# Cours de Statistique pour Licence troisième année de Biologie

Version originale rédigée par Photis Nobelis, modifiée par Myriam Maumy

# Table des matières

# Chapitre 1

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

## 1.1. Modèle

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 1.1.1.** 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.

| Eponges | Durées | 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 1.1.1.** 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 1.1.2.** 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 de durées.

**Remarque 1.1.3.** 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)**. Il est à **effets fixes (fixed effects)**.
2. l'hydrophilie qui est une mesure. Elle est parfois appelée **la réponse (response)**.

**Notations 1.1.1.** *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 :

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

Lorsque *les échantillons sont de même taille* $J$, nous disons que l'expérience est **équilibrée (balanced)**. C'est le cas dans l'**Exemple 1.1.1.** 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 :

$$\overline{y}_i = \frac{1}{J} \sum_{j=1}^{J} y_{ij}, \quad i = 1, \ldots, I, \tag{1.1.1}$$

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

$$s_i^2(y) = \frac{1}{J} \sum_{j=1}^{J} (y_{ij} - \overline{y}_i)^2, \quad i = 1, \ldots, I. \tag{1.1.2}$$

**Remarque 1.1.4.** 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 1.1.1. :** Après calculs, nous avons :

$$\overline{y}_1 = 41,33 \quad \overline{y}_2 = 38,33 \quad \overline{y}_3 = 28,33 \quad \overline{y}_4 = 31,$$

et

$$s_1^2(y) = 1,56 \quad s_2^2(y) = 2,89 \quad s_3^2(y) = 13,56 \quad s_4^2(y) = 2.$$

Le nombre total d'observations est égal à :

$$n = IJ = 12.$$

**Conditions 1.1.1.** *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** $\sigma^2$ *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.

**Test de comparaison 1.1.1.** Nous nous proposons de tester :

$$(H_0) \ : \ \mu_1 = \mu_2 = \ldots = \mu_I$$

contre

$$(H_1) \ : \ \text{Les } \mu_i \ \text{ ne sont pas tous égaux.}$$

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.

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

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

**Propriété 1.2.1.** *La moyenne de toutes les observations est la moyenne des moyennes de chaque échantillon. Ceci s'écrit :*

$$\overline{y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{I} y_{ij} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} = \frac{1}{I} \sum_{i=1}^{I} \overline{y}_i. \tag{1.2.1}$$

**Retour à l'Exemple 1.1.1.** Pour cet exemple, nous constatons cette propriété. En effet, nous avons :

$$\overline{y} = \frac{1}{12} \times 417 = \frac{1}{4}(41,33 + 38,33 + 28,33 + 31) = \frac{1}{4} \times 139 = 34,75,$$

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

**Propriété 1.2.2.** *La variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances. Ceci s'écrit :*

$$s^2(y) = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \overline{y})^2 = \frac{1}{I} \sum_{i=1}^{I} (\overline{y}_i - \overline{y})^2 + \frac{1}{I} \sum_{i=1}^{I} s_i^2(y). \tag{1.2.2}$$

**Retour à l'Exemple 1.1.1.** Pour cet exemple, un calcul simple nous donne :

$$s^2(y) = 32,85.$$

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

$$\frac{1}{I} \sum_{i=1}^{I} (\overline{y}_i - \overline{y})^2 = \frac{1}{4} \left( (41,33 - 34,75)^2 + (38,33 - 34,75)^2 + (28,33 - 34,75)^2 + (31 - 34,75)^2 \right) = 27,85,$$

que la moyenne des variances est égale à :

$$\frac{1}{I}\sum_{i=1}^{I} s_i^2(y) = \frac{1}{4}(1,56 + 2,89 + 13,56 + 2) = 5.$$

En faisant la somme des deux derniers résultats, nous retrouvons bien la valeur de $32,85$ que nous avons obtenue par le calcul simple. Donc la relation (1.2.2) est bien vérifiée.

**Remarque 1.2.1.** En multipliant les deux membres par $n$ de l'équation (1.2.2), 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 :

$$SC_{Tot} = SC_F + SC_R. \tag{1.2.3}$$

**Retour à l'Exemple 1.1.1** Dans cet exemple, nous avons d'une part

$$SC_{Tot} = 394,25$$

et d'autre part

$$SC_F = 334,25 \quad \text{et} \quad SC_R = 60.$$

Donc lorsque nous faisons la somme des deux derniers résulats nous retrouvons bien la valeur du premier résultat. Donc la relation (1.2.3) est bien vérifiée.

**Définition 1.2.1.** *Nous appelons* **variation totale (total variation)** *le terme :*

$$SC_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2. \tag{1.2.4}$$

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

**Définition 1.2.2.** *Nous appelons* **variation due au facteur (variation between)** *le terme :*

$$SC_F = J\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2. \tag{1.2.5}$$

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

**Définition 1.2.3.** *Nous appelons* **variation résiduelle (variation within)** *le terme :*

$$SC_R = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij} - \overline{y}_i)^2\right). \tag{1.2.6}$$

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

**Principe du test :** Si l'hypothèse nulle ($H_0$) est vraie alors la quantité $SC_F$ doit être petite par rapport à la quantité $SC_R$. Par contre, si l'hypothèse alternative ($H_1$) 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 1.2.3.** *Nous appelons* **variance due au facteur** *le terme*

$$s_F^2 = \frac{SC_F}{I-1} \tag{1.2.7}$$

*et* **variance résiduelle** *le terme*

$$s_R^2 = \frac{SC_R}{n-I}. \tag{1.2.8}$$

*Si les* **trois Conditions 1.1.1.** *sont satisfaites et si l'hypothèse nulle* $(H_0)$ *est vraie alors*

$$f = \frac{s_F^2}{s_R^2} \tag{1.2.9}$$

*est une réalisation d'une variable* $F$ *qui est distribuée selon 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 1.2.1.** *Pour un seuil donné* $\alpha$ *(=0,05 en général), les tables des lois de Fisher nous donnent une valeur critique* $c$ *telle que* $\mathbb{P}_{(H_0)}(F \leq c) = 1-\alpha$. *Alors nous décidons :*

$$\begin{cases} \text{si } c \leq F_{obs} & (H_1) \text{ est vraie,} \\ \text{si } f < c & (H_0) \text{ est vraie.} \end{cases}$$

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 1.1.1.** 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 | 334,25 | 3 | 111,42 | 14,86 | 4,07 |
| Résiduelle | 60 | 8 | 7,50 | | |
| Totale | 394,25 | 11 | | | |

Pour un seuil de $\alpha = 0,05$, les tables des lois de Fisher nous donnent la valeur critique $c = 4,07$. Nous décidons donc que l'hypothèse alternative $(H_1)$ 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 1.2.2.** 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 1.2.3.** Le risque d'erreur de notre décision est ici le seuil, c'est-à-dire $\alpha = 0,05$. Le risque de deuxième espèce et le risque a posteriori peuvent être évalués, mais avec une démarche complexe.

## 1.3. Vérification des conditions

Nous étudions les possibilités d'évaluer la validité des **trois Conditions 1.1.1.** que nous avons supposées satisfaites.

### 1.3.1. Indépendance.

**Il n'existe pas**, dans un contexte général, **de test statistique 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.

### 1.3.2. Homogénéité.

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

**Hypothèses :**

$$(H_0) : \sigma_1^2 = \sigma_2^2 = \dots \sigma_I^2$$

contre

$$(H_1) : \text{les variances ne sont pas toutes égales.}$$

**Statistique :** nous considérons l'expression :

$$b = \frac{1}{C_1}\left[(n-I)\ln(s_R^2) - \sum_{i=1}^{I}(n_i - 1)\ln(s_{c,i}^2)\right] \tag{1.3.1}$$

où
- la quantité $C_1$ est définie par :

$$C_1 = 1 + \frac{1}{3(I-1)}\Big(\sum_{i=1}^{I}\frac{1}{n_i - 1} - \frac{1}{n-I}\Big), \tag{1.3.2}$$

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

**Propriété 1.3.1.** *Sous l'hypothèse nulle $(H_0)$ le nombre $b$ défini par* (1.3.1) *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.*

**Décision 1.3.1.** *Pour un seuil fixé $\alpha(= 0,05$ en général), les tables du Khi-deux fournissent une valeur critique $c$ telle que $\mathbb{P}_{(H_0)}[B \leq c] = 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} \text{si } c \leq b & (H_1) \text{ est vraie,} \\ \text{si } b < c & (H_0) \text{ est vraie.} \end{cases}$$

**Retour à l'Exemple 1.1.1.** Pour les données de cet exemple, nous avons :

$$C_1 = 1 + \frac{1}{3(4-1)}\left(\sum_{i=1}^{4}\frac{1}{3-1} - \frac{1}{12-4}\right) = 1,2083.$$

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

$$\begin{aligned} b &= \frac{1}{1,2083}\left[(12-4)\ln(7,5) - (3-1)\left(\ln\left(\frac{3}{2}\times 1,56\right) + \ln\left(\frac{3}{2}\times 2,89\right) + \ln\left(\frac{3}{2}\times 13,56\right) + \ln\left(\frac{3}{2}\times 2\right)\right)\right] \\ &= 2,7. \end{aligned}$$

Au seuil de $\alpha = 0,05$ la valeur critique d'un Khi-deux à 3 degrés de liberté, est $c = 7,815$. Comme $b < c$, nous décidons que l'hypothèse nulle $(H_0)$ est vraie, c'est-à-dire que l'hypothèse d'homogénéité est vérifiée.

### 1.3.3. La 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 1.1.1.** Ici, nous avons trois observations pour chaque échantillon..

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

$$E_{ij} = Y_{ij} - \mu_i,$$

alors

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

alors c'est la même loi pour l'ensemble des unités. Les moyennes $\mu_i$ étant inconnues, nous les estimons par les estimateurs bien connus de la moyenne : les $\overline{y}_i$. Les quantités obtenues s'appellent les **résidus (residuals)** et sont notées $\widehat{e}_{ij}$. Les résidus s'expriment par :

$$\widehat{e}_{ij} = y_{ij} - \overline{y}_i, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J. \tag{1.3.3}$$

Les résidus peuvent s'interpréter comme des estimations des erreurs de mesures. Nous pouvons alors tester la normalité, avec le **test de Shapiro-Francia**, sur l'ensemble des résidus. Nous rappelons la procédure avec **l'Exemple 1.1.1**.

**Retour à l'Exemple 1.1.1.** Pour effectuer le test de nomalité, nous construisons le tableau suivant :

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

$$f_{c,ij} = \frac{r_{ij} - 0,375}{n + 0,25}. \tag{1.3.4}$$

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

**Hypothèses :** en notant $\widehat{E}_{ij}$ la variable aléatoire dont $\widehat{e}_{ij}$ est la réalisation

$$(H_0)\ :\ \mathcal{L}(\widehat{E}_{ij}) = \mathcal{N}$$

contre

$$(H_1)\ :\ \mathcal{L}(\widehat{E}_{ij}) \neq \mathcal{N}.$$

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

$$r(\widehat{e}\ ;\ z) = \frac{1}{ns(\widehat{e})s(z)} \sum_{ij} \widehat{e}_{(ij)} z_{ij}. \tag{1.3.5}$$

**Propriété 1.3.2.** *Sous l'hypothèse nulle* $(H_0)$ *le nombre* $r$ *défini par* (1.3.5) *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 1.3.2.** *Pour un seuil fixé* $\alpha(= 0,05$ *en général), les tables de Shapiro-Francia fournissent une valeur critique* $c$ *telle que* $\mathbb{P}_{(H_0)}[R \leq c] = \alpha$. *Alors nous décidons :*

$$\begin{cases} \text{si } r \leq c & (H_1) \text{ est vraie,} \\ \text{si } c < r & (H_0) \text{ est vraie.} \end{cases}$$

**Retour à l'Exemple 1.1.1.** Pour cet exemple, les tables nous donnent, avec $n = 12$, la valeur critique $c = 0,9261$. Mais $r = 0,9853$. Comme $c < r$, l'hypothèse nulle $(H_0)$ est vraie, c'est-à-dire que **nous décidons que l'hypothèse de normalité est satisfaite.**

## 1.4. Comparaisons multiples

Lorsque pour la comparaison des moyennes théoriques la décision est "l'hypothèse alternative $(H_1)$ 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**.

### 1.4.1. 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** :

$$(H_0)\ :\ \mu_{(i)} = \mu_{(i')}$$

contre

$$(H_1)\ :\ \mu_{(i')} > \mu_{(i)}.$$

**Statistique :** nous considérons le rapport :

$$t_{i',i} = \frac{\overline{y}_{(i')} - \overline{y}_{(i)}}{\sqrt{\dfrac{s_R^2}{2}\left(\dfrac{1}{n_{i'}} + \dfrac{1}{n_i}\right)}}. \tag{1.4.1}$$

**Propriété 1.4.1.** *Le rapport* $t_{i',i}$ *défini par* (1.4.1) *est la réalisation d'une variable aléatoire* $T$ *qui, si l'hypothèse nulle* $(H_0)$ *est vraie, suit une loi appelée* **Etendue Studentisée (studentized range)** *et que nous notons* $\widetilde{T}_{n-I,I}$.

**Décision 1.4.1.** *Pour un seuil fixé* $\alpha(= 0,05$ *en général), les tables de l'étendue studentisée fournissent une valeur critique* $c$ *telle que* $\mathbb{P}_{(H_0)}[T \leq c] = 1 - \alpha$. *Alors nous décidons :*

$$\begin{cases} \textit{si } c \leq t_{i',i} & (H_1) \textit{ est vraie,} \\ \textit{si } t_{i',i} < c & (H_0) \textit{ est vraie.} \end{cases}$$

**Remarque 1.4.1.** 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 1.1.1.** Nous avons les moyennes dans l'ordre :

$$\overline{y}_{(1)} = 28,33 \quad \overline{y}_{(2)} = 31 \quad \overline{y}_{(3)} = 38,33 \quad \overline{y}_{(4)} = 41,33.$$

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

$$s_R^2 = 7,5,$$

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

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 :

$$\sqrt{\frac{7,5}{2}\left(\frac{1}{3} + \frac{1}{3}\right)} = 1,5811.$$

Les tables de l'étendue studentisée nous donne 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 |
|---|---|---|---|---|
| $\mu_{3 \text{ mois}} = \mu_{12 \text{ mois}}$ | 13,00 | 1,5811 | 8,2221 | $\mu_{3 \text{ mois}} > \mu_{12 \text{ mois}}$ |
| $\mu_{3 \text{ mois}} = \mu_{24 \text{ mois}}$ | 10,33 | 1,5811 | 6,5353 | $\mu_{3 \text{ mois}} > \mu_{24 \text{ mois}}$ |
| $\mu_{3 \text{ mois}} = \mu_{6 \text{ mois}}$ | 3,00 | 1,5811 | 1,8976 | $\mu_{3 \text{ mois}} = \mu_{6 \text{ mois}}$ |
| $\mu_{6 \text{ mois}} = \mu_{12 \text{ mois}}$ | 10,00 | 1,5811 | 6,3246 | $\mu_{6 \text{ mois}} > \mu_{12 \text{ mois}}$ |
| $\mu_{6 \text{ mois}} = \mu_{24 \text{ mois}}$ | 7,33 | 1,5811 | 4,6379 | $\mu_{6 \text{ mois}} > \mu_{24 \text{ mois}}$ |
| $\mu_{24 \text{ mois}} = \mu_{12 \text{ mois}}$ | 2,67 | 1,5811 | 1,6866 | $\mu_{24 \text{ mois}} = \mu_{12 \text{ mois}}$ |

**Remarque 1.4.2.** 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 sont $\frac{t_{i',i}}{\sqrt{2}}$ et peuvent donc être positives ou négatives. Ce sont les $p$ valeurs qui permettent de prendre les décisions.

*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, noté par la même lettre (A, ou B, ...).*

**Retour à l'Exemple 1.1.1.** 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 |

### 1.4.2. 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 :**

$$(H_0) \ : \ \mu_0 = \mu_i$$

contre

$$(H_1) \ : \ \mu_0 \neq \mu_i.$$

**Statistique :** nous considérons le rapport :

$$t_{0,i} = \frac{\overline{y}_0 - \overline{y}_i}{\sqrt{(\frac{1}{n_0} + \frac{1}{n_i})s_R^2}}. \tag{1.4.2}$$

**Propriété 1.4.2.** *Le rapport $t_{0,i}$ défini par* (1.4.2) *est la réalisation d'une variable aléatoire $T$ qui, lorsque l'hypothèse nulle $(H_0)$ est vraie, suit une loi dite de* **Dunnett**, *est notée $\mathcal{D}_{n-I,I}$.*

**Décision 1.4.2.** *Pour un seuil fixé $\alpha(= 0,05$ en général), les tables de Dunnett fournissent une valeur critique $c$ telle que $\mathbb{P}_{(H_0)}[-c \leq T \leq c] = 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} si \ t_{0,i} \leq -c, \ ou \ si \ c \leq t_{0,i} & (H_1) \ est \ vraie, \\ si \ \ -c < t_{0,i} < c & (H_0) \ est \ vraie. \end{cases}$$

**Remarque 1.4.3.** Notons que nous avons également des tests de ce type unilatéraux.

## 1.5. Risque de deuxième espèce

Lorsque nous décidons que l'hypothèse nulle $(H_0)$ est vraie pour le test 1.2.1 (voir Décision 1.2.1) 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 $\beta$. Celui-ci est égal à la probabilité de décider l'hypothèse nulle $(H_0)$ vraie alors qu'en réalité c'est l'hypothèse alternative $(H_1)$ qui l'est. Fixons dans l'hypothèse alternative $(H_1)$ des moyennes théoriques différentes $\mu_1, \ \mu_2, \ldots, \ \mu_I$.

Pour **calculer le risque** $\beta$ nous utilisons des abaques et l'expression :

$$\phi = \sqrt{\frac{J}{I\sigma^2} \sum_{i=1}^{I} (\mu_i - \overline{\mu})^2} \tag{1.5.1}$$

où

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

En général il est impossible de calculer la quantité $\phi$ car il faudrait connaître la variance $\sigma^2$. 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 $\phi$. Les **abaques** pour $\nu_1 = I - 1, \ \nu_2 = n - I$ et $\phi$ 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 $\mu_i$. Dans ce cas il peut fonder son calcul du risque $\beta$ 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 $s_F^2 - s_R^2 \geq 0$)

$$\phi = \sqrt{\frac{(I-1)(s_F^2 - s_R^2)}{I s_R^2}}. \tag{1.5.2}$$

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 $\beta$ donné. **En général la valeur de 0,20 est considérée comme satisfaisante.**

## 1.6. 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 $\lambda$. Lorsque nous avons à faire à des proportions nous pouvons utiliser la transformation décrite dans la propriété suivante :

**Propriété 1.6.1.** *Si pour une variable $Y$ nous avons $\mathcal{L}(Y) = \mathcal{B}(n\ ;\ p)$ alors nous obtenons le résultat :*

$$\lim_{n \longrightarrow +\infty} \mathcal{L}\left(\sqrt{n}\left(\left(\arcsin\left(\sqrt{\frac{Y}{n}}\right) - \arcsin(\sqrt{2p})\right)\right)\right) = \mathcal{N}(0\ ;\ 1). \tag{1.6.1}$$

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, les conditions ne sont pas satisfaites, alors il faut utiliser le **test non paramétrique Kruskal-Wallis** (cf paragraphe 8).

## 1.7. Facteurs aléatoires

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

**Exemple 1.7.1.** 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

- $\mu$ 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 :**

$$(H_0) : \sigma_A = 0$$

contre

$$(H_1) : \sigma_A \neq 0.$$

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 $\beta$ plus compliqué à évaluer.

## 1.8. 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 :**

$$(H_0) : \mathcal{L}_1 = \ldots = \mathcal{L}_I$$

contre

$$(H_1) : \text{Les lois } \mathcal{L}_i \text{ sont différentes.}$$

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 :

$$R_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}, \quad i = 1, \ldots, I. \tag{1.8.1}$$

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 :

$$\overline{R} = \frac{1}{n} \sum_{i=1}^{I} n_i R_{i\bullet} = \frac{n+1}{2}. \tag{1.8.2}$$

Cette quantité s'appelle la moyenne générale.

### 1.8.1. 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 :

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{I} n_i \left( R_{i\bullet} - \frac{n+1}{2} \right)^2. \tag{1.8.3}$$

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 (1.8.3) peut s'écrire plus simplement :

$$K = \left( \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} \right) - 3(n+1). \tag{1.8.4}$$

**Remarque 1.8.1.** 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 1.8.1.** *Lorsque $H_0$ est vraie et lorsque $n \longrightarrow +\infty$ nous avons :*

$$\mathcal{L}_{H_0}(K) = \chi^2_{I-1},$$

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

Il existe des tables exactes de $K$, sous l'hypothèse nulle ($H_0$), mais nous utiliserons la loi asymptotique ci-dessus.

**Décision 1.8.1.** *Pour un seuil fixé $\alpha$ ($= 0,05$ en général), les tables du $\chi^2_{I-1}$ nous fournissent une valeur critique $c$ telle que $\mathbb{P}[K < c] \geq 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} (H_1) \text{ est vraie si} & c \leq K, \\ (H_0) \text{ est vraie si} & K < c. \end{cases}$$

**Exemple 1.8.1.** 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 :

$$(H_0) :\ \mathcal{L}_A = \mathcal{L}_B = \mathcal{L}_C$$

contre

$$(H_1) :\ \mathcal{L}_A,\ \mathcal{L}_B,\ \mathcal{L}_C \text{ sont différentes.}$$

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

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Echantillon $A$ | 62 | 64 | 67 | 70 | | | 73 | | | | | | |
| Echantillon $B$ | | | | | | | | 77 | | 80 | 81 | | 84 |
| Echantillon $C$ | | | | | 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 :

| | $R_i$ | $R_i/n_i$ | $R_i^2$ |
|---|---|---|---|
| Echantillon $A$ | 16 | 3,2 | 256 |
| Echantillon $B$ | 42 | 10,5 | 1764 |
| Echantillon $C$ | 33 | 8,25 | 1089 |

En utilisant l'expression (1.8.4), nous pouvons alors calculer la statistique $K$ du **test de Kruskal-Wallis** :

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

Le seuil $\alpha$ étant fixé et égal à $0,05$, les tables du Khi-deux à $I-1=3-1=2$ degrés de liberté nous donnent la valeur critique 5,991. Comme $c<K$, nous décidons que l'hypothèse alternative $(H_1)$ est vraie, c'est-à-dire que **nous décidons que les lois $\mathcal{L}_i$ sont différentes.**

### 1.8.2. 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 :

$$K^*=\frac{\left(\frac{12}{n(n+1)}\sum_{i=1}^{I}\frac{R_i^2}{n_i}\right)-3(n+1)}{1-\frac{1}{n^3-n}\sum_{h=1}^{n}(d_h^3-d_h)}, \tag{1.8.5}$$

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.

## 1.9. Quelques précisions sur les comparaisons multiples

Dans le cas où la décision a été : "$(H_1)$ 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 $\alpha$ 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

$$\left|\frac{R_i}{n_i}-\frac{R_{i'}}{n_{i'}}\right|>c(\alpha,n-I)\sqrt{S^2\times\left(\frac{n-1-K}{n-I}\right)}\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)}, \tag{1.9.1}$$

où

- dans le cas sans ex æquo

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

- dans le cas avec ex æquo

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

**Retour à l'Exemple 1.8.1.** Nous organisons les comparaisons multiples de la manière suivante. Nous devons calculer le terme :

$$c(5\%, 13-3)\sqrt{\frac{13(13+1)}{12} \times \left(\frac{13-1-8,403}{13-3}\right)} \approx 5,2039.$$

Ce terme doit être multiplié par :

$$\sqrt{\left(\frac{1}{5}+\frac{1}{4}\right)} \approx 0,6708$$

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

$$\sqrt{\left(\frac{1}{4}+\frac{1}{4}\right)} \approx 0,7071.$$

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 :

$$3,4908$$

et

$$3,6797.$$

Nous effectuons ces comparaisons dans le tableau suivant :

| Comparaisons | $\left\|\frac{R_i}{n_i}-\frac{R_{i'}}{n_{i'}}\right\|$ | c | Décision |
|---|---|---|---|
| $\mathcal{L}_A$ avec $\mathcal{L}_B$ | 7,3 | 3,4908 | $\mathcal{L}_A \neq \mathcal{L}_B$ |
| $\mathcal{L}_B$ avec $\mathcal{L}_C$ | 2,25 | 3,6797 | $\mathcal{L}_B = \mathcal{L}_C$ |
| $\mathcal{L}_A$ avec $\mathcal{L}_C$ | 5,05 | 3,4908 | $\mathcal{L}_A \neq \mathcal{L}_C$ |

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

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

Ceci complète l'**Exemple 1.8.1.**.

**Exemple 1.9.1.** 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 :

**Hypothèses :**

$(H_0)$ : les traitements sont identiques

contre

$(H_1)$ : 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<br>36 |
| Groupe $B$ | 24 | | | | | | 34<br>34 | 35 | |
| Groupe $C$ | | 26 | 28 | 30 | 32 | 33 | | | 36 |
| Groupe $D$ | | | | 30 | 32<br>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<br>42 | | 46 | | 50 |
| Groupe $B$ | 37 | | 40 | 42<br>42 | 44 | | | 50 |
| Groupe $C$ | | 38 | 40 | 42 | 44 | | | |
| Groupe $D$ | | | 40 | 42 | | 46<br>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** :

$$\begin{aligned} K &= \frac{12}{40(40+1)}\left(\frac{220^2}{10}+\frac{222^2}{10}+\frac{159^2}{10}+\frac{219^2}{10}\right)-3(40+1) \\ &\approx 2,0678. \end{aligned}$$

Mais la présence d'ex æquo nécessite la correction de cette statistique $K$ par :

$$1-\frac{1}{40^3-40}\left(3(2^3-2)+3(3^3-3)+3(4^3-4)+(6^3-6)\right)=0,9924.$$

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

$$K^*=\frac{K}{0,9924}=2,0836.$$

Nous admettons que nous pouvons approcher la loi $\mathcal{L}_{H_0}(K^*)$ par une loi du $\chi^2_{4-1}$. Les tables de cette dernière nous donnent, pour un seuil $\alpha$ égal à $0,05$, la valeur critique $c = 7,81$. Comme $K^* < c$, **nous décidons ainsi que l'hypothèse nulle** $(H_0)$ **est vraie, c'est-à-dire qu'il n'y a pas de différence entre les traitements.**

# Chapitre 2

# Analyse de régression linéaire : Corrélation linéaire - Régression linéaire simple

## 2.1. Introduction

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

**Exemple 2.1.1.** 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 personnes.

| Age $X$ | Cholestérol $Y$ |
|---|---|
| 46 | 181 |
| 52 | 228 |
| 39 | 182 |
| 65 | 249 |
| 54 | 259 |
| 33 | 201 |
| 49 | 121 |
| 76 | 339 |
| 71 | 224 |
| 41 | 112 |
| 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 $X$, il correspondra qu'une seule valeur de $Y$. 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 $n = 11$.

**Exemple 2.1.2.** 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 $X$ il correspond plusieurs valeurs de $Y$ (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 $J = 3$, $I = 4$ et par conséquent $n = IJ = 12$.

Ce plan 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 $x_i$, qu'il est possible d'utiliser pour ce deuxième cas les notations du premier plan.

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

1. Exprimer l'intensité de la liaison éventuelle entre $X$ 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 2.1.1.** Il est très important, avant l'expérimentation, de bien préciser les objectifs et ensuite choisir le plan expérimental approprié.

## 2.2. Le coefficient de corrélation linéaire

Considérons un couple de variables aléatoires $(X, Y)$. Nous supposons que ce couple est distribué selon une loi **normale**. Cette loi dépend de 5 paramètres. Les 4 premiers, notés $\mu_X$, $\mu_Y$ et $\sigma_X$, $\sigma_Y$ sont, respectivement, les moyennes et les écarts-type théoriques des variables $X$ et $Y$.

**Remarque 2.2.1.** Il ne suffit pas de tester la normalité de chaque variable séparement, 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. Dans toute la suite la normalité d'un couple sera toujours admise.

**Définition 2.2.1.** *Le cinquième paramètre noté $\varrho = \varrho(X, Y)$ est appelé* **coefficient de corrélation linéaire** *théorique de $X$ et de $Y$. En fait ce coefficient se déduit d'un autre paramètre, appelé la* **covariance** *théorique de $X$ et de $Y$, noté $\mathrm{Cov}(X, Y)$, qui est égal à la moyenne théorique moins le produit des moyennes théoriques $\mu_X$ et $\mu_Y$. Nous avons :*

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

**Propriétés 2.2.1.**
- *$\varrho$ est bien un coefficient, il ne dépend pas des unités des variables observées.*
- *$\varrho$ est compris entre -1 et 1. Si $\varrho > 0$ (respectivement $\varrho < 0$), les variables $X$ et $Y$ varient dans le même sens (resp. le sens contraire).*
- *Si $\varrho = -1$ ou $1$, alors il existe deux nombres fixes $a$ et $b$, tels que*

$$Y = a + bX.$$

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

**Remarque 2.2.2.** Dans le cas de couple $(X, Y)$ suivant une loi normale, ce coefficient décrit la totalité de la liaison entre les deux variables.

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

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

Nous utilisons la notation la plus simple. Nous estimons en premier lieu la covariance théorique par la covariance observée :

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

**Définition 2.2.2.** *Nous appelons* **coefficient de corrélation observé**, *la statistique définie par :*

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

*Le nombre $r$ est une réalisation de la variable $R$.*

**Interprétation 2.1.1.** Comme nous le verrons par la suite, nous interprétons $r^2(x,\ y)$ comme la part de variation de $Y$ expliquée par (ou due à) une expression linéaire en $X$, c'est-à-dire $a + bX$.

**Retour à l'Exemple 2.1.1.** Nous calculons le coefficient de corrélation $r(x, y)$. Nous avons :

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

et

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

Pour les variances nous avons :

$$s^2(x) = \left(\frac{32834}{11}\right) - 53,091^2 = 166,255 \quad \text{et} \quad s(x) = \sqrt{166,255} = 12,894,$$

et

$$s^2(y) = \left(\frac{515355}{11}\right) - 207,727^2 = 3699,948 \quad \text{et} \quad s(y) = \sqrt{3699,948} = 60,827.$$

Enfin, la somme des produits est égale à :

$$\sum_{i=1}^{11} x_i y_i = 127235,$$

et la covariance à :

$$\mathrm{Cov}(x,y) = \left(\frac{127235}{11}\right) - 53,091 \times 207,727 = 538,384.$$

Nous pouvons à présent calculer le coefficient de corrélation :

$$r(x,y) = \frac{538,384}{12,894 \times 60,827} = 0,686.$$

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

**Retour à l'Exemple 2.1.2.** Nous calculons le coefficient de corrélation $r(x,y)$. Nous avons :

$$\overline{x} = 6,5$$

et

$$\overline{y} = 281,4.$$

Pour les écarts-type nous avons :

$$s(x) = 3,84$$

et

$$s(y) = 145,15.$$

De plus la covariance est égale à :

$$\mathrm{Cov}(x,y) = \left(\frac{15631}{12}\right) - (6,5 \times 281,4) = -526,52.$$

Nous pouvons à présent calculer le coefficient de corrélation :

$$r(x,y) = \frac{-526,52}{3,84 \times 145,15} = -0,945$$

et

$$r^2(x,y) = 0,892.$$

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

**Remarque 2.2.3.** Toutes les méthodes statistiques ne permettent de mettre en évidence, le cas échéant, qu'une liaison entre les variables $X$ et $Y$ et en aucun cas cas une relation de cause à effet. Cette dernière n'est qu'une interprétation donnée par l'utilisateur.

## 2.3. Tests d'hypothèses

Nous avons vu qu'une corrélation nulle est équivalente, dans le cas gaussien à l'indépendance. C'est pourquoi nous présentons le test d'hypothèses :

$$(H_0) : \varrho(X,Y) = 0$$

contre

$$(H_1) : \varrho(X,Y) \neq 0.$$

La loi de $R$, sous l'hypothèse nulle $(H_0)$ est très difficile à calculer. Nous allons utiliser une transformation qui nous ramène à une loi connue.

**Propriété 2.3.1.** *Si l'hypothèse nulle* $(H_0)$ *est vraie et dans le cadre de la normalité, en posant*

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

*nous avons* $\mathcal{L}(t) = \mathcal{T}_{n-2}$*, loi de Student à* $n-2$ *degrés de liberté.*

**Décision 2.3.1.** *Pour un seuil fixé* $\alpha(= 0,05$ *en général), les tables de Student, à* $n-2$ *degrés de liberté, nous fournissent une valeur critique* $c$ *telle que* $\mathbb{P}(-c \leq \mathcal{T}_{n-2} \leq c) = 1-\alpha$*. Si nous utilisons un logiciel, par exemple le logiciel* **Minitab***, celui-ci nous fournit une p-valeur. Alors nous décidons*

$$\begin{cases} (H_0) \text{ est vraie} & \text{si } -c < t < c, \qquad \text{ou si } p > \alpha \\ (H_1) \text{ est vraie} & \text{si } t \leq -c \text{ ou } c \leq t, \quad \text{ou si } p \leq \alpha. \end{cases}$$

**Retour à l'Exemple 2.1.1.** Nous testons :

**Hypothèses :**

$$(H_0) : \text{Le cholestérol est indépendant de l'âge}$$

contre

$$(H_1) : \text{Le cholestérol est dépendant de l'âge.}$$

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

$$t_{obs} = \frac{0,686\sqrt{11-2}}{\sqrt{1-0,471}} = 2,8296.$$

Les tables de la loi de Student à 9 degrés nous donnent, pour $\alpha = 0,05$, la valeur critique $c = 2,201$. Nous décidons donc l'hypothèse alternative $(H_1)$, c'est-à-dire "le cholestérol est dépendant de l'âge" est vraie. Le risque de cette décision est aussi le seuil de test, c'est-à-dire $\alpha = 0,05$.

**Retour à l'Exemple 2.1.2.** Dans cet exemple il ne faut pas effectuer le test précédent. En effet la variable $X$ est totalement fixée par l'expérimentateur ; elle n'est pas aléatoire. Ainsi nous n'avons de loi normale pour le couple. Mais nous verrons par la suite comment tester l'indépendance dans ce cas.

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

**Hypothèses :**

$$(H_0) : \varrho = \varrho_0$$

contre

$$(H_1) : \varrho \neq \varrho_0.$$

**Propriété 2.3.2.** *Posons :*

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

*et*

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

*R. A. Fisher a montré que, si l'hypothèse nulle* $(H_0)$ *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* $n \geq 30$*, nous pouvons utiliser cette loi asymptotique pour réaliser le test.*

**Décision 2.3.2.** *Pour un seuil fixé $\alpha(=0,05$ en général), les tables de la loi normale centrée et réduite nous fournissent une valeur critique $c(=1,96)$ telle que $\mathbb{P}[-c \leq \mathcal{N}(0;1) \leq c] = 1-\alpha$. Alors nous décidons*

$$\begin{cases} (H_1) \text{ est vraie si} & \sqrt{n-3}(z-z_0) \leq -c \text{ ou } c \leq \sqrt{n-3}(z-z_0), \\ (H_0) \text{ est vraie si} & -c < \sqrt{n-3}(z-z_0) < c. \end{cases}$$

**Remarque 2.3.1.** 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 2.1.1.** Posons $\varrho_0 = 0,75$ et faisons le test suivant :

**Hypothèses :**

$(H_0)$ : Le cholestérol et l'âge ont une corrélation égale à 0,75

contre

$(H_1)$ : Le cholestérol et l'âge ont une corrélation qui n'est pas égale à 0,75.

Des résultats numériques précédents nous obtenons :

$$z = \frac{1}{2}\ln\left(\frac{1+0,686}{1-0,686}\right) = 0,8404$$

et

$$z_0 = \frac{1}{2}\ln\left(\frac{1+0,75}{1-0,75}\right) = 0,9730.$$

D'où

$$\sqrt{11-3}(z-z_0) = -0,375.$$

Cette valeur comparée à $1,96$ nous permet de décider que l'hypothèse nulle $(H_0)$ "*Le cholestérol et l'âge ont une corrélation égale à 0,75*" est vraie. Notons que le risque associé à cette décision est inconnu. De plus elle n'est pas fondée dans la mesure où $n = 11 < 30$. Nous avons effectué ce test uniquement pour présenter un exemple d'application.

## 2.4. Intervalle de confiance

La même approximation permet de construire un intervalle de confiance de $\varrho$ au seuil $\alpha$ égal à $0,05$ :

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

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 2.1.1.** Notons que $n = 11$ est très faible pour utiliser une loi asymptotique, mais nous construisons néanmoins un intervalle de confiance de $\rho$ pour montrer le déroulement des calculs. Nous avons :

$$z = \frac{1}{2}\ln\left(\frac{1+0,686}{1-0,686}\right) = 0,840.$$

Les bornes de l'intervalle dépendent de :

$$z_1 = 0,840 - \frac{1,96}{\sqrt{11-3}} = 0,840 - 0,693 = 0,147$$

et

$$z_2 = 0,840 + \frac{1,96}{\sqrt{11-3}} = 0,840 + 0,693 = 1,533.$$

Les formules correspondant à la fonction inverse de

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

nous donnent :

$$r_1 = \frac{e^{0,147} - e^{-0,147}}{e^{0,147} + e^{-0,147}} = 0,146$$

et

$$r_2 = \frac{e^{1,533} - e^{-1,533}}{e^{1,533} + e^{-1,533}} = 0,911.$$

En conclusion nous affirmons que $[0,146; 0,911]$ est parmi les 95% environ d'intervalles que nous pouvons construire avec cette méthode, qui contiennent la vraie valeur inconnue $\rho$.

**Remarque 2.4.1.** Remarquons que l'intervalle est très large. Ceci est dû, entre autres, au fait que la taille de l'échantillon, $n = 11$, est très faible.

## 2.5. Le rapport de corrélation

Nous nous plaçons dans le cadre du **plan expérimental 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

$$n = IJ$$

où

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

observations. Nous rappelons l'équation de l'Analyse de la Variance :

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

Nous avions abrégé cette écriture (cf. Chapitre 1, équation (1.2.3)) en :

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

avec les notations

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

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

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

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

Nous posons :

**Définition 2.5.1.** *Nous appelons* **coefficient de détermination** *le nombre* $r^2(x, y)$.

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 sont non corrélées linéairement. Un autre coefficient permettant l'étude de la dépendance de $Y$ par rapport à $x$, avec ce plan, est le suivant :

**Définition 2.5.2.** *Nous appelons* **rapport de corrélation** *de $Y$ en $x$ la quantité :*

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

**Remarques 2.5.1.** • 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 $Y$ "expliquée" par la variation de $x$ (et pas par une fonction linéaire de $x$).

- L'équation de l'Analyse de la Variance implique que :

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

- 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 $x$.

- 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 $Y$ par rapport à $x$. Nous avons également la relation suivante :

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

**Interprétation 2.5.1.** *Pour décrire le type de dépendance de $Y$ par rapport à $x$, 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 $Y$ ne dépend pas de $x$.*
- *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 $Y$ dépend fonctionnellement de $x$ 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 $Y$ dépend partiellement de $x$, mais cette liaison partielle n'est pas de la forme d'une droite.*
- *si $r^2(x, y)$ et $\eta^2(y \mid x)$ sont moyens, alors il sera admis que $Y$ dépend partiellement de $x$ 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 $Y$ dépend fonctionnellement de $x$, 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 $\eta^2$ n'a de sens que si nous avons appliqué **le plan expérimental 2**.

**Retour à l'Exemple 2.1.1.** Notons que nous ne pouvons pas calculer ici le rapport de corrélation $\eta^2(y \mid x)$. En effet nous avons appliqué **le plan 1** et pour chaque valeur $x_i$ nous avons observé qu'une seule valeur $y_i$.

**Retour à l'Exemple 2.1.2.** Le rapport de corrélation $\eta^2(y \mid x)$ peut être calculé dans cet exemple. Nous avons **le plan 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 $x$ est qualitatif :

| 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 :

$$\eta^2(y \mid x) = \frac{242621,73}{252857,56} = 0,9595.$$

La concentration de l'hormone de croissance dépend fonctionnellement du temps écoulé après l'injection, mais comme $r^2(x,y) = 0,892$ nous ne pouvons pas approcher cette fonction par une droite.

## 2.6. 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 2.6.1.** *Par* **régression linéaire simple***, nous entendons le modèle suivant :*

$$Y_i = a + bx_i + E_i, \qquad (i = 1, \ldots, n),$$

*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 $E_i$ 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 $\sigma^2$.*

Considérons le modèle d'Analyse de la Variance à un facteur :

$$Y_i = \mu_i + E_i, \qquad (i = 1, \ldots, n),$$

où les $\mu_i$ sont des paramètres fixes et inconnus, et $E_i$ des variables aléatoires centrées. La régression stipule simplement que les $\mu_i$ dépendent des nombres connus $x_i$ et d'une manière linéaire des paramètres inconnus $a$ et $b$ :

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

**Remarque 2.6.1.** Nous avons la même définition avec le plan expérimental 2 :

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

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

## 2.7. La méthode des moindres carrés ordinaire

Nous estimons les deux paramètres inconnus $a$ et $b$.

**Définition 2.7.1.** *Nous appelons méthode des moindres carrés 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 2.7.1.** Nous cherchons donc la droite en $x$ qui passe globalement le plus près possible des $y$.

**Définition 2.7.2.** *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 $\{(x_i,\ \overline{y}_i)\}$ 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 :

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

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

**Définition 2.7.3.** *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 2.7.2.** Nous avons deux propriétés de la droite de régression :
- la somme des erreurs est nulle,

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

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

**Remarque 2.7.3.** Nous estimons la droite de régression ; les coefficients sont donnés par :

$$\widehat{b} = 0,686\frac{60,827}{12,894} = 3,236\ ,$$

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

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

$$\text{Cholestérol} = 35,915 + 3,236 \times \text{Age}.$$

**Remarque 2.7.4.** Les coefficients sont donnés par :

$$\widehat{b} = -0,945\frac{145,15}{3,84} = -35,721\ ,$$

$$\widehat{a} = 281,4 + 35,721 \times 6,54 = 515,01\ .$$

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

$$\text{Concentration} = 515,01 - 35,721 \times \text{Minutes}.$$

## 2.8. 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 :

$$(H_0)\ :\ \mu_i = a + b\ x_i,\ \forall i = 1, \ldots, I$$

contre

$$(H_1)\ :\ \text{au moins un } \mu_i \neq a + bx_i.$$

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

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

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 $\sigma^2$. 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 de 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 $s^2$ | $F_{obs}$ |
|---|---|---|---|---|
| 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 | $SC_R = \sum(y_{ij} - \overline{y}_i)^2$ $= (1 - \eta^2(y \mid x))ns^2(y)$ | $n-I$ | $s^2_R = \dfrac{SC_R}{n-I}$ | $f_M = \dfrac{s^2_{M\|RG}}{s^2_R}$ |
| 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 2.8.1.** *Pour un seuil donné $\alpha(= 0,05$ en général), les tables des lois de Fisher nous donnent une valeur critique $c$ telle que $\mathbb{P}[\mathcal{F}_{I-2,n-I} \leq c] = 1 - \alpha$. Alors nous décidons*

$$\begin{cases} H_1 \text{ est vraie si} & f_M \geq c, \\ H_0 \text{ est vraie si} & f_M < c. \end{cases}$$

**Remarque 2.8.1.** 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

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

**Retour à l'Exemple 2.1.2.** Nous avons déjà calculé :

$$\eta^2(y|x) = 0,9595, \quad \text{et} \quad r^2(x,\ y) = 0,892.$$

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

| Variation | $SC$ | $ddl$ | $s^2$ | f |
|---|---|---|---|---|
| De la droite | 17226 | 2 | 8613 | 6,73 |
| Résiduelle | 10236 | 8 | 1279 | |

Les tables nous donnent, pour 2 et 8 degrés de liberté, la valeur critique $c = 4,46$ ; nous en déduisons : "$(H_1)$ est vraie", c'est-à-dire, nous n'avons pas le modèle de la droite, la concentration ne varie pas comme une droite du temps. Si nous effectuons le changement de variable $u_{ij} = Log(y_{ij})$, nous obtenons :

$$\overline{u} = 5,4935,\ s(u) = 0,55398,\ \frac{1}{12}\sum_{ij} x_i u_{ij} = 33,623.$$

Comme nous avons déjà calculé $\overline{x} = 6,5$ et $s(x) = 3,84$, nous en déduisons :

$$cov(x,\ u) = 33,623 - 6,5 \times 5,4935 = -2,08475,$$

$$r(x,\ u) = \frac{-2,08475}{3,84 \times 0,55398} = -0,98$$

et

$$r^2(x,\ u) = 0,96041.$$

Nous pouvons à présent calculer les estimations de $a$ et de $b$, et donner la droite de régression :

$$Log(\text{Concentration}) = 6,41 - 0,1411 \times \text{Minutes}.$$

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

| Variation | $SC$ | $ddl$ | $s^2$ | $F_{obs}$ |
|---|---|---|---|---|
| Due au Facteur | 3,5703 | 3 | 1,1901 | 84,67 |
| Résiduelle | 0,1125 | 8 | 0,0141 | |
| Totale | 3,6827 | 11 | | |

Ce tableau nous permet de calculer :

$$\eta^2(u|x) = \frac{3,5703}{3,6827} = 0,9695.$$

Rappelons que nous avions obtenu :

$$\eta^2(y|x) = 0,9595 \quad \text{et} \quad r^2(x,\ y) = 0,892.$$

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

$$H_0\ :\ \mu(U_{ij}) = a + b\ x_i \quad \text{contre} \quad H_1\ :\ \text{au moins un } \mu(U_{ij}) \neq a + b\ x_i.$$

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 |
|---|---|---|---|---|
| De la droite | 0,0363 | 2 | 0,0181 | 1,29 |
| Résiduelle | 0,1125 | 8 | 0,0141 | |

La valeur $f = 1,29$ comparée à la même valeur critique $c = 4,46$ nous permet de décider que l'hypothèse $(H_0)$ 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'homogénéïté, le modèle de la droite 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 $\sigma^2$.

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

$$(H_0)\ :\ Y\text{ne dépend pas de } x \quad \text{contre} \quad H_1\ :\ Y\text{dépend de } x.$$

Ceci peut se dire :

$$(H_0)\ :\ b = 0 \quad \text{contre} \quad H_1\ :\ b \neq 0.$$

La statistique de test s'écrit :

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

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

$$(H_0)\ :\ \varrho = 0 \quad \text{contre} \quad H_1\ :\ \varrho \neq 0,$$

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.

## 2.9. Vérification des conditions

### 2.9.1. La normalité

Nous considérons les résidus de l'analyse de régression (cf. Définition 2.4.) :

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

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.

### 2.9.2. 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.

### 2.9.3. 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 cours 1, 3.3.1.).

**Retour à l'Exemple 2.1.1.** 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.

**Exemple 2.1.2.** 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.

## 2.10. 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 :

$$Y_i = a + bx_i + E_i, \qquad i = 1, \ldots, n,$$

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

**Propriétés 2.10.1.** *La meilleure estimation de $\sigma^2$ 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 2.10.2.** *Nous avons les résultats suivants :*

$$1) \quad \mathcal{L}\left(\frac{\widehat{A} - a}{S_A}\right) = \mathcal{T}_{n-2},$$

$$2) \quad \mathcal{L}\left(\frac{\widehat{B} - b}{S_B}\right) = \mathcal{T}_{n-2},$$

$$3) \quad \mathcal{L}\left(\frac{\widehat{Y}(x) - a - bx}{S_{\widehat{Y}(x)}}\right) = \mathcal{T}_{n-2},$$

*avec*

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

### 2.10.1. Intervalles de confiance

**Propriétés 2.10.3.** *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 $\alpha$. Alors un intervalle de confiance de $a$ au seuil $\alpha$ est donné par :*

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

*Un intervalle de confiance de $b$ au seuil $\alpha$ est donné par :*

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

*Soit $x$ fixéAlors un intervalle de confiance de $y(x) = a + bx$ au seuil $\alpha$ 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 :*

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

**Remarque 2.10.1.** Notons que pour obtenir des intervalles de faible amplitude, nous devons avoir une taille d'échantillonnage $n$ 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 2.1.1.** 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 $r^2 = 0,471$ et que $s^2(y) = 3699,948$ ). Nous avons ainsi :

$$s_B^2 = \frac{2392,22}{11 \times 166,255} = 1,3081$$

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 $b$ :

$$[3,236 - 2,2622 \times 1,143 \ ; \ 3,236 + 2,2622 \times 1,143] = [0,6503 \ ; \ 5,8217] \ ;$$

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

$$Log(\widehat{y(10)}) = \widehat{a} + \widehat{b}10 = 6,41 - 0,1411 \times 10 = 4,999 \quad \text{et} \quad \widehat{y(10)} = e^{4,999} = 148,265.$$

Pour construire un intervalle de confiance, nous avons $s_{R|M}^2 = 0,0149$, $\overline{x} = 6,5$ et $s^2(x) = 14,7456$. La propriété 5.2. nous donne :

$$s_{\widehat{u(10)}}^2 = \frac{0,0149}{12}\left(1 + \frac{(10-6,5)^2}{14,7456}\right) = 0,002273 \quad \text{et} \quad s_{\widehat{u(10)}} = 0,0477.$$

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

$$[4,999 - 2,2281 \times 0,0477 \ ; \ 4,999 + 20,2281 \times 0,0477] = [4,8927 \ ; \ 5,1053]\,,$$

Pour $y(10)$ nous en déduisons :

$$\left[e^{4,8927} \ ; \ e^{5,1053}\right] = [133,3130 \ ; \ 164,8903]\,.$$

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

### 2.10.2. Tests d'hypothèses .

Nous testons les hypothèses :

$$(H_0) \ : \ a = a_0$$

contre

$$(H_1) : \ a \neq a_0.$$

**Statistique** : Nous considérons la statistique suivante

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

où $s_A$ est donné par la Propriété 5.2.

**Propriétés 2.10.4.** *Le nombre $t$ est la réalisation d'une variable $T$ dont la loi, lorsque l'hypothèse nulle $(H_0)$ est vraie, est une loi de Student à $n-2$ degrés de liberté.*

**Décision 2.10.1.** *Pour un seuil $\alpha$ (=0,05 en général), les tables du $\mathcal{T}$ à $n-2$ degrés de liberté nous fournissent une valeur critique bilatérale $c$ telle que $P(-c < \mathcal{T}_{n-2} < c) = 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} (H_1) \text{ est vraie si} & t \leq -c \text{ ou si } t \geq c, \\ (H_0) \text{ est vraie si} & -c < t < c. \end{cases}$$

Nous testons les hypothèses :

$$(H_0) : \ b = b_0$$

contre

$$(H_1) : \ b \neq b_0.$$

**Statistique** : Nous considérons la statistique suivante

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

où $s_B$ est donné par la propriété 5.2.

**Propriétés 2.10.5.** *Le nombre $t$ est la réalisation d'une variable $T$ dont la loi, lorsque l'hypothèse nulle $(H_0)$ est vraie, est une loi de Student à $n-2$ degrés de liberté.*

**Décision 2.10.2.** *Pour un seuil $\alpha$ (=0,05 en général), les tables du $\mathcal{T}$ à $n-2$ degrés de liberté nous fournissent une valeur critique bilatérale $c$ telle que $\mathbb{P}[-c < \mathcal{T}_{n-2} < c] = 1 - \alpha$. Alors nous décidons :*

$$\begin{cases} (H_1) \text{ est vraie si} & t \leq -c \text{ ou si } t \geq c, \\ (H_0) \text{ est vraie si} & -c < t < c. \end{cases}$$

**Remarque 2.10.2.** Les logiciels donnent en général la $p$-valeur de ces deux tests pour $a_0 = 0$ et pour $b_0 = 0$.