Les calculs du §2 nous donnent une estimation ponctuelle de $y(10)$ par :

$$\ln(\widehat{y(10)}) = \widehat{a} + \widehat{b} \times 10 = 6,412 - 0,141 \times 10 = 5,002 \quad \text{et} \quad \widehat{y(10)} = e^{5,002} = 148,710.$$

Pour construire un intervalle de confiance, nous avons $s_{R|M}^2 = 0,0149, \overline{x} = 6,50$ et $s^2(x) = 14,7456$. La propriété 3.10.2. nous donne :

$$s_{\widehat{u(10)}}^2 = \frac{0,0149}{12}\left(1 + \frac{(10-6,50)^2}{14,7456}\right) = 0,002273 \quad \text{et} \quad s_{\widehat{u(10)}} = 0,0477.$$

La valeur critique bilatérale d'une loi de Student au seuil de $5\% = 0,05$ à 10 degrés de liberté est $c_{10} = 2,2281$. La proposition 5.3. nous donne alors un intervalle de confiance de $u(10)$:

$$[5,002 - 2,2281 \times 0,0477 \ ; \ 5,002 + 20,2281 \times 0,0477] = [4,8930 \ ; \ 5,1056],$$

Pour $y(10)$ nous en déduisons :

$$\left[e^{4,8927} \ ; \ e^{5,1053}\right] = [133,3530 \ ; \ 164,9430].$$

La faiblesse de la variance résiduelle par rapport au modèle, nous fournit un intervalle relativement étroit.

3.10.2. Tests d'hypothèses .

Nous testons les hypothèses :

$$(\mathcal{H}_0) \ : \ a = a_0$$

contre

$$(\mathcal{H}_1) : \ a \neq a_0.$$

Statistique : Nous considérons la statistique suivante

$$t = \frac{\widehat{a} - a_0}{s_A}$$

où s_A est donné par la propriété 3.10.3.

Propriétés 3.10.4. *Le nombre t est la réalisation d'une variable T dont la loi, lorsque l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, est une loi de Student à $n-2$ degrés de liberté.*

Décision 3.10.1. *Pour un seuil α (=5%=0,05 en général), les tables du $\mathcal{T}$ à $n-2$ degrés de liberté nous fournissent une valeur critique bilatérale c telle que $\mathbb{P}_{(\mathcal{H}_0)}[-c < \mathcal{T}_{n-2} < c] = 1-\alpha$. Alors nous décidons :*

$$\begin{cases} (\mathcal{H}_1) \text{ est vraie si} & t \leqslant -c \text{ ou si } t \geqslant c, \\ (\mathcal{H}_0) \text{ est vraie si} & -c < t < c. \end{cases}$$

Nous testons les hypothèses :

$$(\mathcal{H}_0) : \ b = b_0$$

contre

$$(\mathcal{H}_1) : \ b \neq b_0.$$

Statistique : Nous considérons la statistique suivante

$$t = \frac{\widehat{b} - b_0}{s_B}$$

où s_B est donné par la propriété 3.10.3.

Propriétés 3.10.5. *Le nombre t est la réalisation d'une variable T dont la loi, lorsque l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, est une loi de Student à $n-2$ degrés de liberté.*

Décision 3.10.2. *Pour un seuil α (=5%=0,05 en général), les tables de la loi de Student, à $n-2$ degrés de liberté, nous fournissent une valeur critique bilatérale c telle que $\mathbb{P}_{(\mathcal{H}_0)}[-c < \mathcal{T}_{n-2} < c] = 1-\alpha$. Alors nous décidons :*

$$\begin{cases} (\mathcal{H}_1) \text{ est vraie si} & t \leqslant -c \text{ ou si } t \geqslant c, \\ (\mathcal{H}_0) \text{ est vraie si} & -c < t < c. \end{cases}$$

Remarque 3.10.2. Les logiciels donnent en général la $P-$valeur de ces deux tests pour $a_0 = 0$ et pour $b_0 = 0$.