STT-1920
Méthodes statistiques

Claude Bélisle
Département de mathématiques et de statistique
Université Laval

Janvier 2011

Avant-propos

Cet ouvrage est utilisé comme manuel de référence pour le cours STT-1920 *Méthodes statistiques* offert par le département de mathématiques et de statistique de l'Université Laval. Ce cours s'adresse principalement aux étudiants des programmes de baccalauréat en agronomie et de baccalauréat en biologie pour qui il est cours obligatoire. Il s'adresse également aux étudiants des programmes de baccalauréat en sciences et technologie des aliments, de baccalauréat en kinésiologie et de baccalauréat en service social pour qui il est cours à option.

Les sept chapitres de cet ouvrage correspondent au contenu du cours STT-1920 tel que décidé en 2006 lors de discussions entre un groupe d'enseignants du département de mathématiques et de statistique et les directeurs des programmes concernés. En donnant ce cours, l'enseignant devrait donc s'assurer de couvrir les sept chapitres. Pour y arriver, il devra probablement omettre quelques sections. Dans la table des matières, les sections marquées d'une étoile (★) peuvent être ou bien omises ou bien laissées en lecture, sans compromettre la continuité et l'intégrité du cours.

Une série d'exercices apparaît à la fin de chaque chapitre. Il est important que les étudiants fassent ces exercices pour s'assurer d'avoir bien compris la matière. Certains exercices nécessitent l'utilisation d'un logiciel d'analyse statistique. Le logiciel recommandé pour le cours STT-1920 est le logiciel R, disponible gratuitement à partir du site Internet `www.r-project.org/`. Les étudiants auront l'occasion de s'initier à ce logiciel durant la séance hebdomadaire de travaux pratiques qui s'ajoute aux trois heures d'enseignement magistral.

Je remercie tous ceux et celles qui, de loin ou de près, ont contribué à la réalisation de cet ouvrage. Je remercie en particulier Madame Emmanuelle Reny-Nolin, directrice du *Centre de dépannage et d'apprentissage en mathématiques et en statistique* de l'Université Laval. Emmanuelle a lu avec attention la première édition de cet ouvrage et ses nombreux commentaires ont grandement contribué à améliorer le texte. Je remercie également Jean-Hubert Smith Lacroix pour ses nombreux commentaires sur la dernière révision du texte. Enfin, je remercie les nombreux étudiants des programmes d'agronomie et de biologie qui ont suivi le cours avec moi durant les sessions d'hiver 2006, d'automne 2006 et d'automne 2007 et qui m'ont fait part de plusieurs commentaires constructifs.

Claude Bélisle
9 décembre 2009

Table des matières

1 La statistique descriptive **1**
- 1.1 Introduction . . . 1
- 1.2 Population, échantillon et variable statistique . . . 1
- 1.3 L'histogramme . . . 3
- 1.4 La moyenne et l'écart-type . . . 9
- 1.5 Les quantiles . . . 11
- 1.6 Le diagramme en boîte . . . 14
- 1.7 Diagramme en pointes de tarte et diagramme en bâtons . . . 16
- 1.8 Les données bivariées . . . 19
- 1.9 Exercices . . . 20

2 La théorie des probabilités **27**
- 2.1 Introduction . . . 27
- 2.2 Les principales propriétés des probabilités . . . 30
- 2.3 Notions d'analyse combinatoire . . . 34
- 2.4 ★ Probabilité conditionnelle et indépendance . . . 40
- 2.5 Variables aléatoires et distributions . . . 45
- 2.6 Quelques lois de probabilité classiques . . . 53
- 2.7 ★ Le cas multidimensionnel . . . 65
- 2.8 Exercices . . . 73

3 Estimation et tests d'hypothèses : problèmes à un échantillon **84**
- 3.1 Théorie générale de l'estimation . . . 84
- 3.2 L'estimation d'une moyenne . . . 86
- 3.3 L'estimation d'une variance . . . 87
- 3.4 L'estimation d'une proportion . . . 88
- 3.5 Intervalle de confiance pour une moyenne . . . 90
- 3.6 Intervalle de confiance pour une variance . . . 93
- 3.7 Intervalle de confiance pour une proportion . . . 97
- 3.8 Calcul de taille d'échantillon . . . 97
- 3.9 Tests d'hypothèses : un exemple illustratif . . . 99
- 3.10 Tests d'hypothèses : théorie générale . . . 100
- 3.11 Tests sur une moyenne . . . 104
- 3.12 ★ Tests sur une variance . . . 107
- 3.13 Tests sur une proportion . . . 110

3.14 Graphe quantile-quantile et test de Shapiro et Wilk . . . 114
3.15 Exercices . . . 120

4 Estimation et tests d'hypothèses : problèmes à deux échantillons 128
4.1 Introduction . . . 128
4.2 Comparaison de deux variances . . . 129
4.3 Comparaison de deux moyennes : le cas où $\sigma_1^2 = \sigma_2^2$ . . . 136
4.4 Comparaison de deux moyennes : le cas où $\sigma_1^2 \neq \sigma_2^2$ . . . 140
4.5 Comparaison de deux moyennes : le cas où les données sont appariées . . . 142
4.6 Comparaison de deux proportions . . . 145
4.7 ★ Le test de Wilcoxon-Mann-Whitney . . . 147
4.8 Exercices . . . 153

5 Introduction à l'analyse de la variance 160
5.1 Introduction . . . 160
5.2 L'estimation des paramètres du modèle . . . 161
5.3 Le test d'égalité des moyennes théoriques . . . 163
5.4 Une interprétation du rapport MST_R/MSE . . . 164
5.5 Un exemple illustratif . . . 166
5.6 Décomposition de la somme des carrés et table d'anova . . . 167
5.7 Un autre exemple illustratif . . . 169
5.8 Inférence pour une combinaison linéaire des moyennes . . . 171
5.9 ★ Description alternative du modèle d'anova . . . 172
5.10 L'hypothèse d'homogénéité des variances . . . 174
5.11 L'hypothèse de normalité . . . 178
5.12 Transformation des données . . . 178
5.13 ★ Le test de Kruskal et Wallis . . . 179
5.14 ★ L'anova à deux facteurs . . . 182
5.15 Exercices . . . 187

6 Introduction à la régression 194
6.1 Deux exemples illustratifs . . . 194
6.2 Le modèle classique de régression linéaire simple . . . 197
6.3 Estimation des paramètres du modèle . . . 198
6.4 Intervalle de confiance et tests d'hypothèses . . . 199
6.5 Retour à l'exemple 1 . . . 202
6.6 Retour à l'exemple 2 . . . 203
6.7 Décomposition de la somme des carrés et table d'anova pour la régression . . . 204
6.8 ★ Intervalle de prédiction . . . 208
6.9 Le coefficient de corrélation . . . 209
6.10 ★ Le lien entre r et MSE . . . 217
6.11 La vérification des hypothèses du modèle . . . 218
6.12 ★ La régression linéaire multiple . . . 220
6.13 Exercices . . . 220

7 Tableaux de fréquences et tests du khi-deux **226**
7.1 Test d'adéquation pour une variable discrète 226
7.2 Test d'homogénéité de I populations 229
7.3 Test d'indépendance de deux variables discrètes 234
7.4 ★ Quelques remarques 236
7.5 Exercices 240

A Tables de distributions **242**
A.1 La loi normale 242
A.2 La loi de Student 244
A.3 La loi du khi-deux 246
A.4 La loi de Fisher 248

Chapitre 1

La statistique descriptive

1.1 Introduction

Deux types de méthodes statistiques seront étudiées dans ce cours : les méthodes de la statistique descriptive et les méthodes de la statistique inférentielle .

La statistique descriptive

La *statistique descriptive*, c'est l'art de résumer un ensemble de données de façon à mettre en évidence l'information qu'elles contiennent. Le but du présent chapitre est de présenter les principaux concepts et les principales méthodes de la statistique descriptive.

La statistique inférentielle

La *statistique inférentielle*, c'est l'art de tirer des conclusions au sujet d'une population à partir d'un échantillon provenant de cette population. La statistique inférentielle fait appel à la théorie des probabilités. Un bref survol de la théorie des probabilités sera présenté au chapitre 2. Les chapitres subséquents porteront sur les principales méthodes de la statistique inférentielle.

1.2 Population, échantillon et variable statistique

Voici un exemple élémentaire pour illustrer les concepts de population, d'échantillon et de variable statistique.

Exemple 1. Un chercheur étudie une certaine espèce de serpents. Il s'intéresse en particulier au poids des serpents à la naissance. Il obtient les poids de 147 serpents nouveau-nés. Ces poids sont mesurés en grammes. Le Tableau 1 donne la liste de ces 147 poids.

Dans cet exemple, la *population* qui nous intéresse est l'ensemble de tous les serpents de l'espèce sous considération. L'*échantillon* est l'ensemble des 147 serpents obtenus par le chercheur. La *variable statistique* qui nous intéresse est la variable « poids à la naissance ». Parfois on s'intéressera à plusieurs variables. Dans le présent exemple on pourrait considérer la variable « longueur à la naissance » mesurée, par exemple, en centimètres.

On pourrait aussi considérer plusieurs populations. Par exemple, un chercheur pourrait vouloir comparer différentes espèces de serpents.

33.90	34.87	34.49	34.16	35.14	34.10	34.41
33.10	35.59	35.20	34.35	33.87	35.18	34.54
35.71	34.74	36.42	34.68	34.05	34.76	35.49
35.44	36.03	34.19	35.49	35.30	34.26	35.36
35.83	33.64	34.83	36.11	35.32	37.37	34.78
36.66	33.91	33.55	34.91	34.17	34.96	34.71
34.31	35.78	34.86	34.19	35.50	32.62	33.20
35.52	34.59	35.32	36.21	35.61	36.14	34.31
34.55	37.61	35.86	33.75	34.77	34.37	33.68
35.70	35.65	35.67	34.20	34.11	35.18	35.07
35.74	36.11	34.99	35.05	34.36	36.00	34.27
35.05	35.97	33.07	34.76	34.23	35.13	34.56
35.54	36.52	35.04	35.50	33.68	34.17	32.99
33.68	35.78	33.40	34.31	34.81	35.53	33.99
34.77	36.37	35.79	33.65	35.10	33.97	36.08
35.03	35.65	35.30	36.27	35.97	33.53	36.68
36.11	34.67	34.04	34.51	35.39	35.25	35.45
34.22	35.07	34.49	34.11	34.69	34.80	33.95
33.31	34.19	36.03	35.44	36.73	35.14	34.03
34.16	33.94	32.83	33.56	36.84	32.96	35.34
35.44	35.57	34.26	34.89	34.34	34.02	35.56

TABLEAU 1 : Les poids (en grammes) de 147 serpents nouveau-nés.

Dans les sections qui vont suivre, nous allons examiner les données du Tableau 1 et nous allons essayer de mettre en évidence l'information qu'elles contiennent.

Les différents types de variables statistiques :

La plupart des variables statistiques qu'on rencontre dans la pratique appartiennent à l'une ou l'autre des quatre catégories suivantes. Nous aurons l'occasion de travailler avec ces différents types de variables tout au long de ce cours.

- VARIABLE QUANTITATIVE CONTINUE. Il s'agit d'une variable à valeurs numériques. La variable peut prendre n'importe quelle valeur à l'intérieur d'un intervalle. Voici quelques exemples typiques de variables quantitatives continues : le poids d'un individu, le temps de germination d'une graine de semence, la durée de vie d'un animal, le volume d'une pomme, la hauteur d'un plant, etc.

- Variable quantitative discrète. Il s'agit d'une variable à valeurs numériques avec seulement quelques valeurs possibles. Dans la plupart des exemples qu'on rencontre en pratique, ces valeurs possibles sont des entiers positifs. Voici quelques exemples typiques de variables quantitatives discrètes : le nombre de louveteaux dans la portée d'une louve, le nombre de fruits sur un plant, le nombre d'ordinateurs dans une maison unifamiliale, etc.

- Variable qualitative ordinale. Il s'agit d'une variable possédant un nombre fini de valeurs possibles. Ces valeurs possibles sont non numériques mais elles sont dans un certain ordre naturel. Un exemple : le niveau de satisfaction par rapport à un certain produit pourrait être une variable avec valeurs possibles « pas du tout satisfait », « moyennement satisfait », « très satisfait ». Un autre exemple : une variable dont les valeurs possibles seraient « enfant », « adolescent », « adulte » et « personne âgée ».

- Variable qualitative nominale. Il s'agit d'une variable possédant un nombre fini de valeurs possibles. Ces valeurs possibles sont non numériques et elles ne sont pas ordonnées de façon naturelle. Voici quelques exemples typiques de variables qualitatives nominales : la nationalité d'un individu, le programme d'étude d'un étudiant de l'Université Laval, le sexe d'une personne, etc.

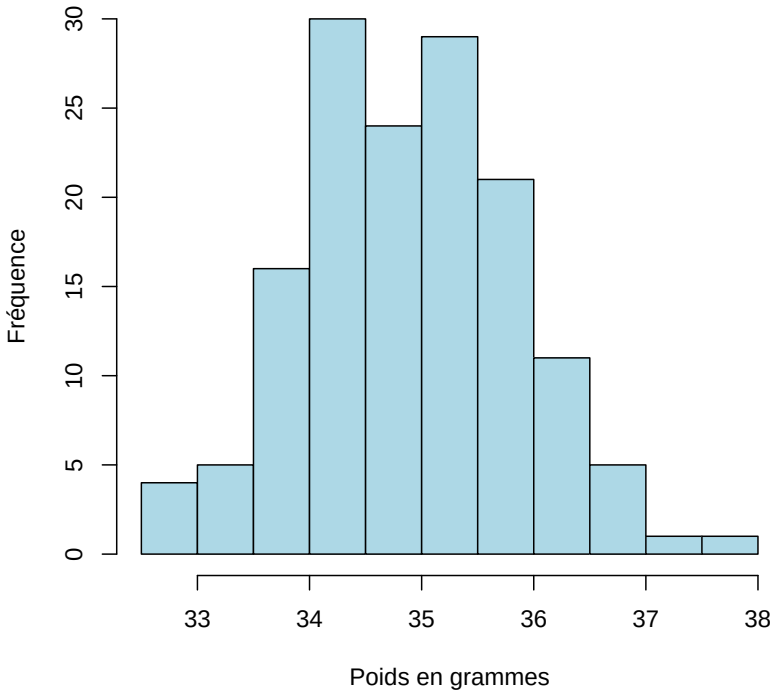
1.3 L'histogramme

L'histogramme est une représentation graphique qui nous permet de *voir* la forme de la distribution des données. On l'utilise surtout lorsqu'on est en présence d'une variable quantitative continue. Pour construire l'histogramme, on groupe d'abord les données en classes convenablement choisies. Voici un choix raisonnable de classes, ainsi que les fréquences correspondantes, pour l'exemple des serpents.

Poids	Fréquence
$(32.5, 33.0]$	4
$(33.0, 33.5]$	5
$(33.5, 34.0]$	16
$(34.0, 34.5]$	30
$(34.5, 35.0]$	24
$(35.0, 35.5]$	29
$(35.5, 36.0]$	21
$(36.0, 36.5]$	11
$(36.5, 37.0]$	5
$(37.0, 37.5]$	1
$(37.5, 38.0]$	1

Tableau 2 : Les poids (en g) de 147 serpents.

Les données présentées dans le Tableau 1 sont parfois appelées *les données brutes*. Le Tableau 2 contient quant à lui *les données groupées*. À partir des données groupées, on obtient l'histogramme que voici.

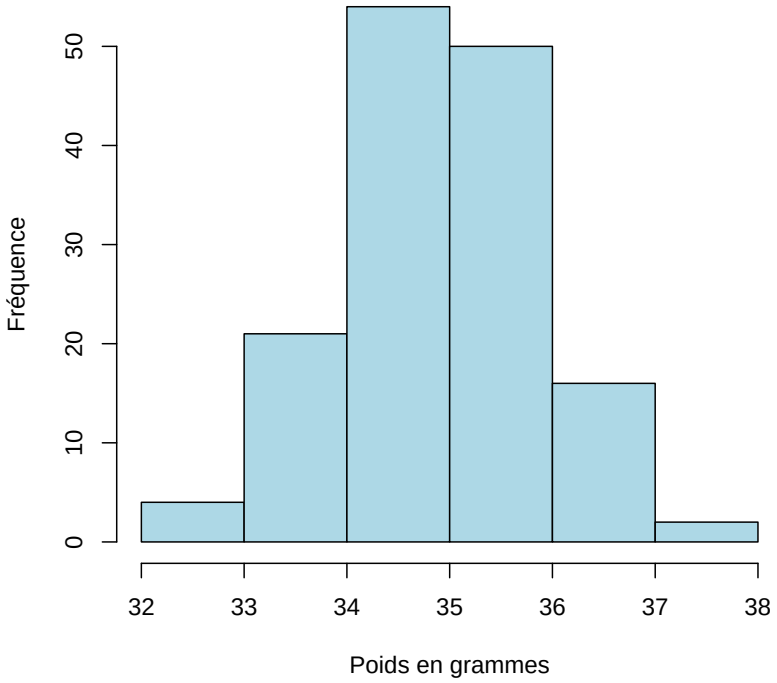


Il est important de noter que le choix des classes est un peu arbitraire. Avec un trop grand nombre de classes on risque d'obtenir un histogramme trop chaotique. Avec un trop petit nombre de classes on risque de ne pas voir la vraie forme de la distribution. Avec l'exemple des serpents, on aurait pu grouper les données de la façon suivante :

Poids	Fréquence
$(32.0, 33.0]$	4
$(33.0, 34.0]$	21
$(34.0, 35.0]$	54
$(35.0, 36.0]$	50
$(36.0, 37.0]$	16
$(37.0, 38.0]$	2

Tableau 3 : Les poids (en g) de 147 serpents ; seulement 6 classes.

On obtient alors l'histogramme suivant :



Avec le logiciel R :

Les données brutes de l'exemple précédent sont disponibles sur le site web du cours dans un fichier EXCEL appelé `serpents.xls`. Voici une méthode simple pour importer ces données dans R. On ouvre le fichier `serpents.xls` dans EXCEL, on sélectionne les 147 poids apparaissant dans la première colonne et on clique sur *copier*. Les données sont maintenant sur le presse-papier (le *clipboard*). Dans R, on tape la commande suivante :

```
donnees <- read.table("clipboard")
```

On vient ainsi de créer dans R une matrice de dimensions 147 par 1 appelée `donnees`. Pour nos analyses, il est plus simple de travailler avec un vecteur-ligne plutôt qu'avec une matrice à une colonne. On tape la commande

```
poids <- donnees[,1]
```

On vient ainsi de créer dans R un vecteur-ligne appelé `poids`. Ce vecteur-ligne contient nos 147 poids de serpents. Pour créer un histogramme des poids, l'étudiant devrait essayer successivement les commandes suivantes et examiner ce qui se produit :

```
hist(poids)
```

```
hist(poids, main="Histogramme des poids")
```

```
hist(poids, main="",xlab="Poids en grammes",ylab="Fréquence")
```

```
hist(poids, main="",xlab="Poids en grammes",ylab="",nclass=5)

hist(poids,main="",xlab="",ylab="",breaks=c(32,33.5,35,36.5,38))

hist(poids, main="",xlab="",ylab="", col="red")
```

Par défaut, R utilise un algorithme plutôt complexe pour déterminer le nombre de classes et les bornes (points de séparation) de ces classes. Le résultat est presque toujours satisfaisant. Avec l'option `nclass`, on suggère un nombre de classe et R essaie de tenir compte de notre suggestion. Avec l'option `breaks`, on force R à utiliser les classes qu'on veut.

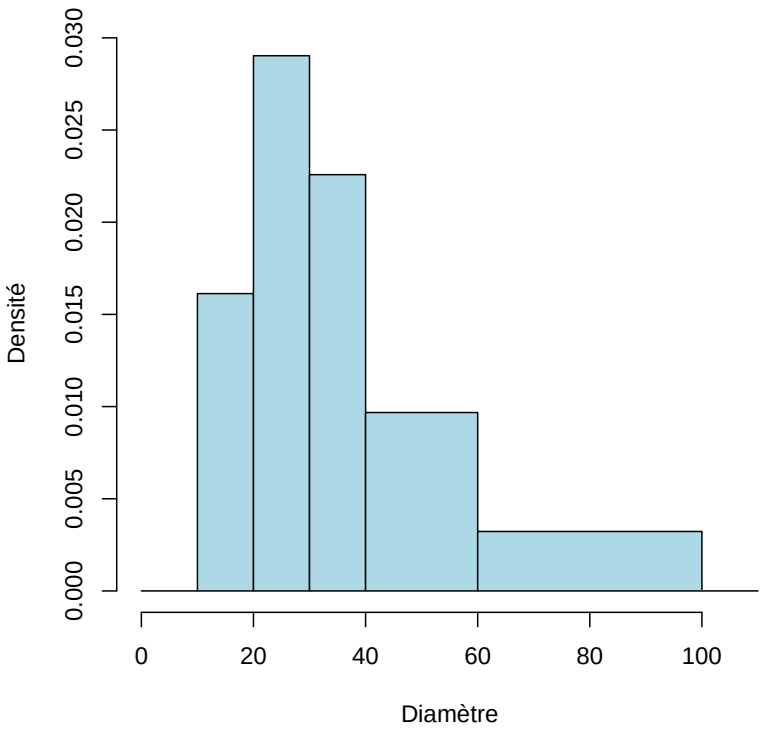
Histogramme avec classes inégales :

Dans l'exemple des serpents, les classes que nous avons utilisées étaient toutes de la même largeur. Il peut arriver que les observations nous soient présentées sous forme d'un tableau de données groupées dont les classes ne sont pas toutes de la même largeur. Lorsqu'on construit notre histogramme, il faut alors s'assurer que ce sont les surfaces (et non les hauteurs) des boîtes qui sont proportionnelles aux fréquences. Voici un exemple illustratif.

Exemple 2 : Chez Anti-Rouille Métropolitain, on a mesuré les diamètres de 155 gouttelettes d'huile émises par un appareil utilisé pour l'application d'une protection anti-rouille pour véhicules motorisés. Le tableau suivant résume nos observations :

Diamètre des gouttes	Fréquence observée
$(10, 20]$	25
$(20, 30]$	45
$(30, 40]$	35
$(40, 60]$	30
$(60, 100]$	20

Ces diamètres sont mesurés en microns. L'histogramme obtenu avec le logiciel R est reproduit à la page suivante. L'axe vertical de cet histogramme est différent de celui de l'histogramme des poids des serpents. Avec les serpents, les classes étaient toutes de la même largeur. Le logiciel R produisait alors un histogramme de fréquences absolues : sur l'axe vertical gradué on lisait les fréquences absolues des classes. Dans l'histogramme de la page 4, la hauteur de la boîte correspondant à la classe $(33.5, 34.0]$ est donc égale à 16 et c'est ce qu'on peut lire sur l'axe vertical gradué. Dans l'exemple des gouttelettes d'huile, les classes ne sont pas toutes de la même largeur. Le logiciel R produit alors un histogramme de type *densité* : les surfaces des boîtes sont égales aux fréquences relatives des différentes classes. Prenons par exemple la classe $(40, 60]$. Cette classe contient 30 observations. La fréquence relative de cette classe est donc $30/155 = 0.19355$. Dans l'histogramme, la surface de la boîte correspondant à la classe $(40, 60]$ est donc 0.19355. Comme la base de cette boîte rectangulaire mesure $60 - 40 = 20$, sa hauteur est $0.19355/20 = 0.0097$ et c'est ce qu'on peut lire sur l'axe vertical gradué.



Histogramme des fréquences absolues dessiné à la main :

Voici comment on dessine à la main un histogramme de fréquences absolues. Notre tableau des données groupées a l'allure suivante :

Classe	Fréquence
$(c_0, c_1]$	f_1
$(c_1, c_2]$	f_2
$\vdots$	$\vdots$
$(c_{\ell-1}, c_\ell]$	f_ℓ

On suppose ici que les ℓ classes sont toutes de même largeur, c'est-à-dire

$$c_1 - c_0 = c_2 - c_1 = c_3 - c_2 = \cdots = c_\ell - c_{\ell-1}.$$

On dessine l'histogramme de la façon suivante :

1. On trace un axe horizontal gradué.
2. Sur cet axe, on indique les valeurs $c_0, c_1, c_2, ..., c_\ell$.
3. Pour chaque $j \in \{1, 2, ..., \ell\}$, on dessine une boîte rectangulaire dont la base coïncide avec l'intervalle $(c_{j-1}, c_j]$ sur notre axe gradué et dont la hauteur est égale à la fréquence absolue f_j.

Histogramme de type « densité » dessiné à la main :

Pour ce type d'histogramme, il n'est pas nécessaire que les classes soient toutes de la même largeur. On construit d'abord le tableau suivant

Classe	Fréquence absolue	Fréquence relative	Hauteur de la boîte
$(c_0, c_1]$	f_1	$\frac{f_1}{n}$	$h_1 = \frac{f_1}{n(c_1-c_0)}$
$(c_1, c_2]$	f_2	$\frac{f_2}{n}$	$h_2 = \frac{f_2}{n(c_2-c_1)}$
$(c_2, c_3]$	f_3	$\frac{f_3}{n}$	$h_3 = \frac{f_3}{n(c_3-c_2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(c_{j-1}, c_j]$	f_j	$\frac{f_j}{n}$	$h_j = \frac{f_j}{n(c_j-c_{j-1})}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(c_{\ell-1}, c_\ell]$	f_ℓ	$\frac{f_\ell}{n}$	$h_\ell = \frac{f_\ell}{n(c_\ell-c_{\ell-1})}$

On dessine ensuite l'histogramme de la façon suivante :

1. On trace un axe horizontal gradué.
2. Sur cet axe, on indique les valeurs $c_0, c_1, c_2, ..., c_\ell$.
3. Pour chaque $j \in \{1, 2, ..., \ell\}$, on dessine une boîte rectangulaire dont la base coïncide avec l'intervalle $(c_{j-1}, c_j]$ sur notre axe gradué et dont la surface est égale à la fréquence relative f_j/n. Autrement dit, on dessine une boîte rectangulaire dont la base coïncide avec l'intervalle $(c_{j-1}, c_j]$ sur notre axe gradué et dont la hauteur est égale à h_j.

Avec le logiciel R :

Voici comment l'histogramme des diamètres des gouttelettes d'huile de la page 7 a été obtenu avec le logiciel R. Nous n'avons pas les données brutes, nous avons seulement le tableau des données groupées qui apparaît à la page 6. Dans R, on crée un vecteur de données brutes fictives qui respectent les fréquences de notre tableau de données groupées : 25 valeurs entre 10 et 20, 45 valeurs entre 20 et 30, etc. Pour y arriver, on peut utiliser la commande

```
goutte <- c(rep(15,25),rep(25,45),rep(35,35),rep(50,30),rep(80,20))
```

Nous avons ainsi créé un vecteur contenant 25 fois la valeur 15, 45 fois la valeur 25, 35 fois la valeur 35, 30 fois la valeur 50 et 20 fois la valeur 80. Le compte est bon. Nous avons en tout 155 valeurs et nous respectons les fréquences de notre tableau de données groupées. On obtiens ensuite l'histogramme de la page 7 avec la commande

```
hist(goutte, breaks=c(10,20,30,40,60,100), main="", xlab="Diamètre",
ylab="Densité", col="light blue")
```

1.4 La moyenne et l'écart-type

Considérons un échantillon aléatoire de taille n, disons $x_1, x_2, x_3, ..., x_n$. On suppose ici que la variable statistique d'intérêt est de type quantitative. La moyenne (échantillonnale) et l'écart-type (échantillonnal) sont donnés respectivement par

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1.2}$$

La raison pour le dénominateur $n-1$ plutôt que n dans la formule (1.2) sera expliquée à la section 3.3. Dans l'exemple des serpents, on obtient

$$\overline{x} = \frac{33.90 + 33.10 + 35.71 + \cdots + 35.34 + 35.56}{147} = 34.882 \text{ g}$$

$$s = \sqrt{\frac{(33.90 - 34.882)^2 + (33.10 - 34.882)^2 + \cdots + (35.56 - 34.882)^2}{146}} = 0.956 \text{ g}$$

On peut interpréter l'écart-type de la façon suivante. La moyenne des poids de nos 147 serpents est 34.882 g. Certains serpents ont un poids qui est très près de la moyenne. D'autres ont un poids qui est très loin de la moyenne. L'écart-type $s = 0.956$ g représente la distance typique entre les poids x_i et la moyenne des poids $\overline{x}$.

La moyenne $\overline{x}$ est un exemple de ce qu'on appelle une *mesure de tendance centrale*. C'est de loin la mesure de tendance centrale la plus utilisée. L'écart-type s est un exemple de ce qu'on appelle une *mesure de dispersion*. Malgré le fait que la formule (1.2) soit plutôt compliquée, l'écart-type est de loin la mesure de dispersion la plus utilisée. Une alternative conceptuellement plus simple est *l'écart absolu moyen* donné par la formule suivante :

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

Son interprétation est simple. L'écart absolu moyen d est simplement la moyenne des distances entre les observations x_i et la moyenne $\overline{x}$. En pratique, personne n'utilise l'écart absolu moyen d, tout le monde utilise l'écart-type s. La principale raison est que, contrairement à l'écart absolu moyen, l'écart-type possède de belles propriétés mathématiques. Nous y reviendrons aux chapitres 2 et 3. D'un point de vue pratique, il n'y a pas une grande différence entre les deux. Dans l'exemple des serpents, on obtient

$$d = \frac{|33.90 - 34.882| + |33.10 - 34.882| + \cdots + |35.56 - 34.882|}{147} = 0.883 \text{ g}.$$

Avec le logiciel R :

Pour calculer la moyenne (en anglais « *mean* ») et l'écart-type (en anglais « *standard deviation* »), on utilise respectivement les fonctions `mean( )` et `sd( )` du logiciel R. Pour l'exemple des serpents on tape `mean(poids)` et R nous répond 34.88177. Puis on tape `sd(poids)` et R nous répond 0.9560275.

Le coefficient de variation :

Dans le cas de variables à valeurs positives (poids, longueurs, durées de vie, etc.), le rapport $s/\overline{x}$ est souvent utilisé pour mesurer la *dispersion relative*. Ce rapport $s/\overline{x}$ est appelé le *coefficient de variation*. Il est parfois dénoté *CV*. Contrairement à l'écart-type, le coefficient de variation est un nombre sans unité. Il est souvent exprimé en pourcentage. Voici un exemple illustratif. On mesure les poids de 75 rats de laboratoire. On calcule la moyenne et l'écart-type et on obtient $\overline{x}_1 = 110$ grammes et $s_1 = 11$ grammes. On mesure ensuite les poids de 84 lapins, on calcule la moyenne et l'écart-type et on obtient $\overline{x}_2 = 660$ grammes et $s_2 = 22$ grammes. Si on compare les écarts-types s_1 et s_2, on conclut qu'il y a dans les poids des lapins deux fois plus de variation que dans les poids des rats. Mais un écart-type de 22 grammes est relativement petit considérant que la moyenne est 660 grammes alors qu'un écart-type de 11 grammes est relativement grand considérant que la moyenne est 110 grammes ! Plutôt que de comparer les écarts-types, il est probablement plus approprié de comparer les coefficients de variation. Ici on a $s_1/\overline{x}_1 = 11/110 = 0.100$, ou 10%, alors que $s_2/\overline{x}_2 = 22/660 = 0.033$, ou 3.3%. Bref, l'écart-type est plus petit chez les rats mais le coefficient de variation est plus petit chez les lapins.

Calcul de $\overline{x}$ et s à partir des données groupées

Pour pouvoir utiliser les formules (1.1) et (1.2), il faut avoir accès aux données brutes. Lorsqu'on dispose seulement des données groupées, il est possible de calculer des valeurs approximatives pour $\overline{x}$ et s, qu'on notera $\overline{x}_*$ et s_*. Il suffit de faire comme si les f_j observations de la classe $(c_{j-1}, c_j]$ étaient toutes égales à la valeur $m_j = (c_{j-1} + c_j)/2$, le milieu de la classe $(c_{j-1}, c_j]$. On obtient ainsi les approximations suivantes :

$$\overline{x} \approx \overline{x}_* = \frac{1}{n}\sum_{j=1}^{\ell}(f_j \times m_j)$$

$$s \approx s_* = \sqrt{\frac{1}{n-1}\sum_{j=1}^{\ell}(f_j \times (m_j - \overline{x}_*)^2)}$$

Pour l'exemple des gouttelettes d'huile, on obtient

$$\begin{aligned}\overline{x}_* &= \frac{1}{n}\sum_{j=1}^{\ell}(f_j \times m_j) = \frac{1}{155}\sum_{j=1}^{5}(f_j \times m_j) \\ &= \frac{(25\times 15)+(45\times 25)+(35\times 35)+(30\times 50)+(20\times 80)}{155} \\ &= 37.58 \text{ microns}\end{aligned}$$

$$
\begin{aligned}
s_* &= \sqrt{\frac{1}{n-1}\sum_{j=1}^{\ell}(f_j \times (m_j - \overline{x}_*)^2)} \\
&= \sqrt{\frac{1}{154}\sum_{j=1}^{5}(f_j \times (m_j - \overline{x}_*)^2)} \\
&= \sqrt{\frac{1}{154}\left\{(25 \times (15 - 37.58)^2) + \cdots + (20 \times (80 - 37.58)^2)\right\}} \\
&= 19.86 \text{ microns}
\end{aligned}
$$

1.5 Les quantiles

Comme à la section précédente, on considère un ensemble de n nombres qu'on note x_1, x_2, $x_3, ..., x_n$ et qu'on appelle *nos observations*, ou *nos données brutes*, ou *notre échantillon*. Fixons γ, un nombre entre 0 et 1. Grosso modo, le *quantile d'ordre* γ de notre échantillon est un nombre qu'on dénote q_γ et qui possède la propriété suivante : $100\gamma\%$ des observations sont plus petites que q_γ (et par conséquent $100(1-\gamma)\%$ des observations sont plus grandes que q_γ). Dans l'exemple des 147 poids de serpents nouveau-nés, le quantile d'ordre 0.4 est donc la valeur $q_{0.4}$ qui est telle que 40% de nos 147 poids lui sont inférieurs et 60% lui sont supérieurs. Le qualificatif *grosso modo* est important. Si on est en présence d'un échantillon de taille $n = 7$ et si nos sept observations sont

$$22.3 \qquad 17.9 \qquad 20.4 \qquad 24.6 \qquad 19.5 \qquad 26.2 \qquad 18.7$$

alors comment définit-on le quantile d'ordre 0.3 ? On aimerait que $q_{0.3}$ soit un nombre tel que 30% des observations lui sont inférieures (et 70% lui sont supérieures). Il y a plusieurs façons de procéder. Voici l'approche utilisée par le logiciel R. D'abord, on écrit $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$ pour dénoter nos observations placées en ordre croissant. Donc $x_{(1)}$ dénote la plus petite observation, $x_{(2)}$ dénote la deuxième plus petite observation, $x_{(3)}$ dénote la troisième plus petite observation, etc... et enfin $x_{(n)}$ dénote la plus grande observation. Dans notre petit exemple avec $n = 7$, on a

$$(x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7) = (22.3,\ 17.9,\ 20.4,\ 24.6,\ 19.5,\ 26.2,\ 18.7)$$

et on obtient

$$\big(x_{(1)},\ x_{(2)},\ x_{(3)},\ x_{(4)},\ x_{(5)},\ x_{(6)},\ x_{(7)}\big) = (17.9,\ 18.7,\ 19.5,\ 20.4,\ 22.3,\ 24.6,\ 26.2).$$

Le quantile échantillonnal d'ordre γ est défini de la façon suivante :

$$q_\gamma = x_{(1+(n-1)\gamma)}. \tag{1.3}$$

Dans notre petit exemple avec $n = 7$ et $\gamma = 0.3$, on obtient

$$q_{0.3} = x_{(1+(7-1)\times 0.3)} = x_{(2.8)}.$$

Petit problème : il faut préciser ce qu'on entend par $x_{(2.8)}$. La réponse est simple : $x_{(2.8)}$ est situé à 80% du chemin entre $x_{(2)}$ et $x_{(3)}$:

$$x_{(2.8)} = x_{(2)} + 0.8 \times (x_{(3)} - x_{(2)}).$$

On obtient donc

$$\begin{aligned} q_{0.3} &= x_{(1+(7-1)\times 0.3)} \\ &= x_{(2.8)} \\ &= x_{(2)} + 0.8 \times (x_{(3)} - x_{(2)}) \\ &= 18.7 + 0.8 \times (19.5 - 18.7) \\ &= 19.34 \end{aligned}$$

Dans l'exemple des serpents on a $n = 147$. Avec $\gamma = 0.4$ on obtient

$$\begin{aligned} q_{0.4} &= x_{(1+(147-1)\times 0.4)} \\ &= x_{(59.4)} \\ &= x_{(59)} + 0.4 \times (x_{(60)} - x_{(59)}) \\ &= 34.56 + 0.4 \times (34.59 - 34.56) \\ &= 34.572 \end{aligned}$$

Certains quantiles ont des noms particuliers et des notations particulières. Si k est un entier entre 0 et 100, le quantile d'ordre $k/100$ est appelé le k^e centile. Le terme *percentile* est synonyme de centile. Par exemple, le quantile d'ordre 0.17, qu'on note $q_{0.17}$, est aussi appelé le 17^e centile. Les quantiles d'ordre 0.25, 0.50 et 0.75, qu'on note respectivement $q_{0.25}$, $q_{0.50}$ et $q_{0.75}$, et qu'on appelle aussi le 25^e centile, le 50^e centile et le 75^e centile, sont également appelés le premier quartile, le deuxième quartile et le troisième quartile et sont souvent dénotés $Q1$, $Q2$ et $Q3$. Le deuxième quartile est aussi appelé la médiane et est souvent dénoté m. Les quartiles $Q1$, $Q2$ et $Q3$ seront utilisés dans la prochaine section. Examinons-les de plus près. Avec la formule (1.3), on obtient

$$\begin{aligned} Q1 &= x_{(1+(n-1)/4)} \\ Q2 &= x_{(1+(n-1)/2)} \\ Q3 &= x_{(1+3(n-1)/4)}. \end{aligned}$$

L'équation ci-dessus pour $Q2$ peut être réécrite sous la forme suivante :

$$Q2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ est impair} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ est pair.} \end{cases}$$

Bref, si n est impair, la médiane $Q2$ est la *valeur du milieu* alors que si n est pair la médiane $Q2$ est à mi-chemin entre les *deux valeurs du milieu*. Dans le petit exemple ci-dessus avec $n = 7$, on obtient

$$Q2 = x_{\left(\frac{n+1}{2}\right)} = x_{(4)} = 20.4$$

Le résumé à cinq nombres :

Le vecteur $(x_{(1)}, Q1, Q2, Q3, x_{(n)})$ est souvent appelé le *résumé à cinq nombres*. Rappelons que $x_{(1)}$ est simplement la plus petite observation et est souvent dénotée *min* (pour *minimum*). De même, $x_{(n)}$ est simplement la plus grande observation et est souvent dénotée *max* (pour *maximum*).

L'écart interquartile :

La distance entre le premier quartile $Q1$ et le troisième quartile $Q3$ est appelée *l'écart interquartile* et est souvent dénotée EIQ. On a donc $EIQ = Q3 - Q1$. Tout comme l'écart-type s, l'écart interquartile est une mesure de dispersion.

Les rangs centiles :

Imaginez un grand jeux de données, disons $x_1, x_2, ..., x_n$ avec n très grand. Les 99 centiles $q_{0.01}, q_{0.02}, q_{0.03}, ..., q_{0.99}$ découpent l'axe des x en 100 segments. Imaginez qu'on numérote ces 100 segments de 1 à 100 en partant de la gauche. Le *rang centile* d'une observation est le numéro du segment dans lequel se trouve cette observation. Il ne faut pas confondre *centile* et *rang centile*. Dire que le rang centile de l'observation x_j est égal à k, c'est dire que cette observation x_j est située quelque part entre le $(k-1)^e$ centile et le k^e centile. Dans l'exemple des serpents, le Tableau 1 à la page 2 nous indique que le premier serpent pesait 33.90 grammes. La formule (1.3) nous permet de voir que $q_{0.13} = 33.8994$ et $q_{0.14} = 33.9232$. Le rang centile de l'observation 33.90 est donc égal à 14.

Avec le logiciel R :

Dans R, on peut obtenir les quantiles avec la fonction `quantile`. Considérons d'abord le petit exemple de la page 11 avec $n = 7$. Avec la commande

```
yoyo <- c(22.3, 17.9, 20.4, 24.6, 19.5, 26.2, 18.7)
```

on crée le vecteur `yoyo`. Ce vecteur contient nos 7 observations. Pour obtenir le quantile d'ordre 0.3, aussi appelé le 30^e centile, on tape la commande

```
quantile(yoyo, 0.30)
```

et R nous retourne la valeur 19.34 conformément à ce qu'on a obtenu à la page 12 en faisant le calcul à la main. On peut obtenir plusieurs quantiles en une seule commande. Si on tape la commande

```
quantile(poids, c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9))
```

alors R nous retourne le petit tableau suivant :

10%	20%	30%	40%	50%	60%	70%	80%	90%
33.680	34.102	34.268	34.572	34.860	35.140	35.442	35.666	36.092

Pour obtenir le résumé à cinq nombres, il suffit de taper la commande

```
quantile(poids)
```

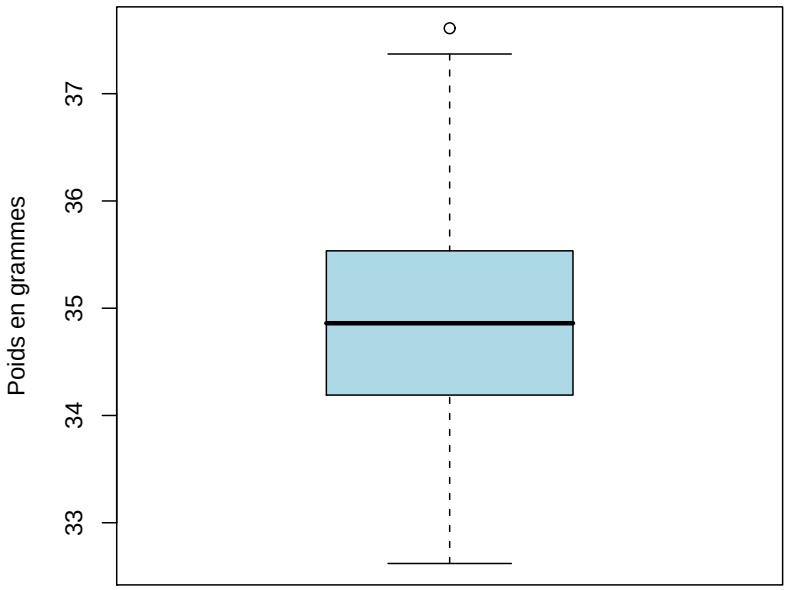
et R nous retourne le petit tableau suivant :

0%	25%	50%	75%	100%
32.620	34.190	34.860	35.535	37.610

La commande `quantile(poids)` nous donne donc le résumé à cinq chiffres. Elle est équivalente à la commande `quantile(poids, c(0, 0.25, 0.50, 0.75, 1))`.

1.6 Le diagramme en boîte

À quelques détails près, le diagramme en boîte (en anglais *boxplot*) est une simple représentation graphique du résumé à cinq nombres. Voici le diagramme en boîte pour l'exemple des serpents :



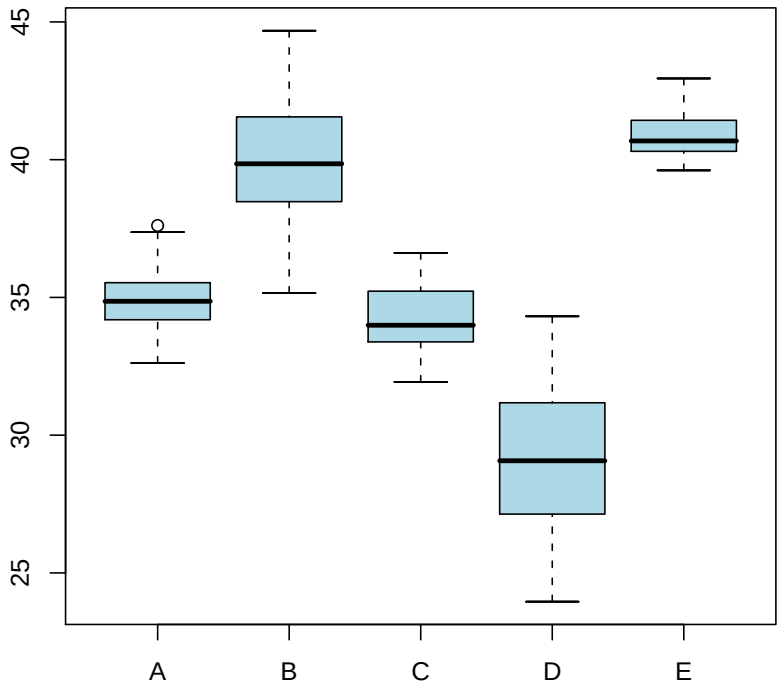
Le diagramme en boîte pour l'exemple des serpents

Ce diagramme en boîte a été construit de la façon suivante.

1. On trace un axe vertical gradué.
2. Sur cet axe vertical, on note la position des valeurs $x_{(1)}, Q1, Q2, Q3, x_{(n)}$.
3. À droite de l'axe vertical, on dessine une boîte rectangulaire de hauteur égale à l'écart EIQ, avec extrémité inférieure située à la hauteur $Q1$ et extrémité supérieure située à la hauteur $Q3$.
4. À travers cette boîte, on trace un segment horizontal à la hauteur $Q2$.
5. À partir du centre du côté inférieur de la boîte, on trace un segment vertical, appelé *bras inférieur*, qui s'étend de la boîte jusqu'en un point situé vis-à-vis la valeur u sur l'axe vertical. Ici u est la valeur de la plus petite observation non aberrante, c'est-à-dire le plus petit x_i satisfaisant $x_i \geq Q1 - (1.5)EIQ$.

6. À partir du centre du côté supérieur de la boîte, on trace un segment vertical, appelé *bras supérieur*, qui s'étend de la boîte jusqu'en un point situé vis-à-vis la valeur v sur l'axe horizontal. Ici v est la valeur de la plus grande observation non aberrante, c'est-à-dire le plus grand x_i satisfaisant $x_i \leq Q3 + (1.5)EIQ$.
7. Les observations aberrantes, s'il y en a, sont alors indiquées par des astérisques ou des points sur le prolongement imaginaire des bras. Une observation est dite *aberrante* si elle est ou bien plus petite que la valeur $Q1 - (1.5)EIQ$, ou bien plus grande que la valeur $Q3 + (1.5)EIQ$.

En général, l'histogramme est une meilleure représentation graphique des données que le diagramme en boîte. Cependant, le diagramme en boîte est un outil très utile lorsqu'on veut comparer plusieurs échantillons. Revenons à l'exemple des serpents. Imaginez qu'on veuille comparer les poids à la naissance pour 5 populations de serpents. Une façon simple de faire cette comparaison serait de tracer les diagrammes en boîte juxtaposés. On obtiendrait quelque chose ressemblant peut-être à ceci :



À l'aide d'un tel graphe on peut tirer plusieurs conclusions, dont les suivantes : ce sont dans les populations B et E que les poids sont les plus grands et dans la population D qu'ils sont les plus petits ; pour les populations B et D, les poids varient beaucoup plus que pour les trois autres populations.

Avec le logiciel R :

Le diagramme en boîte de la page 14 a été obtenu avec la commande suivante :

```
boxplot(poids, xlab="Le diagramme en boîte pour l'exemple des serpents",
        ylab="Poids en grammes", col="light blue")
```

Pour obtenir les diagrammes en boîte juxtaposés ci-dessus, nous avons utilisé la commande suivante :

```
boxplot(serA, serB, serC, serD, serE, names=c("A","B","C","D","E")
        col="light blue")
```

Dans cette commande, `serA` est le nom du vecteur contenant les poids à la naissance pour notre échantillon de serpents provenant de la population A. Il en est de même pour `serB`, `serC`, `serD` et `serE`.

Petit détail technique. La plupart des logiciels de statistique dessinent leurs diagrammes en boîte selon la méthode présentée dans les pages précédentes. Le logiciel R procède autrement. Plutôt que de tracer la limite inférieure de la boîte à la valeur $Q1$, il la trace à l'endroit situé à mi-chemin entre la plus grande observation inférieure à $Q1$ et la plus petite observation supérieure à $Q1$. De même, plutôt que de tracer la limite supérieure de la boîte à la valeur $Q3$, il la trace à l'endroit situé à mi-chemin entre la plus grande observation inférieure à $Q3$ et la plus petite observation supérieure à $Q3$. Dans l'exemple des serpents, les deux méthodes donnent le même diagramme en boîte. Avec les données 1, 2, 3, 7, 7, 7, 8, 9, 10, 20, 21, 24, les méthodes donnent des diagrammes en boîte différents. La boîte du diagramme usuel va de $Q1 = 6$ à $Q3 = 12.5$. La boîte du diagramme produit par R va de $(3 + 7)/2 = 5$ à $(10 + 20)/2 = 15$.

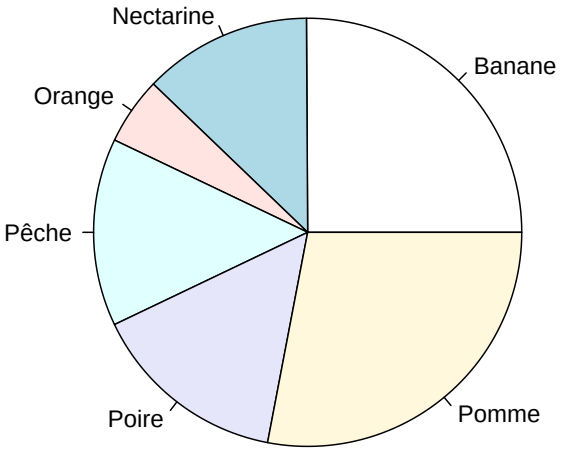
1.7 Diagramme en pointes de tarte et diagramme en bâtons

Les méthodes présentées jusqu'à maintenant sont des méthodes propres aux variables quantitatives. Pour les variables qualitatives, il existe des méthodes plus appropriées. Voici un exemple où la variable d'intérêt est une variable qualitative nominale.

Exemple 3. On a demandé à 275 jeunes écoliers quel était leur fruit préféré parmi les six fruits les plus consommés au Québec : banane, nectarine, orange, pêche, poire, pomme. Voici les résultats de ce petit sondage maison :

Fruit	Fréquence
Banane	69
Nectarine	35
Orange	14
Pêche	39
Poire	41
Pomme	77

On peut représenter graphiquement ces données à l'aide d'un diagramme en pointes de tarte (en anglais *pie chart*) :

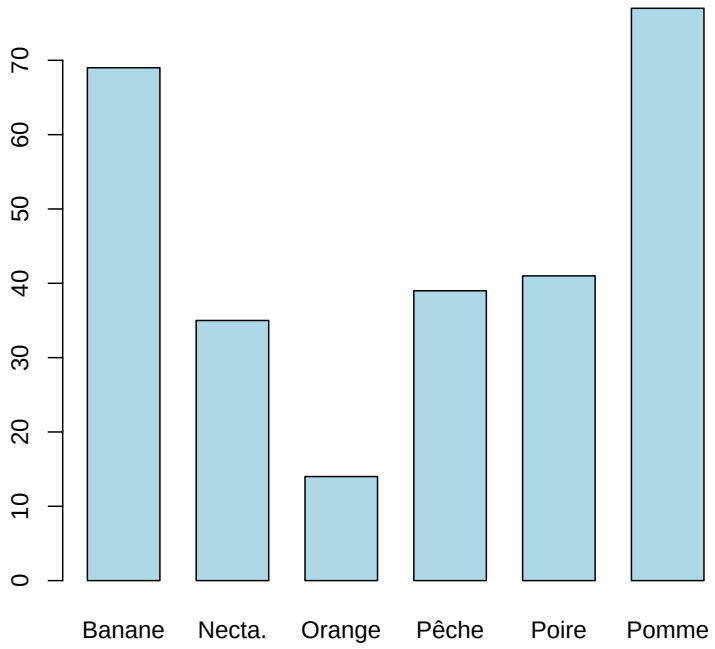


Dans ce diagramme en pointes de tarte, les surfaces des secteurs sont proportionnelles aux fréquences absolues données dans le tableau ci-dessus. Ceci est équivalent à dire que les longueurs des arcs de cercle sont proportionnelles aux fréquences absolues données dans le tableau ci-dessus.

Pour dessiner à la main un diagramme en pointes de tarte, il suffit de compléter le tableau suivant :

Fruit préféré	Fréquence absolue	Fréquence relative	Longueur de l'arc en degrés
Banane	69	$\frac{69}{275}$	$\frac{69}{275} \times 360 = 90.33$
Nectarine	35	$\frac{35}{275}$	$\frac{35}{275} \times 360 = 45.82$
Orange	14	$\frac{14}{275}$	$\frac{14}{275} \times 360 = 18.33$
Pêche	39	$\frac{39}{275}$	$\frac{39}{275} \times 360 = 51.05$
Poire	41	$\frac{41}{275}$	$\frac{41}{275} \times 360 = 53.67$
Pomme	77	$\frac{77}{275}$	$\frac{77}{275} \times 360 = 100.80$
Total	275	1	360

On peut aussi représenter les données de l'exemple 3 à l'aide d'un diagramme en bâtons :



Le diagramme en bâtons (en anglais *bar graph*) est tout simplement un histogramme adapté aux variables qualitatives. Sur l'axe vertical on peut lire les fréquences absolues. Bien que l'utilisation du diagramme en pointes de tarte soit très répandue, elle est habituellement déconseillée par les statisticiens. Selon des études menées par des psychologues, le diagramme en pointes de tarte ne transmet pas l'information aussi efficacement que le diagramme en bâtons. À vous de juger.

Avec le logiciel R :

Voici comment on obtient les diagrammes ci-dessus avec le logiciel R. On crée d'abord un vecteur de longueur 275 qu'on appelle `fruit` et qui contient 69 fois le mot « Banane », 35 fois le mot « Nectarine », etc.

```
fruit <- c(rep("Banane",69),rep("Nectarine",35),rep("Orange",14),
       rep("Pêche",39),rep("Poire",41),rep("Pomme",77))
```

On obtient ensuite notre diagramme en pointes de tarte et notre diagramme en bâtons à l'aide des commandes suivantes :

```
pie(table(fruit))
barplot(table(fruit),space=0.5, col="light blue")
```

La fonction `pie` choisit certaines couleurs par défaut. Il y a une option qui nous permet de choisir les couleurs qu'on veut. Dans la fonction `barplot`, on utilise l'option `space=0.5` pour dire à R que qu'on veut que l'espace entre bâtons adjacents soit 0.5 fois la largeur des bâtons. Dans R, tapez la commande `table(fruit)` et examinez le résultat.

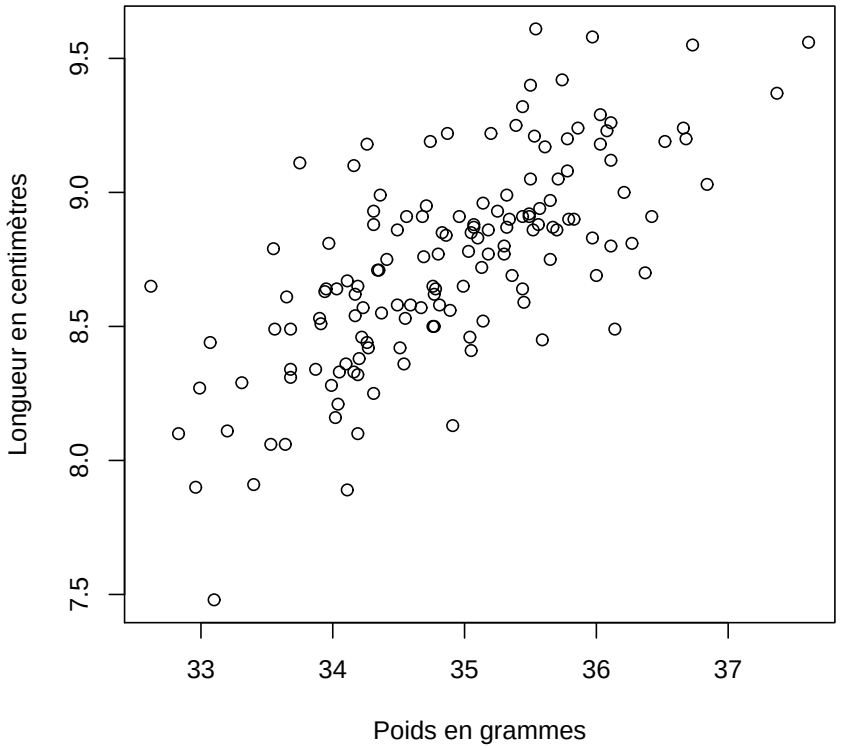
1.8 Les données bivariées[1]

Il arrive souvent qu'on veuille étudier deux ou plusieurs variables en même temps. Considérons brièvement le cas de deux variables. On considère donc une population avec deux variables d'intérêt qu'on appelle tout simplement la variable X et la variable Y. On obtient un échantillon aléatoire de taille n et, pour chacun des n individus de notre échantillon, on observe les deux variables. Nos données sont alors sous la forme de n couples qu'on note de la façon suivante :

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ...(x_n, y_n).$$

On écrit $\overline{x}$ et s_x pour la moyenne et l'écart-type de la variable X et on écrit $\overline{y}$ et s_y pour la moyenne et l'écart-type de la variable Y.

Revenons à l'exemple des serpents. Le fichier `Serpents.xls` disponible sur le site web du cours contient non seulement les poids des serpents mais aussi leurs longueurs. Intuitivement, on s'attend à ce qu'il y ait une association positive entre la variable poids et la variable longueur. Plus un serpent est pesant, plus on s'attend à ce qu'il soit long. Plus un serpent est léger, plus on s'attend à ce qu'il soit court. Pour visualiser cette association, on trace le graphe suivant, appelé graphe de dispersion ou diagramme de dispersion (en anglais *scatterplot*) :



[1]Cette section pourrait être étudiée plus tard, en même temps que le Chapitre 6.

Pour mesurer le degré d'association linéaire entre deux variables quantitatives, on utilise le coefficient de corrélation. Ce coefficient est dénoté r et est calculé de la façon suivante :

$$r \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right).$$

Dans l'exemple des serpents, on obtient $r = 0.707$. Le coefficient de corrélation est toujours un nombre entre -1 et 1. Il mesure le degré d'association linéaire entre les deux variables. Nous y reviendrons aux chapitres 2 et 6.

Avec le logiciel R :

Voici comment on obtient le diagramme de dispersion ci-dessus avec le logiciel R. À partir du fichier `serpents.xls`, on crée dans R le vecteur `poids` (contenant les 147 poids de serpents) et le vecteur `longueurs` (contenant les 147 longueurs de serpents). Pour y arriver, on ouvre d'abord le fichier `serpents.xls` dans EXCEL, puis on sélectionne les données (147 lignes, deux colonnes) et on copie le tout sur le presse-papier. Ensuite on tape les commandes suivantes dans R :

```
serpents <- read.table("clipboard")
poids <- serpents[,1]
longueurs <- serpents[,2]
```

L'objet `serpents` est alors une matrice de dimensions 147 par 2, c'est-à-dire une matrice avec 147 lignes et deux colonnes. La première colonne contient les 147 poids. La deuxième colonne contient les 147 longueurs. Avec la commande `poids <- serpents[,1]`, on dit à R de créer un vecteur contenant la première colonne de la matrice `serpents`. Avec la commande `longueurs <- serpents[,2]`, on dit à R de créer un vecteur contenant la deuxième colonne de la matrice `serpents`. Enfin, pour créer notre diagramme de dispersion, on tape la commande

```
plot(poids, longueurs, main="", xlab="Poids en grammes",
     ylab="Longueur en centimètres")
```

1.9 Exercices

Numéro 1. Déterminez la nature (quantitative continue, quantitative discrète, qualitative ordinale ou qualitative nominale) de chacune des variables suivantes.

(a) La moyenne cumulative d'un étudiant de l'Université Laval.

(b) La cote (A+, A, A-, B+,...) qu'un étudiant reçoit dans un cours.

(c) La section (A, B, S) du cours STT-10400 à laquelle un étudiant de génie électrique est inscrit.

(d) Le nombre de fois qu'un étudiant du baccalauréat en agronomie échoue le cours STT-19909 durant son baccalauréat.

(e) Le nombre de louvetaux dans la portée d'une louve.

(f) Le poids d'un loup à la naissance.

(g) La couleur (blanche, jaune, orange, verte, bleue, marron, noire) de la ceinture d'un adepte de karaté Shotokan.

(h) La couleur du feu de circulation en arrivant à l'intersection du Chemin Ste-Foy lorsque vous roulez sur la rue Myrand en direction nord.

(i) Le salaire annuel d'un professeur de cégep.

(j) Le nombre de tomates produites par un plant de tomate durant le mois d'août.

(k) La nature d'une variable statistique.

Numéro 2. On a mesuré les diamètres de 34 prunes. Voici les résultats, en centimètres.

4.03	4.05	3.96	4.09	4.28	4.04	4.18	4.23	4.14
4.12	4.03	3.94	4.02	4.08	4.13	4.04	3.93	4.08
4.37	4.07	4.11	4.03	4.00	3.97	4.01	4.09	4.06
3.92	4.19	3.96	4.48	4.24	4.06	3.98		

(a) Calculez la moyenne.

(b) Calculez l'écart-type.

(c) Calculez le coefficient de variation.

(d) Dessinez un histogramme approprié.

(e) Calculez le résumé à cinq nombres.

(f) Dessinez le diagramme en boîte.

(g) Qu'arrive-t-il à la moyenne si les valeurs 4.18, 4.23, 4.14 et 4.12 sont remplacées par les valeurs 4.68, 4.73, 4.64 et 4.62 ?

(h) Qu'arrive-t-il à la médiane si les valeurs 4.18, 4.23, 4.14 et 4.12 sont remplacées par les valeurs 4.68, 4.73, 4.64 et 4.62 ?

(i) Calculez le 30^e centile.

Numéro 3. Un biologiste a mesuré les longueurs de 310 perchaudes attrapées dans le Chenal du Moine en août 2006. Le tableau suivant résume ces 310 mesures :

Longueur	Nombre de perchaudes
entre 21.0 et 22.0 cm	50
entre 22.0 et 23.0 cm	90
entre 23.0 et 24.0 cm	70
entre 24.0 et 26.0 cm	60
entre 26.0 et 30.0 cm	40

(a) Représentez ces données à l'aide d'un histogramme.

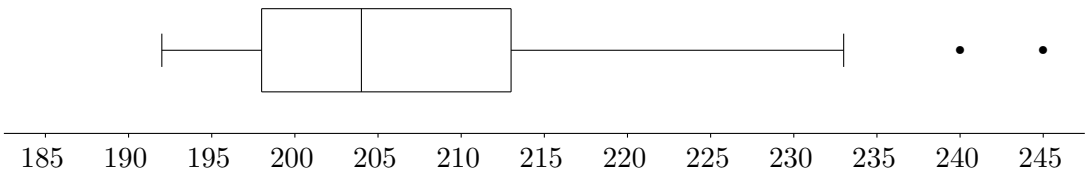
(b) Calculez une approximation pour la moyenne.

(c) Calculez une approximation pour l'écart-type.

(d) Calculez une approximation pour la médiane.

(e) Calculez une approximation pour la proportion dee perchaudes dont la longueur est située entre 25 et 27 cm.

NUMÉRO 4. Le diagramme en boîte suivant représente la distribution des poids (exprimés en livres) des 92 joueurs qui se sont présentés au camp d'entraînement de l'équipe de football du Rouge et Or en juin dernier.



Les énoncés suivants sont-ils vrais ou faux ?

(a) La somme des poids des 3 joueurs les plus pesants est inférieure à 720 livres.

(b) La moyenne de ces 92 poids est 204 livres.

(c) L'écart inter-quartile de ces 92 poids est de 15 livres.

(d) [Pour les étudiants qui sont familiers avec la loi normale] La loi normale est un bon modèle pour décrire cette distribution de poids.

NUMÉRO 5. Parmi les 116 étudiants inscrits au cours STT-20694 à l'automne 2006, 21 ont reçu un A, 36 ont reçu un B, 27 ont reçu un C, 18 ont reçu un D et 14 ont reçu un E.

(a) Représentez ces données à l'aide d'un diagramme en pointes de tarte.

(b) Représentez ces données à l'aide d'un diagramme en bâtons.

NUMÉRO 6. On a mesuré la variable $X =$ `longueur` (de la tête à la queue) et la variable $Y =$ `étendue` (du bout d'une aile à l'autre) de 12 oiseaux d'une certaine espèce. Voici les résultats en mm.

Longueur	156	154	153	153	155	163	157	155	164	158	158	160
Étendue	245	240	240	236	243	247	238	239	248	238	240	244

(a) Tracez le diagramme de dispersion.

(b) Calculez les moyennes $\overline{x}$ et $\overline{y}$.

(c) Calculez les écarts-types s_x et s_y.

(d) Calculez le coefficient de corrélation r.

(e) Sur le diagramme de dispersion, interprétez les résultats obtenus en (b), (c) et (d).

Numéro 7. On a mesuré les diamètres des troncs de 24 plants choisis au hasard dans un champ de blé d'Inde. Voici, en ordre croissant, nos 24 diamètres en millimètres :

10.5	10.9	11.4	11.8	12.1	12.1	12.5	12.7
12.8	13.1	13.3	13.4	13.7	13.9	14.1	14.3
14.5	14.8	15.3	15.4	15.7	16.2	16.4	17.9

(a) Calculez la moyenne et l'écart-type.

(b) Calculez le coefficient de variation.

(c) Dessinez un histogramme. C'est à vous de choisir des classes raisonnables.

(d) Obtenez le résumé à cinq nombres.

(e) Dessinez le diagramme en boîte.

(f) Calculez le quantile d'ordre 0.35.

Numéro 8. Quatre-vingt-dix jeunes pousses d'une certaine variété d'une espèce de plante ont été obtenues. On a mesuré la longueur de la plus grande feuille sur chacune des pousses. Le tableau suivant résume nos 90 observations :

Longueur de la plus grande feuille, en mm	Nombre de pousses
$(25.0, 26.0]$	12
$(26.0, 27.0]$	30
$(27.0, 28.0]$	20
$(28.0, 29.0]$	13
$(29.0, 30.0]$	5
$(30.0, 31.0]$	7
$(31.0, 32.0]$	3

(a) Dessinez l'histogramme.

(b) Obtenez l'approximation $\overline{x}_*$ de la moyenne échantillonnale.

(c) Obtenez l'approximation s_* de l'écart-type échantillonnal.

(d) Obtenez l'approximation $Q2_*$ de la médiane échantillonnale.

(e) Pour quelle pourcentage des pousses la longueur de la plus grande feuille excède-t-elle 29 mm ?

(f) Pour quelle pourcentage des pousses la longueur de la plus grande feuille est-elle inférieure à 26.5 mm ?

Numéro 9. On a mesuré les diamètres des troncs de 247 plants choisis au hasard dans un champ de blé d'Inde. Les données, exprimées en millimètres, se trouvent dans le fichier Excel `diametres-troncs.xls`, disponible sur le site web du cours.

(a) Obtenez la moyenne et l'écart-type.

(b) Obtenez le coefficient de variation.

(c) Obtenez un histogramme.

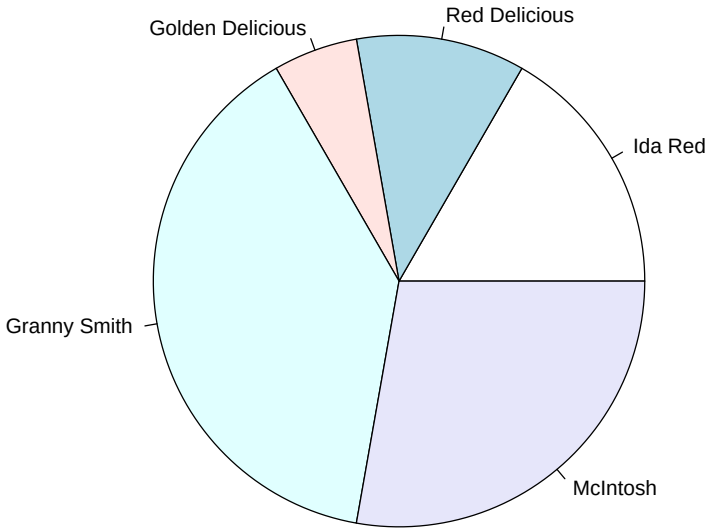
(d) Obtenez le résumé à cinq nombres.

(e) Obtenez le diagramme en boîte.

(f) Obtenez le 80^e centile.

(g) Quel est le rang centile du tronc de diamètre 13.8 mm ?

Numéro 10. Un verger compte 1512 pommiers. Le diagramme suivant représente les proportions de chacune des 5 espèces de pommiers de ce verger :



(a) Complétez le tableau suivant :

Espéce de pommes	Fréquence relative	Nombre de pommiers
Ida Red		
Red Delicious		
Golden Delicious		
Granny Smith		
McIntosh		

(b) Représentez ces données à l'aide d'un diagramme en bâtons.

Numéro 11. Lors d'une journée de pêche sur le lac St-Pierre, un groupe de pêcheurs ont attrapé 9 achigans, 22 barbottes, 13 brochets, 18 dorés et 47 perchaudes. Représentez ces données à l'aide d'un diagramme en pointes de tarte.

Numéro 12. Voici les longueurs, en cm, des 22 barbottes dont il est question au numéro précédent :

32.2	29.5	29.3	31.9	36.9	30.8	40.6	31.9	26.7	37.9	34.4
30.3	32.9	24.4	32.2	28.3	29.2	31.0	30.6	27.3	30.0	27.6

(a) Calculez la moyenne.

(b) Calculez l'écart-type.

(c) Calculez le coefficient de variation.

(d) Dessinez un histogramme approprié.

(e) Calculez le résumé à cinq nombres.

(f) Dessinez le diagramme en boîte.

Numéro 13. Voici les longueurs, en cm, et les poids, en gr, des 9 achigans dont il est question au numéro 11 :

Longueur	31.5	34.8	31.1	37.5	30.4	32.6	36.0	27.6	36.8
Poids	542	680	653	694	614	596	606	541	743

(a) Tracez le diagramme de dispersion.

(b) Calculez les moyennes $\overline{x}$ et $\overline{y}$.

(c) Calculez les écarts-types s_x et s_y.

(d) Calculez le coefficient de corrélation r.

(e) Sur le diagramme de dispersion, interprétez les résultats obtenus en (b), (c) et (d).

Numéro 14. En vous inspirant de votre domaine d'études,

(a) donnez trois exemples de variables quantitatives continues ;

(b) donnez trois exemples de variables quantitatives discrètes ;

(c) donnez trois exemples de variables qualitatives ordinales ;

(d) donnez trois exemples de variables qualitatives nominales.

Chapitre 2

La théorie des probabilités

2.1 Introduction

Le présent chapitre se veut un bref un survol des principales notions de la théorie des probabilités. Voici deux exemples élémentaires qui vont nous aider à comprendre les concepts d'expérience aléatoire, d'ensemble des résultats possibles, d'événement et de probabilité d'un événement.

Exemple 1. On lance une paire de dés. Quelle est la probabilité que la somme des deux dés soit égale à 9 ? Quelle est la probabilité d'obtenir au moins une fois la valeur 6 ?

Solution. Identifions les deux dés : le dé A et le dé B. Voici l'ensemble de tous les résultats possibles pour cette expérience aléatoire :

$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$
$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(5,5)$	$(5,6)$
$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$

Dans ce tableau, le couple (k,ℓ) représente le résultat « *obtenir la face k avec le dé A et la face ℓ avec le dé B* ». Dire qu'on obtient le résultat $(5,4)$ c'est dire qu'on obtient la face 5 avec le dé A et la face 4 avec le dé B. En supposant que les dés sont parfaitement symétriques, les 36 résultats énumérés ci-dessus ont tous la même chance de survenir. La probabilité d'obtenir le résultat $(5,4)$ est donc $1/36$. Parmi les 36 résultats possibles, il y en a 4 pour lesquels la somme des deux dés est égale à 9. Il s'agit des résultats $(6,3), (5,4), (4,5)$, et $(3,6)$. La probabilité que la somme des deux dés soit égale à 9 est donc $4/36$, c'est-à-dire $1/9$. De même, parmi les 36 résultats possibles, il y en a 11 pour lesquels la face 6 apparaît au moins une fois. La probabilité d'obtenir au moins une fois la valeur 6 est donc égale à $11/36$.

EXEMPLE 2. On lance une pièce de monnaie quatre fois. Quelle est la probabilité d'obtenir exactement deux faces et deux piles ?

SOLUTION. Voici l'ensemble de tous les résultats possibles pour l'expérience aléatoire qui consiste à lancer une pièce de monnaie 4 fois :

$FFFF$	$FFFP$	$FFPF$	$FFPP$
$FPFF$	$FPFP$	$FPPF$	$FPPP$
$PFFF$	$PFFP$	$PFPF$	$PFPP$
$PPFF$	$PPFP$	$PPPF$	$PPPP$

On utilise une convention semblable à celle utilisée à l'exemple 1. La notation $FPPF$ signifie « *face au premier lancer, pile au deuxième lancer, pile au troisième lancer et face au quatrième lancer* ». Si la pièce de monnaie est bien balancée, il est raisonnable de conclure que ces 16 résultats possibles ont tous la même probabilité. La probabilité d'obtenir le résultat $FPPF$ est donc $1/16$. Parmi les 16 résultats possibles, il y en a 6 qui donnent lieu à deux piles et deux faces. La probabilité d'obtenir deux piles et deux faces est donc égale à $6/16$, c'est-à-dire $3/8$.

Les principaux concepts.

Le point de départ est le concept d'*expérience aléatoire*. Une expérience aléatoire est une expérience ayant plusieurs *résultats possibles*. On ne peut pas prédire quel résultat surviendra mais on peut dresser la liste de tous les résultats possibles. On écrit $\mathcal{E}$ pour dénoter l'expérience aléatoire et on écrit Ω pour dénoter l'ensemble de tous les résultats possibles. Cet ensemble Ω est parfois appelé l'*ensemble fondamental* de l'expérience $\mathcal{E}$. Dans l'exemple 1 on a

$$\begin{aligned} \mathcal{E} &= \text{« On lance une paire de dés »}, \\ \Omega &= \{(1,1),(1,2),(1,3),...,(6,5),(6,6)\}. \end{aligned}$$

Dans l'exemple 2 on a

$$\begin{aligned} \mathcal{E} &= \text{« On lance une pièce de monnaie quatre fois »}, \\ \Omega &= \{FFFF, FFFP, FFPF, ..., PPPF, PPPP\}. \end{aligned}$$

Un *événement* est un sous-ensemble de l'ensemble de tous les résultats possibles. On utilise les lettres majuscules du début de l'alphabet pour dénoter des événements. Voici deux exemples d'événements relatifs à l'exemple 1 :

$$\begin{aligned} A &= \text{« Obtenir au moins un 6 »} \\ &= \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),(5,6),(4,6),(3,6),(2,6),(1,6)\} \end{aligned}$$

$$\begin{aligned} B &= \text{« Le total des deux dés est 8 »} \\ &= \{(6,2),(5,3),(4,4),(3,5),(2,6)\}. \end{aligned}$$

Voici trois exemples d'événements relatifs à l'exemple 2 :

$$\begin{aligned} C &= \text{« Obtenir une pile et trois faces »} \\ &= \{FFFP, FFPF, FPFF, PFFF\} \\ D &= \text{« Obtenir deux piles et deux faces »} \\ &= \{FFPP, FPFP, FPPF, PFFP, PFPF, PPFF\} \\ E &= \text{« Obtenir quatre résultats identiques »} \\ &= \{FFFF, PPPP\}. \end{aligned}$$

À chaque événement on associe sa *probabilité*. La probabilité d'un événement A, dénotée $\mathbb{P}[A]$, est un nombre entre 0 et 1 qui représente la fréquence relative à long terme avec laquelle l'événement A se réaliserait si on répétait l'expérience un très grand nombre de fois. Dans l'exemple 1, si on suppose que les dés sont bien balancés, alors les 36 résultats possibles ont tous la même probabilité. Il n'y a pas de raison pour que le résultat $(3, 5)$ soit plus probable ou moins probable que le résultat $(2, 1)$. Chacun des résultats possibles a donc une probabilité égale à $1/36$. Pour les événements A et B décrits ci-dessus, on obtient $\mathbb{P}[A] = 11/36$ et $\mathbb{P}[B] = 5/36$. Dans l'exemple 2, si on suppose que la pièce de monnaie est bien balancée, alors les 16 résultats possibles ont tous la même probabilité. Chacun des résultats possibles a donc une probabilité égale à $1/16$. Pour les événements C, D et E décrits ci-dessus, on obtient $\mathbb{P}[C] = 4/16 = 1/4$, $\mathbb{P}[D] = 6/16 = 3/8$ et $\mathbb{P}[E] = 2/16 = 1/8$.

Interprétation de la probabilité d'un événement

Dans le présent document, on s'intéresse seulement à l'aspect pratique de la théorie des probabilités. Il n'est pas question de s'attarder ici au sens philosophique de la notion de probabilité. Pour nous, la probabilité $\mathbb{P}[B] = 5/36$ obtenue dans l'exemple 1 est interprétée de la façon suivante : si on répète un très grand nombre de fois l'expérience aléatoire qui consiste à lancer une paire de dés bien balancés, alors on s'attend à ce que la somme des deux dés soit égale à 8 en moyenne 5 fois sur 36. Pour nous, la probabilité d'un événement est donc la fréquence relative avec laquelle cet événement se réaliserait si on répétait notre expérience aléatoire un très grand nombre de fois.

Le language ensembliste et diagramme de Venn

En théorie des probabilités on utilise souvent les diagrammes de Venn pour illustrer diverses notions. Dans un tel diagramme, une boîte rectangulaire représente l'ensemble Ω et les événements qui nous intéressent sont illustrés avec, par exemple, des disques à l'intérieur de cette boîte rectangulaire. Le diagramme de Venn de la Figure 1 représente l'ensemble Ω ainsi qu'un événement dénoté A. On imagine que Ω est l'ensemble de tous les résultats possibles d'une certaine expérience aléatoire $\mathcal{E}$. Lorsqu'on réalise cette expérience, un résultat survient, disons le résultat ω. Si ce ω appartient à l'ensemble A, on dit que l'événement A s'est réalisé. Si ce ω n'appartient pas à l'ensemble A, on dit que l'événement A ne s'est pas réalisé. La Figure 1 illustre le cas où l'événement A s'est réalisé. La notation A^c est utilisée pour dénoter le complément de l'ensemble A.

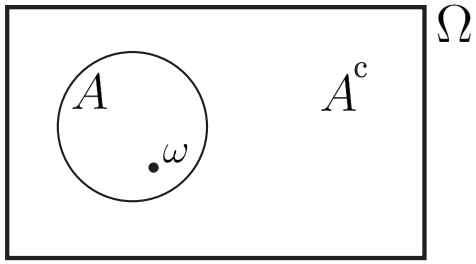


FIGURE 1. L'expérience a donné lieu au résultat ω.
Puisque $\omega \in A$, on dit que l'événement A s'est réalisé.

Si A et B sont des événements, on écrit $A \cup B$ pour dénoter l'union des ensembles A et B. Dire que l'événement $A \cup B$ se réalise c'est dire qu'au moins un des événements A et B se réalise. Le diagramme de Venn de la Figure 2 illustre l'événement $A \cup B$.

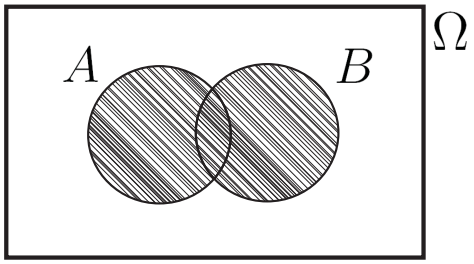


FIGURE 2. L'événement $A \cup B$.

Si A et B sont des événements, on écrit $A \cap B$ pour dénoter l'intersection des ensembles A et B. Dire que l'événement $A \cap B$ se réalise c'est dire que les événements A et B se sont tous les deux réalisés. Le diagramme de Venn de la Figure 3 illustre l'événement $A \cap B$.

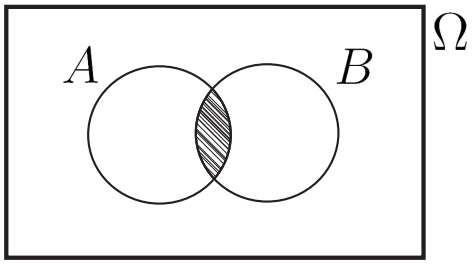


FIGURE 3. L'événement $A \cap B$.

2.2 Les principales propriétés des probabilités

Dans cette section, nous présentons les six principales propriétés des probabilités. Les trois premières propriétés sont en fait les trois axiomes sur lesquels toute la théorie mathématique des probabilités repose. On les appelle les axiomes de Kolmogorov. Les trois autres propriétés sont des conséquences de ces trois axiomes.

Propriété 1 : le premier axiome de Kolmogorov

Pour tout événement A, on a $\quad 0 \leq \mathbb{P}[A] \leq 1.$

Propriété 2 : le deuxième axiome de Kolmogorov

$$\mathbb{P}[\emptyset] = 0 \quad \text{et} \quad \mathbb{P}[\Omega] = 1.$$

Rappelons ici que le symbole $\emptyset$ est utilisé pour dénoter l'ensemble vide.

Propriété 3 : le troixième axiome de Kolmogorov

Si $A_1, A_2, ..., A_n$ sont des événements mutuellement exclusifs, alors

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \mathbb{P}[A_i]$$

c'est-à-dire

$$\mathbb{P}\left[A_1 \cup A_2 \cup \cdots \cup A_n\right] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + \cdots + \mathbb{P}[A_n].$$

Si $A_1, A_2, A_3, ...$ sont des événements mutuellement exclusifs, alors

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$$

c'est-à-dire

$$\mathbb{P}\left[A_1 \cup A_2 \cup A_3 \cup \cdots\right] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + \mathbb{P}[A_3] + \cdots$$

Pour le non mathématicien, les axiomes de Kolmogorov peuvent être vus comme étant une description mathématique de l'interprétation *fréquence relative à long terme* présentée à la section 2.1. Pour le mathématicien, les trois axiomes de Kolmogorov constituent une base à partir de laquelle toute la théorie des probabilité est construite.

Propriété 4 : le cas équiprobable

Si Ω est un ensemble fini et si les résultats possibles ont tous la même probabilité, alors pour tout événement A on a

$$\mathbb{P}[A] = \frac{\text{cardinal de } A}{\text{cardinal de } \Omega}.$$

Rappelons que le *cardinal* d'un ensemble fini est simplement le nombre d'éléments que contient cet ensemble. On a équiprobabilité dans chacun des deux exemples élémentaires présentés à la section 2.1. Voici deux exemples un peu moins élémentaires.

Exemple 3. On tire une main de poker, c'est-à-dire cinq cartes, au hasard à partir d'un jeu ordinaire de 52 cartes. Calculez la probabilité d'obtenir une *main pleine*, c'est-à-dire une main de poker comprenant une paire et un triple. Ici, Ω est l'ensemble de toutes les mains de poker possibles. Chaque main de poker a la même probabilité d'être obtenue. La probabilité désirée est donc, d'après la propriété 4,

$$\frac{\text{nombre total de mains pleines possibles}}{\text{nombre total de mains de poker possibles}} = \frac{3\,744}{2\,598\,960} \approx 0.00144.$$

Exemple 4. La 6/49 est sans doute la loterie la plus populaire au Canada. Pour 2\$, le joueur achète un billet de 6/49, c'est-à-dire une *combinaison* de six nombres différents choisis parmi les nombres 1 à 49. Lors du tirage, une combinaison est obtenue au hasard. Toutes les combinaisons ont la même chance d'être obtenues. Le montant d'argent que le joueur gagne dépend de plusieurs facteurs. Dans le cas particulier où la combinaison du joueur compte exactement trois des six nombres de la combinaison gagnante, le joueur gagne 10\$. Quelle est la probabilité que le joueur gagne 10\$? Nous sommes ici dans le cas équiprobable et la probabilité désirée est donc

$$\frac{\text{cardinal de } A}{\text{cardinal de } \Omega} = \frac{246\,820}{13\,983\,816} \approx 0.0177.$$

Ici Ω dénote l'ensemble de toutes les combinaisons possibles de 6 nombres choisis parmi les nombres 1 à 49 et A dénote l'ensemble de toutes les combinaisons qui ont exactement trois nombres en commun avec la combinaison gagnante.

Dans la prochaine section, nous étudierons certaines techniques de dénombrement qui nous permettrons d'arriver, avec beaucoup de facilité, aux réponses numériques des deux exemples précédents.

Propriété 5 : la complémentation

Pour tout événement A on a $\quad \mathbb{P}[A] = 1 - \mathbb{P}[A^c]$.

Rappelons à nouveau que si A est un sous-ensemble de Ω, alors le *complément* de A, dénoté A^c, est défini comme étant l'ensemble de tous les éléments qui appartiennent à Ω mais pas à A. Dans l'exemple 1, nous avons vu que si on lance une paire de dés alors la probabilité d'obtenir au moins une fois la valeur six est 11/36. Nous aurions pu arriver à ce résultat en utilisant la propriété de complémentation. Le complément de l'événement $A =$ « *obtenir au moins un six* » est l'événement $A^c =$ « *n'obtenir aucun six* ». On obtient donc

$$\mathbb{P}[\text{« } \textit{obtenir au moins un six} \text{ »}] = 1 - \mathbb{P}[\text{« } \textit{n'obtenir aucun six} \text{ »}] = 1 - \frac{25}{36} = \frac{11}{36}.$$

Dans cet exemple, on n'avait pas besoin de passer par la complémentation ; on examine le schéma présenté au début du présent chapitre et on note sans difficulté que le cardinal de A est 11. Pour apprécier l'utilité de la propriété de complémentation, il suffit de considérer un exemple où l'ensemble Ω est plus difficile à visualiser.

Exemple 5. On lance 8 dés biens balancés. Quelle est la probabilité d'obtenir au moins une fois la valeur 6 ? Écrivons Ω pour dénoter l'ensemble de tous les résultats possibles de cette expérience aléatoire et écrivons A pour dénoter l'événement « *obtenir au moins une fois la valeur 6* ». Le cardinal de Ω est $6^8 = 6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 = 1\,679\,616$. On peut voir ça de la façon suivante. Les résultats possibles de cette expérience aléatoire sont les vecteurs de la forme $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ avec $x_i \in \{1, 2, 3, 4, 5, 6\}$ pour chaque i. Il y a donc en tout 6^8 résultats possibles. Le cardinal de A^c est 5^8. On peut voir ça de la façon suivante. L'événement A^c comprend tous les vecteurs de la forme $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ avec $x_i \in \{1, 2, 3, 4, 5\}$ pour chaque i. Il y a donc en tout 5^8 résultats possibles qui appartiennent à A^c. On obtient donc

$$\mathbb{P}[A] \;=\; 1 - \mathbb{P}[A^c] \;=\; 1 - \frac{\text{cardinal de } A^c}{\text{cardinal de } \Omega} \;=\; 1 - \frac{5^8}{6^8} \;\approx\; 0.7674.$$

Propriété 6 : la formule de Poincaré

Pour tout événements A et B on a $\quad \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

Pour tout événements A, B et C on a

$$\begin{aligned} \mathbb{P}[A \cup B \cup C] \;&=\; \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C] \\ &\quad - \mathbb{P}[A \cap B] - \mathbb{P}[A \cap C] - \mathbb{P}[B \cap C] \\ &\quad + \mathbb{P}[A \cap B \cap C]. \end{aligned}$$

La Figure 4 ci-dessous nous aide à comprendre la formule de Poincaré dans le cas d'une union de deux événements. L'étudiant devrait essayer de faire un shéma analogue pour le cas d'une union de trois événements.

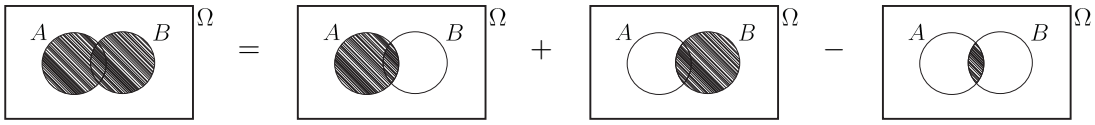


Figure 4. Illustration de l'équation $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

L'étudiant peut deviner la formule analogue pour la probabilité d'une union de quatre événements, de cinq événements, etc. L'identité de Poincaré est parfois appelée la formule d'inclusion-exclusion. On comprend pourquoi en examinant la Figure 4 et son analogue pour le cas à trois événements. Voici un exemple élémentaire pour illustrer la propriété de Poincaré.

Exemple 6. On lance un dé 3 fois. À l'aide de l'identité de Poincaré, calculez la probabilité d'obtenir au moins une fois la valeur 6. La probabilité désirée est $\mathbb{P}[A \cup B \cup C]$, avec

$$\begin{aligned} A &= \text{l'événement « obtenir un 6 au premier lancer »,} \\ B &= \text{l'événement « obtenir un 6 au deuxième lancer »,} \\ C &= \text{l'événement « obtenir un 6 au troisième lancer ».} \end{aligned}$$

On obtient

$$\begin{aligned} \mathbb{P}[A] = \mathbb{P}[B] = \mathbb{P}[C] &= 1/6, \\ \mathbb{P}[A \cap B] = \mathbb{P}[B \cap C] = \mathbb{P}[A \cap C] &= 1/36, \\ \mathbb{P}[A \cap B \cap C] &= 1/216 \end{aligned}$$

et l'identité de Poincaré nous donne

$$\mathbb{P}[A \cup B \cup C] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \frac{1}{36} - \frac{1}{36} - \frac{1}{36} + \frac{1}{216} = \frac{91}{216}.$$

On aurait pu obtenir le même résultat en utilisant la propriété de complémentation, comme à l'exemple 5 :

$$\mathbb{P}[\text{« au moins un 6 »}] = 1 - \mathbb{P}[\text{« aucun 6 »}] = 1 - \frac{5^3}{6^3} = 1 - \frac{125}{216} = \frac{91}{216}.$$

2.3 Notions d'analyse combinatoire

2.3.1 Introduction

Si l'ensemble Ω de tous les résultats possibles d'une expérience aléatoire $\mathcal{E}$ est un ensemble fini et si ces résultats possibles ont tous la même probabilité, alors on dit qu'on est dans le cas *équiprobable*. Nous avons vu à la section précédente que dans le cas équiprobable la probabilité d'un événement A est donnée par l'équation

$$\mathbb{P}[A] = \frac{\text{cardinal de } A}{\text{cardinal de } \Omega}.$$

Considérons par exemple l'expérience aléatoire qui consiste à choisir une main de poker au hasard, c'est-à-dire l'expérience aléatoire qui consiste à choisir 5 cartes au hasard à partir d'un jeu ordinaire de 52 cartes. Supposons qu'on veuille calculer la probabilité d'obtenir un brelan, c'est-à-dire une main de poker comprenant un triple (3 cartes de la même valeur) et 2 cartes de valeurs distinctes entre elles et distinctes de la valeur commune des 3 cartes du triple. Exemple de brelan : la main de poker qui comprend le 5 de pique, le 5 de coeur, le 5 de carreau, le valet de coeur et le 7 de carreau. Ici, Ω est l'ensemble de toutes les mains de poker possibles et l'événement A qui nous intéresse est l'ensemble de tous les brelans. Nous sommes dans le cas équiprobable et la probabilité désirée est donc

$$\mathbb{P}[A] = \frac{\text{cardinal de } A}{\text{cardinal de } \Omega} = \frac{\text{nombre total de brelans possibles}}{\text{nombre total de mains de poker possibles}}.$$

Pour calculer cette probabilité, il faut être capable de dénombrer l'ensemble de toutes les mains de poker possibles ainsi que l'ensemble de toutes les mains de poker qui sont des brelans. L'ensemble des techniques de dénombrement permettant de faire ce genre de calcul s'appelle l'*analyse combinatoire*. Dans la présente section, nous faisons un survol des principales notions d'analyse combinatoire.

2.3.2 Le principe fondamental du dénombrement

Imaginez une procédure comprenant k étapes. Ces k étapes doivent être réalisées dans un certain ordre chronologique. Supposons que les conditions suivantes sont satisfaites :

- Il y a n_1 façons différentes de réaliser la première étape.
- Peu importe la façon choisie pour réaliser la première étape, il y a n_2 façons différentes de réaliser la deuxième étape.
- Peu importe les façons choisies pour réaliser les deux premières étapes, il y a n_3 façons différentes de réaliser la troisième étape.

  $\vdots$
- Peu importe les façons choisies pour réaliser les $k-1$ premières étapes, il y a n_k façons différentes de réaliser la k^e étape.

Alors il y a en tout $n_1 \times n_2 \times \cdots \times n_k$ façons différentes de réaliser cette procédure. Ce résultat élémentaire s'appelle le *principe fondamental du dénombrement*. Nous avons utilisé ce principe de dénombrement à l'exemple 5. Voici quelques exemples additionnels.

Exemple 7. Autrefois les codes régionaux utilisés pour la téléphonie en Amérique du Nord étaient tous de la forme (a, b, c) avec $a \in \{2, 3, 4, 5, 6, 7, 8, 9\}$, $b \in \{0, 1\}$ et $c \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Combien de codes régionaux étaient alors possibles ?

Solution : $8 \times 2 \times 9 = 144$.

Exemple 8. Sur les plaques d'immatriculation des véhicules automobiles du Michigan, il y a 3 lettres suivies de 3 chiffres. Il y a en tout combien de numéros de plaque possibles au Michigan ? Il y a en tout combien de numéros de plaque comprenant 3 lettres différentes et 3 chiffres différents ? Si on choisit un numéro de plaque au hasard, quelle est la probabilité d'obtenir un numéro de plaque comprenant 3 lettres différentes et 3 chiffres différents ?

Solution : D'après le principe fondamental du dénombrement, il y a en tout $26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17\,576\,000$ numéros de plaque différents. Parmi ces numéros de plaque, il y en a $26 \times 25 \times 24 \times 10 \times 9 \times 8 = 11\,232\,000$ qui comprennent 3 lettres différentes et 3 chiffres différents. En supposant l'équiprobabilité, la probabilité d'obtenir un numéro de plaque à 3 lettres différentes et 3 chiffres différents est

$$\frac{11\,232\,000}{17\,576\,000} \approx 0.6391.$$

Exemple 9. Combien de mots à n lettres peut-on former avec un alphabet de ℓ lettres ?

Solution : Lorsqu'on écrit un tel mot, on a ℓ choix possibles pour la première lettre, ℓ choix possibles pour la deuxième lettre, etc. Il y a donc en tout ℓ^n mots possibles.

2.3.3 Permutations

Une *permutation* de n objets est un arrangement ordonné de ces n objets. L'exemples suivant clarifie cette définition.

EXEMPLE 10. Considérons les lettres A, B, C et D. Voici la liste de toutes les permutations possibles de ces quatre lettres :

$$\begin{array}{cccccc} ABCD & ABDC & ACBD & ACDB & ADBC & ADCB \\ BACD & BADC & BCAD & BCDA & BDAC & BDCA \\ CABD & CADB & CBAD & CBDA & CDAB & CDBA \\ DABC & DACB & DBAC & DBCA & DCAB & DCBA \end{array}$$

On note qu'il y a en tout 24 permutations possibles des lettres A, B, C et D. Nous aurions pu déterminer ce nombre en observant que lorsqu'on écrit une permutation des lettres A, B, C et D, on a 4 choix possibles pour la première lettre, puis 3 choix possibles pour la deuxième lettre, ensuite 2 choix possibles pour la troisième lettre et enfin un seul choix possible pour la dernière lettre. On a donc en tout $4 \times 3 \times 2 \times 1 = 24$ permutations possibles.

Plus généralement, le principe fondamental du dénombrement nous dit qu'il y a en tout

$$n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

permutations possibles de n objets.

LA NOTATION FACTORIELLE : Le produit $n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$ est dénoté $n!$ et est appelé « *n factoriel* ». On a donc

$$n \textit{ factoriel } = \ n! \ = \ n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1.$$

Par convention, on pose $0! = 1$. Nous verrons pourquoi à la page suivante.

EXEMPLE 11. Sur le bureau de Michel, il y a 12 livres : 4 livres de mathématiques, 3 livres de chimie, 3 livres de physique et 2 livres de biologie. Il faut ranger ces livres sur une tablette.

(a) De combien de façons différentes peut-on ranger ces 12 livres sur la tablette ?

(b) Si les livres sont rangés au hasard, quelle est la probabilité que les livres se retrouvent groupés par matière ?

SOLUTION : Il y a $12! = 479\,001\,600$ façons de ranger les livres. Si les livres sont rangés au hasard, la probabilité qu'ils se retrouvent groupés par matière est

$$\frac{4!\,4!\,3!\,3!\,2!}{12!} = \frac{41\,472}{479\,001\,600} = 0.0000866$$

Le premier facteur au numérateur représente le nombre de permutations des 4 matières alors que les facteurs suivants représentent, pour chaque matière, le nombre de permutations des livres de cette matière.

Il arrive qu'on s'intéresse aux arrangements ordonnés de r objets choisis parmi un groupe de n objets. Ces arrangements s'appellent des *permutations de r objets choisis parmi un groupe de n objets.*

Exemple 12. Considérons les lettres A, B, C et D. Voici la liste de toutes les permutations possibles de deux lettres choisies parmi ces quatre lettres

$$\begin{array}{cccccc} AB & AC & AD & BC & BD & CD \\ BA & CA & DA & CB & DB & DC \end{array}$$

Il y a donc 12 permutations possibles de 2 lettres choisies parmi les lettres A, B, C et D. Il fallait s'y attendre : lorsqu'on écrit une permutation de 2 lettres choisies parmi les lettres A, B, C et D, on a 4 choix pour la première lettre, puis 3 choix pour la deuxième lettre, donc en tout $4 \times 3 = 12$ permutations possibles. Notez que l'ordre compte : AB et BA sont deux permutations différentes. Plus généralement, le nombre total de permutations de r objets choisis parmi un groupe de n objets est donné par

$$n \times (n-1) \times (n-2) \times \cdots \times (n-(r-1)).$$

La notation factorielle nous permet de simplifier cette expression :

$$\begin{aligned} & n \times (n-1) \times \cdots \times (n-(r-1)) \\ & \quad = \quad n \times (n-1) \times \cdots \times (n-(r-1)) \times \frac{(n-r)!}{(n-r)!} \\ & \quad = \quad \frac{n \times (n-1) \times \cdots \times (n-(r-1)) \times (n-r) \times \cdots \times 3 \times 2 \times 1}{(n-r)!} \\ & \quad = \quad \frac{n!}{(n-r)!} \end{aligned}$$

Notation : Le nombre de permutations de r objets choisis parmi un groupe de n objets est parfois dénoté $P_{n,r}$. On a donc

$$P_{n,r} = \frac{n!}{(n-r)!} \tag{2.1}$$

Le lecteur peut maintenant apprécier le choix de la définition $0! = 1$. En effet, $P_{n,n}$ étant le nombre de permutations de n objets pris n à la fois, il faut avoir $P_{n,n} = n!$. Donc, pour que l'équation (2.1) soit valide avec $r = n$, il faut que $0!$ soit défini comme étant égal à 1.

Exemple 13. À la finale du 100 mètres, il y a 8 coureurs qui s'affrontent pour les médailles d'or, d'argent et de bronze. De combien de façons différentes ces médailles peuvent-elles être attribuées ?

Réponse : $P_{8,3} = 8!/(8-3)! = 8!/5! = 336$.

2.3.4 Combinaisons

Considérons, comme à la section précédente, un groupe de n objets. Un ensemble de r objets choisis parmi ces n objets s'appelle une *combinaison de r parmi n*. Contrairement au cas des permutations, ici on ne tient pas compte de l'ordre. En langage ensembliste, une combinaison de r parmi n est simplement un sous-ensemble de cardinal r obtenu à partir d'un ensemble de cardinal n.

EXEMPLE 14. Considérons les lettres A, B, C et D. Voici la liste de toutes les combinaisons possibles de 2 lettres choisies parmi ces 4 lettres :

$$\{A,B\} \quad \{A,C\} \quad \{A,D\} \quad \{B,C\} \quad \{B,D\} \quad \{C,D\}$$

L'ordre ne compte pas. Ainsi, $\{A,B\}$ et $\{B,A\}$ dénotent la même combinaison.

NOTATION : Le nombre de combinaisons possibles de r objets choisis parmi un groupe de n objets est dénoté $C_{n,r}$.

L'exemple 12 nous montre que $P_{4,2} = 12$ et l'exemple 14 nous montre que $C_{4,2} = 6$. On a donc $P_{4,2} = 2 \times C_{4,2}$. On aurait pu obtenir ce résultat en notant qu'à chaque combinaison de 2 parmi 4 correspondent 2 permutations de 2 parmi 4. Par exemple, à la combinaison $\{B,C\}$ correspondent les permutations BC et CB. Il y a donc, dans cet exemple, deux fois plus de permutations que de combinaisons. Plus généralement, si on considère les combinaisons de r objets parmi n, on note qu'à chaque combinaison correspondent $r!$ permutations. Le nombre de permutations de r parmi n est donc $r!$ fois plus grand que le nombre de combinaisons de r parmi n. On a donc

$$P_{n,r} = r!\, C_{n,r}.$$

Puisque $P_{n,r} = n!/(n-r)!$, on conclut que

$$C_{n,r} = \frac{n!}{r!\,(n-r)!}. \tag{2.2}$$

Avec $r = 0$, l'équation (2.2) nous donne $C_{n,0} = 1$. Ceci correspond au fait qu'il y a une seule façon de choisir zéro objet parmi un groupe de n objets : on n'en prend aucun ! Avec $r = 1$, l'équation (2.2) nous donne $C_{n,1} = n$. Ceci correspond au fait qu'il y a n façons différentes de choisir un objet parmi un groupe de n objets : on peut prendre l'objet numéro 1, ou bien l'objet numéro 2, ou bien l'objet numéro 3, etc.

EXEMPLE 15. Le nombre total de combinaisons possibles à la 6/49 est

$$C_{49,6} = \frac{49!}{6!\,(49-6)!} = \frac{49!}{6!\,43!} = 13\,983\,816.$$

Donc, lorsqu'on achète un billet de 6/49, la probabilité de gagner le gros lot est $\frac{1}{13\,983\,816}$. Le nombre 13 983 816 correspond à peu près au nombre de pièces de dix sous qu'il faut pour couvrir un terrain de football.

EXEMPLE 16. Combien de mots à huit lettres peut-on écrire si on doit utiliser trois fois la lettre A et cinq fois la lettre B ? Il y en a autant qu'il y a de façons de choisir, parmi les huit positions des lettres, les trois positions où on va placer la lettre A. Il y a donc

$$C_{8,3} = \frac{8!}{3!\,5!} = 56$$

mots à huit lettres comprenant trois fois la lettre A et 5 fois la lettre B.

Coefficient binomiaux : Le nombre de combinaisons $C_{n,r}$ est aussi appelé le r^e coefficient binomial d'ordre n et est souvent dénoté $\binom{n}{r}$. On a donc

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

Retour à l'exemple 3. Le nombre total de mains de poker différentes est simplement le nombre de façons différentes de choisir 5 cartes parmi 52 cartes :

$$\binom{52}{5} = \frac{52!}{5! \times (52-5)!} = \frac{52!}{5! \times 47!} = 2\,598\,960.$$

Pour *construire* une main pleine, on peut procéder de la façon suivante :

1. On choisit la valeur qu'on utilise pour le triple. Il y a $\binom{13}{1} = 13$ façons.
2. On choisit les trois couleurs qu'on utilise pour le triple. Il y a $\binom{4}{3} = 4$ façons.
3. On choisit la valeur qu'on utilise pour le double. Il y a $\binom{12}{1} = 12$ façons.
4. On choisit les deux couleurs qu'on utilise pour le double. Il y a $\binom{4}{2} = 6$ façons.

Le nombre de mains pleines est donc

$$\binom{13}{1} \times \binom{4}{3} \times \binom{12}{1} \times \binom{4}{2} = 13 \times 4 \times 12 \times 6 = 3744.$$

La probabilité désirée est donc

$$\frac{\text{nombre total de mains pleines possibles}}{\text{nombre total de mains de poker possibles}} = \frac{3\,744}{2\,598\,960} \approx 0.00144.$$

Retour à l'exemple 4. Le nombre total de combinaisons à la lotto 6/49 est

$$\binom{49}{6} = \frac{49!}{6! \times (49-6)!} = \frac{49!}{6! \times 43!} = 13\,983\,816.$$

Pour construire une combinaison ayant exactement trois nombres en commun avec la combinaison gagnante, il suffit de choisir trois nombres parmi les six nombres de la combinaison gagnante et trois nombres parmi les 43 nombres n'appartenant pas à la combinaison gagnante. Le nombre de combinaisons ayant exactement trois nombres en commun avec la combinaison gagnante est donc

$$\binom{6}{3} \times \binom{43}{3} = 20 \times 12\,341 = 246\,820$$

La probabilité désirée est donc

$$\frac{\text{nombre de combinaisons avec trois bons nombres}}{\text{nombre total de combinaisons}} = \frac{246\,820}{13\,983\,816} \approx 0.0177.$$

2.3.5 Avec le logiciel R

Avec le logiciel R, on peut calculer $n!$ grâce à la fonction `factorial`. Par exemple, si on tape `factorial(4)`, le logiciel R nous retourne la valeur 24. Si on tape `factorial(12)`, le logiciel R nous retourne la valeur 479 001 900. Si `a, b, c` et `d` sont des entiers positifs et si on tape `factorial(c(a, b, c, d))`, le logiciel R nous retourne un vecteur contenant les nombres `a!, b!, c!` et `d!`. Par exemple, si on tape `factorial(c(0,1,2,3,4,5))`, le logiciel R nous retourne le vecteur $(1, 1, 2, 6, 24, 120)$.

On peut aussi calculer des coefficients binomiaux grâce à la commande `choose(n,k)`. Si on tape `choose(6,2)`, le logiciel R nous retourne la valeur 15. Si on tape la commande `choose(6, 0 : 6)`, le logiciel R nous retourne le vecteur `(1, 6, 15, 20, 15, 6, 1)`. La notation $0 : 6$ est une façon abrégée d'écrire `c(0, 1, 2, 3, 4, 5, 6)`.

Le logiciel R nous permet aussi de faire des simulations de tirages au hasard grâce à la fonction `sample`. Pour tirer au hasard une combinaison de la lotto 6/49, on tape `sample(1 : 49, 6)`. Le logicial R fait alors 6 tirages sans remise à partir de l'ensemble $\{1, 2, 3, ..., 49\}$. Pour faire des tirages avec remise, on utilise l'option `replace = T`. Par exemple, pour simuler 6000 lancers d'un dé bien balancé, on tape `sample(1 : 6, 6000, replace=T)`. Le logiciel R simule alors 6000 lancers d'un dé bien balancé. Comme le dé est bien balancé, on s'attend à ce que chacune des 6 faces du dé surviennent à peu près environ 1000 fois. On peut en faire la vérification à l'aide des commandes suivantes :

```
x <- sample(1 : 6, 6000, replace=T)
```

```
table(x)
```

La première commande produit un vecteur contenant les résultats de 6000 tirages avec remise à partir de l'ensemble $\{1, 2, 3, 4, 5, 6\}$. La deuxième commande produit, par exemple, le tableau suivant :

1	2	3	4	5	6
975	1009	997	1020	1021	978

2.4 Probabilité conditionnelle et indépendance[1]

2.4.1 Probabilité conditionnelle

La notion de probabilité conditionnelle est une des notions les plus importantes en théorie des probabilités. Avant de l'introduire de façon formelle, considérons un exemple illustratif. On lance à deux reprises un dé bien balancé. Quelle est la probabilité d'obtenir au moins une fois la valeur six ? Pour résoudre ce problème élémentaire, il suffit d'observer que

- l'expérience aléatoire « lancer un dé deux fois » donne lieu à 36 résultats possibles ;
- ces 36 résultats possibles sont *équiprobables* : ils ont tous la même probabilité ;
- parmi ces 36 résultats, il y en a 11 pour lesquels on obtient au moins un six.

[1] On peut omettre cette section si on manque de temps

La réponse est donc 11/36. Maintenant, on lance le dé deux fois et on nous annonce que la somme des deux lancers est 8. Étant donnée cette information, quelle est la probabilité d'avoir obtenu au moins un six ? Pour résoudre ce problème, on note d'abord qu'il y a 5 résultats possibles pour lesquels la somme des deux lancers est 8 :

$$(2,6),(3,5),(4,4),(5,3),(6,2).$$

Parmi ces 5 résultats possibles, il y en a 2 pour lesquels il y a un six : $(2,6)$ et $(6,2)$. La réponse est donc 2/5. Examinons notre démarche d'un peu plus près. Posons

$$\begin{aligned} A &= \text{l'événement « obtenir au moins un six »} \\ &= \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6),(6,5),(6,4),(6,3),(6,2),(6,1)\}, \\ B &= \text{l'événement « la somme des deux lancers est 8 »} \\ &= \{(2,6),(3,5),(4,4),(5,3),(6,2)\}, \end{aligned}$$

et notons que

$$A \cap B = \{(2,6),(6,2)\}.$$

Notre réponse 2/5 peut donc s'écrire sous la forme

$$\frac{2}{5} = \frac{2/36}{5/36} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Cet exemple justifie la définition suivante. :

Définition de probabilité conditionnelle. Soit $\mathcal{E}$, une expérience aléatoire avec ensemble fondamental Ω. Soient A et B, des événements. Supposons que $\mathbb{P}[B] > 0$. La *probabilité conditionnelle* de A sachant B, dénotée $\mathbb{P}[A|B]$, est définie par

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

La Figure 5 illustre cette définition. Sachant que l'événement B s'est réalisé, la probabilité que l'événement A se soit réalisé est égale à la probabilité de l'intersection $A \cap B$ (région ombragée) divisée par la probabilité de B (région encerclée par une ligne épaisse).

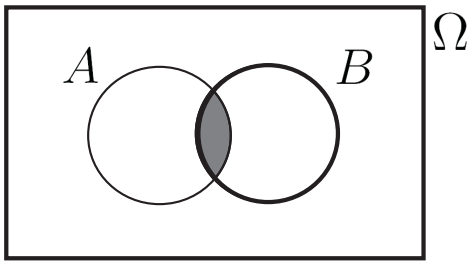


Figure 5. La probabilité conditionnelle de A sachant B.

2.4.2 Les principales propriétés des probabilités conditionnelles

On peut montrer que les six propriétés présentées à la section 2.1 sont également valides pour les probabilités conditionnelles. Plus précisément, on a les propriétés suivantes :

1. $0 \leq \mathbb{P}[A|B] \leq 1$.
2. $\mathbb{P}[\emptyset|B] = 0$ et $\mathbb{P}[\Omega|B] = 1$.
3. Si $A_1, A_2, A_3, ...$ sont mutuellement exclusifs, alors $\mathbb{P}[\bigcup A_i|B] = \sum \mathbb{P}[A_i|B]$.
4. Dans le cas équiprobable, $\mathbb{P}[A|B] =$ (cardinal de $A \cap B$)/(cardinal de B).
5. $\mathbb{P}[A|B] = 1 - \mathbb{P}[A^c|B]$.
6. $\mathbb{P}[E \cup F|B] = \mathbb{P}[E|B] + \mathbb{P}[F|B] - \mathbb{P}[E \cap F|B]$.

2.4.3 La règle de multiplication

Dans certains problèmes, les probabilités $\mathbb{P}[A \cap B]$ et $\mathbb{P}[B]$ sont ou bien données ou bien faciles à calculer. On peut alors utiliser la définition

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

pour calculer la probabilité conditionnelle $\mathbb{P}[A|B]$. Mais souvent ce sont la probabilité conditionnelle $\mathbb{P}[A|B]$ et la probabilité $\mathbb{P}[B]$ qui sont ou bien données ou bien facile à calculer. On peut alors calculer $\mathbb{P}[A \cap B]$ à l'aide de la *règle de multiplication* :

$$\mathbb{P}[A \cap B] = \mathbb{P}[B]\ \mathbb{P}[A|B]. \tag{2.3}$$

Les rôles de A et B étant symétriques dans l'expression $\mathbb{P}[A \cap B]$, on peut aussi écrire la règle de multiplication sous la forme

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\ \mathbb{P}[B|A]. \tag{2.4}$$

Dans la pratique, ce sont les données du problème qui nous indiquent laquelle des deux équations précédentes on doit utiliser. Souvent il y a un ordre chronologique qui suggère le bon choix.

Exemple 17. Un panier contient cinq boules noires et trois boules blanches. On tire deux boules au hasard et sans remise à partir du panier. Quelle est la probabilité d'obtenir deux boules noires ?

Solution. Si on pose

$$\begin{aligned} A &= \text{l'événement « obtenir une boule noire au premier tirage »,} \\ B &= \text{l'événement « obtenir une boule noire au deuxième tirage »,} \end{aligned}$$

alors la probabilité désirée est simplement $\mathbb{P}[A \cap B]$. Une simple application de la règle de multiplication nous donne

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B|A] = \frac{5}{8} \times \frac{4}{7} = \frac{5}{14}.$$

Une généralisation élémentaire de la règle de multiplication nous permet de traiter les intersections de plus de deux événements. Dans le cas d'une intersection de trois événements, disons $A \cap B \cap C$, on obtient

$$\mathbb{P}[A \cap B \cap C] = \mathbb{P}[A]\ \mathbb{P}[B|A]\ \mathbb{P}[C\,|A \cap B]$$

et dans le cas d'une intersection de quatre événements, disons $A \cap B \cap C \cap D$, on obtient

$$\mathbb{P}[A \cap B \cap C \cap D] = \mathbb{P}[A]\ \mathbb{P}[B|A]\ \mathbb{P}[C\,|A \cap B]\ \mathbb{P}[D\,|A \cap B \cap C].$$

Exemple 18. Un panier contient cinq boules bleues, six boules blanches et sept boules rouges. On tire quatre boules au hasard et sans remise à partir du panier. Quelle est la probabilité d'obtenir quatre boules de la même couleur ?

Solution. Si on pose

$$\begin{aligned} D &= \text{l'événement « obtenir 4 boules de la même couleur »,} \\ A &= \text{l'événement « obtenir 4 boules bleues »,} \\ B &= \text{l'événement « obtenir 4 boules blanches »,} \\ C &= \text{l'événement « obtenir 4 boules rouges »,} \end{aligned}$$

alors on a $D = A \cup B \cup C$. Les événements A, B, et C étant mutuellement exclusifs, on obtient

$$\mathbb{P}[D] = \mathbb{P}[A \cup B \cup C] = \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C].$$

La règle de multiplication nous donne

$$\begin{aligned} \mathbb{P}[A] &= \frac{5}{18} \times \frac{4}{17} \times \frac{3}{16} \times \frac{2}{15} = \frac{1}{612}, \\ \mathbb{P}[B] &= \frac{6}{18} \times \frac{5}{17} \times \frac{4}{16} \times \frac{3}{15} = \frac{3}{612}, \\ \mathbb{P}[C] &= \frac{7}{18} \times \frac{6}{17} \times \frac{5}{16} \times \frac{4}{15} = \frac{7}{612}, \end{aligned}$$

et on obtient donc

$$\mathbb{P}[D] = \frac{1}{612} + \frac{3}{612} + \frac{7}{612} = \frac{11}{612}.$$

2.4.4 Événements indépendants

Les événements A et B sont dits *indépendants* si on a $\mathbb{P}[A|B] = \mathbb{P}[A]$ et $\mathbb{P}[B|A] = \mathbb{P}[B]$. Le scénario classique donnant lieu à des évévements indépendants est le scénario des tirages avec remise. Voici un exemple illustratif.

Exemple 19. On fait deux tirages avec remise à partir d'un panier contenant cinq boules noires et trois boules blanches et on s'intéresse aux événements suivants :

$$\begin{aligned} A &= \text{l'événement « obtenir une boule blanche au premier tirage »,} \\ B &= \text{l'événement « obtenir une boule noire au deuxième tirage »,} \end{aligned}$$

On a $\mathbb{P}[B] = 3/8$ et $\mathbb{P}[B] = 5/8$. Considérons les situations suivantes :

SITUATION 1 : On a fait les deux tirages avec remise et on vous informe qu'on a obtenu une boule blanche au premier tirage. Étant donné cette information, quelle est la probabilité d'avoir obtenu une boule noire au deuxième tirage ? Réponse : puisqu'on fait des tirages avec remise, la connaissance du résultat du premier tirage ne nous donne aucune information au sujet du résultat du deuxième tirage et on a $\mathbb{P}[B|A] = \mathbb{P}[B] = 5/8$.

SITUATION 2 : On fait les deux tirages avec remise et on vous informe qu'on a obtenu une boule noire au deuxième tirage. Étant donné cette information, quelle est la probabilité d'avoir obtenu une boule blanche au premier tirage ? Réponse : puisqu'on fait des tirages avec remise, la connaissance du résultat du deuxième tirage ne nous donne aucune information au sujet du résultat du premier tirage et on a $\mathbb{P}[A|B] = \mathbb{P}[A] = 3/8$.

Dans l'exemple 19, les événements A et B sont donc indépendants.

Définition mathématique de l'indépendance de deux événements :

La définition d'indépendance donnée à la page précédente est très intuitive. Toutefois, les mathématiciens préfèrent définir l'indépendance de la façon suivante. Les événements A et B sont dits indépendants si on a

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B].$$

On peut montrer que cette définition est équivalente à la définition donnée à la page précédente. Un des avantages de cette définition est qu'elle se prête bien aux généralisations. Par exemple, on dit que les événements A, B et C sont indépendants si les quatre conditions suivantes sont satisfaites :

$$\begin{aligned} \mathbb{P}[A \cap B \cap C] &= \mathbb{P}[A] \cdot \mathbb{P}[B] \cdot \mathbb{P}[C] \\ \mathbb{P}[A \cap B] &= \mathbb{P}[A] \cdot \mathbb{P}[B] \\ \mathbb{P}[A \cap C] &= \mathbb{P}[A] \cdot \mathbb{P}[C] \\ \mathbb{P}[B \cap C] &= \mathbb{P}[B] \cdot \mathbb{P}[C] \end{aligned}$$

Dans les applications qui nous concernent, c'est le contexte qui nous dit si des événements sont indépendants ou non. Dans l'exemple 17, il est clair que les événements A et B ne sont pas indépendants. Dans l'exemple 19, il est clair que les événements A et B sont indépendants.

2.5 Variables aléatoires et distributions

Une variable aléatoire est une quantité dont la valeur numérique dépend du résultat d'une expérience aléatoire. Voici quelques exemples de variables aléatoires :

T = « le total obtenu lors du lancer d'une paire de dés »

N = « le nombre de boules rouges obtenues lorsqu'on fait trois tirages sans remise à partir d'un panier contenant cinq boules rouges et six boules noires »

V = « le nombre de lancers nécessaires pour obtenir un premier SIX dans une séquence de lancers d'un dé bien balancé »

X = « la distance entre le point choisi et le centre du disque lorsqu'on choisit un point au hasard sur un disque de rayon égal à cinq centimètres »

Y = « la quantité de précipitation (en mm) en juillet prochain à Québec »

Z = « le diamètre hauteur de poitrine (dhp) d'une épinette noire choisie au hasard ».

Les trois premiers exemples sont des exemples de *variables aléatoires discrètes*. Les trois autres sont des exemples de *variables aléatoires continues*.

Variable aléatoire discrète

Une variable aléatoire discrète est une variable aléatoire dont l'ensemble des valeurs possibles est un ensemble fini ou un ensemble infini dénombrable. Les variables T et N ci-dessus sont des exemples de variables aléatoires possédant seulement un nombre fini de valeurs possibles. La variable V est un exemple d'une variable aléatoire possédant un nombre infini dénombrable de valeurs possibles, l'ensemble de tous les entiers positifs. Si X est une variable aléatoire discrète, la fonction

$$p(x) = \mathbb{P}[X = x]$$

est appelée la *distribution* de la variable X. On dit aussi *la distribution de probabilité* ou encore *la fonction de probabilité* de la variable X. Cette fonction satisfait toujours les deux conditions suivantes :

$$p(x) \geq 0 \quad \text{pour tout } x \tag{2.5}$$

$$\sum_x p(x) = 1. \tag{2.6}$$

Si B est un ensemble de nombres réels, alors la probabilité que la variable aléatoire X prenne une valeur dans B est donnée par l'équation

$$\mathbb{P}[X \in B] = \sum_{x \in B} p(x). \tag{2.7}$$

La *moyenne* de la variable aléatoire discrète X, que l'on dénote μ, ou parfois μ_X, est définie par l'équation

$$\mu = \sum_x x p(x). \tag{2.8}$$

La moyenne μ est aussi appelée l'*espérance* de X, ou l'*espérance mathématique* de X, et est également dénotée $\mathbb{E}[X]$. La *variance* de la variable aléatoire discrète X, que l'on dénote σ^2, ou parfois σ_X^2, est définie par l'équation

$$\sigma^2 = \sum_x (x-\mu)^2 p(x). \tag{2.9}$$

La variance σ^2 est aussi dénotée $\mathbb{V}\mathrm{ar}[X]$. La racine carrée (positive) de la variance, c'est-à-dire la quantité σ, est appelée l'*écart-type* de la variable aléatoire X.

On interprète la moyenne et l'écart-type d'une variable aléatoire de la même façon qu'on interprétait la moyenne et l'écart-type d'une variable statistique au chapitre 1. Donc, si on répète notre expérience aléatoire un très grand nombre de fois et si on observe notre variable X à chaque fois, alors les valeurs observées auront tendance à être aux alentours de μ, plus ou moins à peu près σ.

Notez que nous faisons une distinction entre *variable statistique* et *variable aléatoire*. On parle de variable statistique lorsqu'on est en présence d'une population. On parle de variable aléatoire lorsqu'on est en présence d'une expérience aléatoire.

Trois exemples de variables aléatoires discrètes

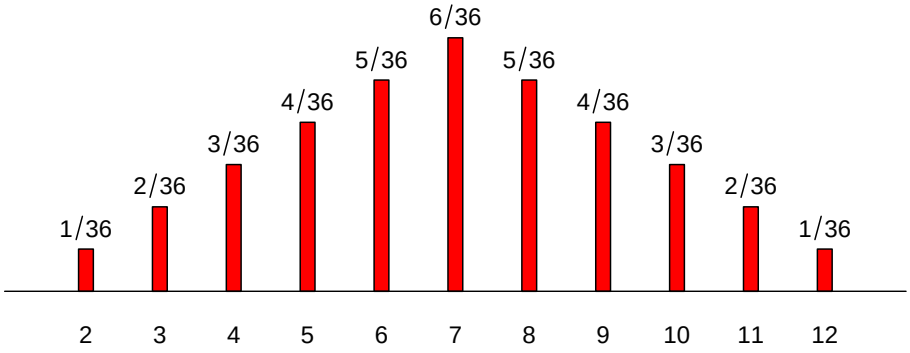
Exemple 20. On reprend l'exemple 1 du présent chapitre. Autrement dit, on considère l'expérience aléatoire qui consiste à lancer une paire de dés bien balancés. On s'intéresse au total des deux dés, c'est-à-dire la variable aléatoire T définie ci-dessus. Il s'agit bel et bien d'une variable aléatoire au sens de la définition donnée au début de la présente section : T est une quantité dont la valeur numérique dépend du résultat d'une expérience aléatoire. Par exemple, si on obtient le résultat $(3,5)$ alors on aura $T = 8$; si on obtient le résultat $(3,1)$ alors on aura $T = 4$. Le tableau suivant nous donne la *distribution* de la variable aléatoire T, c'est-à-dire la liste de toutes les valeurs possibles de T ainsi que les probabilités associées à chacune de ces valeurs possibles.

k	2	3	4	5	6	7	8	9	10	11	12
$p(k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Ces probabilités ont été obtenues en examinant le tableau présenté au tout début de la section 2.1. Par exemple, la probabilité $\mathbb{P}[T = 5]$ a été obtenue de la façon suivante :

$$\mathbb{P}[T=5] = \mathbb{P}[\{(4,1),(3,2),(2,3),(1,4)\}] = \frac{4}{36}.$$

Pour apprécier la *forme* de la distribution de probabilité de la variable T, on trace le graphe de la fonction $p(k)$:



Ce graphe étant symétrique autour de la valeur 7, on conclut que la moyenne est $\mu = 7$. On peut aussi obtenir μ à partir de l'équation (2.8), de la façon suivante :

$$\begin{aligned}\mu &= \sum_x xp(x) = \sum_{k=2}^{12} kp(k) \\ &= \left(2 \times \frac{1}{36}\right) + \left(3 \times \frac{2}{36}\right) + \left(4 \times \frac{3}{36}\right) + \cdots + \left(12 \times \frac{1}{36}\right) = 7.\end{aligned}$$

Pour l'écart-type, on obtient

$$\begin{aligned}\sigma &= \sqrt{\sum_x (x-\mu)^2 p(x)} = \sqrt{\sum_{k=2}^{12} (k-\mu)^2 p(k)} \\ &= \sqrt{\left((2-7)^2 \times \frac{1}{36}\right) + \left((3-7)^2 \times \frac{2}{36}\right) + \cdots + \left((12-7)^2 \times \frac{1}{36}\right)} \\ &= \sqrt{\frac{35}{12}} \approx 1.71.\end{aligned}$$

À la lumière de ces calculs, il est raisonnable d'affirmer que lorsqu'on lance une paire de dés bien balancés, on s'attend à ce que la somme des deux dés soit environ 7, plus ou moins environ 1.71. Pour calculer, par exemple, $\mathbb{P}[4 \leq T \leq 7]$, on utilise l'équation (2.7) et on obtient

$$\mathbb{P}[\,4 \leq T \leq 7\,] = p(4) + p(5) + p(6) + p(7) = \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} = \frac{18}{36} = \frac{1}{2}.$$

Exemple 21. Un panier contient cinq boules rouges et six boules noires. On fait trois tirages sans remise et on considère la variable aléatoire

$$N = \text{« le nombre de boules rouges parmi les trois boules tirées »}.$$

Le tableau suivant nous donne la *distribution* de la variable aléatoire N :

k	0	1	2	3
$p(k)$	4/33	15/33	12/33	2/33

Ces probabilités ont été obtenues de la façon suivante :

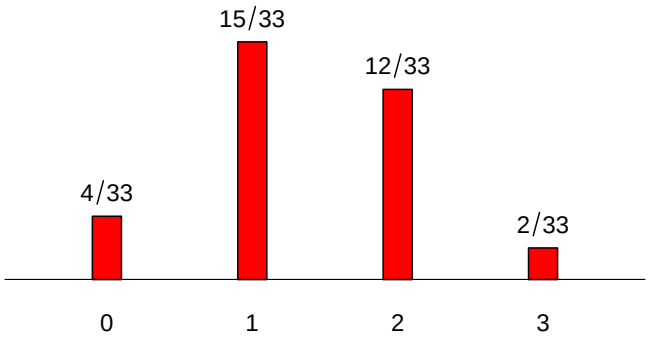
$$p(0) = \mathbb{P}[N=0] = \frac{\binom{5}{0}\binom{6}{3}}{\binom{11}{3}} = \frac{4}{33}$$

$$p(1) = \mathbb{P}[N=1] = \frac{\binom{5}{1}\binom{6}{2}}{\binom{11}{3}} = \frac{15}{33}$$

$$p(2) = \mathbb{P}[N=2] = \frac{\binom{5}{2}\binom{6}{1}}{\binom{11}{3}} = \frac{12}{33}$$

$$p(3) = \mathbb{P}[N=3] = \frac{\binom{5}{3}\binom{6}{0}}{\binom{11}{3}} = \frac{2}{33}$$

Voici le graphique de cette distribution :



Cette fois-ci on obtient

$$\mu = \left(0 \times \frac{4}{33}\right) + \left(1 \times \frac{15}{33}\right) + \left(2 \times \frac{12}{33}\right) + \left(3 \times \frac{2}{33}\right) = \frac{45}{33} = \frac{15}{11}$$

$$\sigma = \sqrt{\left(\left(0 - \frac{15}{11}\right)^2 \cdot \frac{4}{33}\right) + \left(\left(1 - \frac{15}{11}\right)^2 \cdot \frac{15}{33}\right) + \left(\left(2 - \frac{15}{11}\right)^2 \cdot \frac{12}{33}\right) + \left(\left(3 - \frac{15}{11}\right)^2 \cdot \frac{2}{33}\right)}$$
$$= \sqrt{72/121} = 0.7714.$$

Exemple 22. On lance un dé bien balancé jusqu'à ce qu'on obtienne la valeur six pour la première fois. On considère la variable aléatoire

$V =$ « le nombre total de lancers nécessaires pour obtenir notre premier six ».

L'ensemble des valeurs possibles de la variable V est l'ensemble de tous les entiers positifs. Si on pose

$$A_j = \text{« obtenir un SIX au } j^e \text{ lancer, »}$$

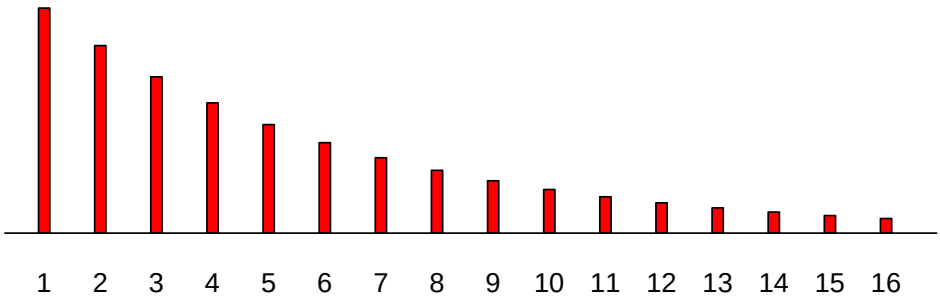
alors, puisque les événements A_1^c, A_2^c, A_3^c,... sont indépendants, on obtient

$$\begin{aligned}
\mathbb{P}[V = k] &= \mathbb{P}[A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k] \\
&= \mathbb{P}[A_1^c] \cdot \mathbb{P}[A_2^c] \cdot \mathbb{P}[A_{k-1}^c] \cdot \mathbb{P}[A_k] \\
&= \frac{5}{6} \times \frac{5}{6} \times \cdots \times \frac{5}{6} \times \frac{1}{6} \;=\; \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}.
\end{aligned}$$

La distribution de probabilité de la variable Y est donc donnée par

$$p(k) = \begin{cases} \dfrac{1}{6}\left(\dfrac{5}{6}\right)^{k-1} & \text{si } k \in \{1, 2, 3, ...\} \\ \\ 0 & \text{si } k \notin \{1, 2, 3, ...\} \end{cases}$$

Voici l'allure de cette distribution (tronquée à $k = 16$) :



Pour calculer, par exemple, $\mathbb{P}[V > 4]$ on peut procéder de la façon suivante :

$$\begin{aligned}
\mathbb{P}[V > 4] &= 1 - \mathbb{P}[V \leq 4] \\
&= \mathbb{P}[V = 1] + \mathbb{P}[V = 2] + \mathbb{P}[V = 3] + \mathbb{P}[V = 4] \\
&= \frac{1}{6} + \frac{5}{36} + \frac{25}{216} + \frac{125}{1296} \\
&= \frac{671}{1296} \;\approx\; 0.5177.
\end{aligned}$$

On peut montrer (voir la *loi géométrique* à la section 2.6.2) que $\mathbb{E}[V] = 6$ et $\mathbb{V}\text{ar}[V] = 30$. L'écart-type σ est donc $\sqrt{30} \approx 5.5$.

Variable aléatoire continue

Si X est une variable aléatoire continue, alors la fonction

$$f(x) = \frac{d}{dx}\, \mathbb{P}[X \leq x] \tag{2.10}$$

est appelée la *densité* de la variable X. On dit aussi *la densité de probabilité* de la variable aléatoire X. Cette fonction satisfait toujours les deux conditions suivantes :

$$f(x) \geq 0 \quad \text{pour tout } x \tag{2.11}$$

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1. \tag{2.12}$$

Si B est un ensemble de nombres réels, alors la probabilité que la variable aléatoire X prenne une valeur dans B est donnée par l'équation

$$\mathbb{P}[X \in B] = \int_B f(x)\, dx. \tag{2.13}$$

La *moyenne* de la variable aléatoire continue X, que l'on dénote μ, ou parfois μ_X, est définie par l'équation

$$\mu = \int_{-\infty}^{+\infty} x\, f(x)\, dx. \tag{2.14}$$

Comme dans le cas discret, la moyenne μ est aussi appelée l'*espérance* de X, ou l'*espérance mathématique* de X et est également dénotée $\mathbb{E}[X]$. La *variance* de la variable aléatoire continue X, que l'on dénote σ^2, ou parfois σ_X^2, est définie par l'équation

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\, dx. \tag{2.15}$$

Comme dans le cas discret, la variance σ^2 est aussi dénotée $\mathbb{V}\mathrm{ar}[X]$. La racine carrée (positive) de la variance, c'est-à-dire la quantité σ, est appelée l'*écart-type* de la variable aléatoire X. L'interprétation de la moyenne et de l'écart-type dans le cas continu est la même que dans le cas discret. Notez le parallèle entre les équations (2.11) à (2.15) et les équations (2.5) à (2.9).

La fonction

$$F(x) = \mathbb{P}[X \leq x]$$

est appelé la fonction de répartition de la variable X. L'équation (2.10) peut donc sécrire sous la forme

$$f(x) = \frac{d}{dx}\, F(x). \tag{2.16}$$

L'équation (2.13) avec $B = (-\infty, x]$ nous permet d'obtenir

$$F(x) = \mathbb{P}[X \leq x] = \mathbb{P}[X \in (-\infty, x]] = \int_{-\infty}^{x} f(u)\, du.$$

On a donc

$$F(x) = \int_{-\infty}^{x} f(u)\, du. \tag{2.17}$$

Bref, dans le cas continu, la densité de probabilité $f(x)$ et la fonction de répartition $F(x)$ sont deux façons de décrire la distribution d'une variable aléatoire. L'équation (2.16) nous

permet de calculer $f(x)$ à partir de $F(x)$ et l'équation (2.17) nous permet de calculer $F(x)$ à partir de $f(x)$.

Un exemple de variable aléatoire continue

EXEMPLE 23. On choisit un point au hasard et de façon uniforme sur un disque de rayon égal à cinq centimètres et on pose

$$X = \text{« } \textit{la distance entre le point choisi et le centre du disque.} \text{ »}$$

Calculons la fonction de répartition de la variable X. Si $x < 0$, alors on a bien sûr $F(x) = \mathbb{P}[X \leq x] = 0$. Si $x > 5$, alors on a bien sûr $F(x) = \mathbb{P}[X \leq x] = 1$. Pour $0 \leq x \leq 5$ on obtient

$$\begin{aligned} F(x) &= \mathbb{P}[X \leq x] \\ &= \frac{\text{surface du disque de rayon x}}{\text{surface du disque de rayon 5}} \\ &= \frac{\pi\, x^2}{\pi\, 5^2} = \frac{x^2}{25}. \end{aligned}$$

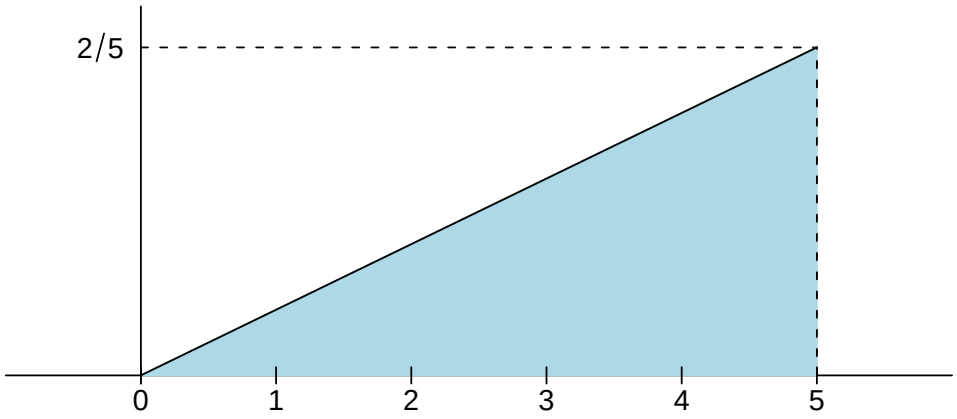
On a donc

$$F(x) = \begin{cases} 0 & \text{si } -\infty < x < 0 \\ x^2/25 & \text{si } 0 \leq x \leq 5 \\ 1 & \text{si } 5 < x < \infty \end{cases}$$

À l'aide de l'équation (2.16) on obtient la densité de la variable X :

$$f(x) = \begin{cases} 2x/25 & \text{si } 0 \leq x \leq 5 \\ 0 & \text{sinon.} \end{cases}$$

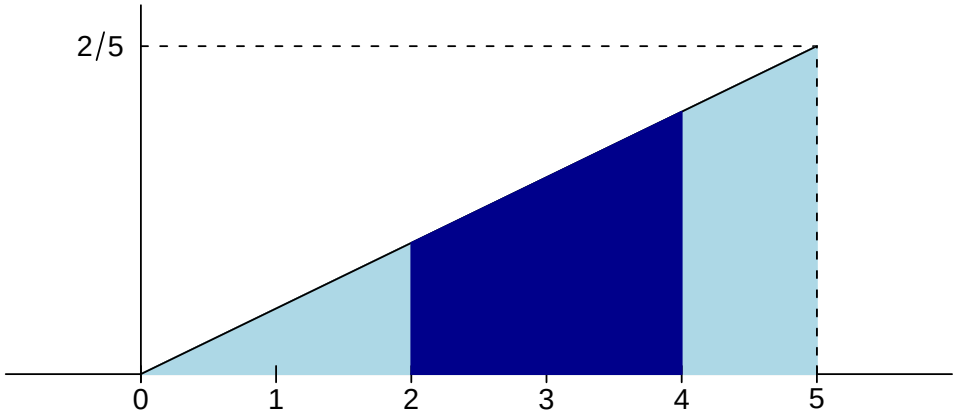
Voici le graphe de cette densité de probabilité :



Pour calculer $\mathbb{P}[2 \le X \le 4]$, $\mathbb{E}[X]$ et $\mathbb{V}\mathrm{ar}[X]$, on utilise les équations (2.13), (2.14) et (2.15) :

$$\begin{aligned}
\mathbb{P}[2 \le X \le 4] &= \int_2^4 f(x)\,dx = \int_2^4 \frac{2x}{25}\,dx = \frac{x^2}{25}\Big|_2^4 = \frac{12}{25} \\
\mathbb{E}[X] &= \int_{-\infty}^{+\infty} x\,f(x)\,dx = \int_0^5 x\,\frac{2x}{25}\,dx = \frac{2x^3}{75}\Big|_0^5 = \frac{10}{3} \\
\mathbb{V}\mathrm{ar}[X] &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx = \int_0^5 (x-(10/3))^2\,\frac{2x}{25}\,dx = \frac{25}{18}.
\end{aligned}$$

La probabilité $\mathbb{P}[2 \le X \le 4]$ est représentée par la surface de la région ombragée foncée dans le graphe ci-dessous :



Propriétés de l'espérance

Nous avons vu que l'espérance d'une variable aléatoire est définie de la façon suivante :

$$\mu_X = \mathbb{E}[X] = \begin{cases} \sum_k k\,p(k) & \text{si } X \text{ est une variable discrète} \\ \int_{-\infty}^{+\infty} x\,f(x)\,dx & \text{si } X \text{ est une variable continue.} \end{cases}$$

Plus généralement, l'espérance d'une fonction d'une variable aléatoire, disons la fonction $g(X)$, est donnée par

$$\mu_{g(X)} = \mathbb{E}[g(X)] = \begin{cases} \sum_k g(k)\,p(k) & \text{si } X \text{ est une variable discrète} \\ \int_{-\infty}^{+\infty} g(x)\,f(x)\,dx & \text{si } X \text{ est une variable continue.} \end{cases} \tag{2.18}$$

En comparant les équations (2.9) et (2.15) avec l'équation (2.18), on observe que $\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[(X-\mu_X)^2]$. La variance de la variable X est donc l'espérance de la variable $(X-\mu_X)^2$.

Voici les quatre principales propriétés de l'espérance et de la variance :

(a) $\mathbb{E}[aX+b] = a\mathbb{E}[X] + b$.

(b) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

(c) $\mathbb{V}\text{ar}[aX+b] = a^2\mathbb{V}\text{ar}[X]$.

(d) $\mathbb{V}\text{ar}[X+Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y]$ si les variables X et Y sont indépendantes.

Les propriétés (a) et (c) sont des conséquences de l'équation (2.18). Les propriétés (b) et (d) sont des conséquences de la version bivariée de l'équation (2.18). Pour la définition d'indépendance de variables aléatoires, voir la section 2.7.

Pour calculer la variance d'une variable aléatoire, on utilise souvent le fait que

$$\begin{aligned}
\mathbb{V}\text{ar}[X] &= \mathbb{E}[(X-\mu_X)^2] \\
&= \mathbb{E}[X^2 - 2\mu_X X + \mu_X^2] \\
&= \mathbb{E}[X^2] - 2\mu_X\mathbb{E}[X] + \mu_X^2 \\
&= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

Bref, on a toujours

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \tag{2.19}$$

Cette équation est parfois appelée la *formule raccourci* pour calculer la variance. Revenons à l'exemple 23. On a obtenu $\mathbb{E}[X] = 10/3$. On peut calculer $\mathbb{E}[X^2]$ de la même façon :

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2\, f(x)\, dx \;=\; \int_0^5 x^2\, \frac{2x}{25}\, dx \;=\; \left.\frac{2x^4}{100}\right|_0^5 \;=\; \frac{25}{2}.$$

La variance est donc

$$\mathbb{V}\text{ar}[X] \;=\; \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \;=\; \frac{25}{2} - \left(\frac{10}{3}\right)^2 \;=\; \frac{25}{18}.$$

2.6 Quelques lois de probabilité classiques

Dans la présente section, nous présentons quelques distributions de probabilité classiques. Ces distributions de probabilité sont souvent appelées *lois de probabilité*. Pour chaque distribution, nous présentons la fonction de probabilité ou la densité de probabilité, selon que l'on est dans le cas discrèt ou dans le cas continu. Nous expliquons dans quel contexte la distribution est utilisée. Enfin, nous présentons les principales propriétés de la distribution.

2.6.1 La loi binomiale

Une épreuve de Bernoulli est une expérience aléatoire avec seulement deux résultats possibles, disons S, pour succès, et E, pour échec. La probabilité de succès est dénotée p. La probabilité d'échec est donc $1-p$.

Fixons n, un entier positif, et considérons une séquence de n épreuves de Bernoulli indépendantes les unes des autres et ayant toutes la même probabilité de succès p. Posons

$$Y = \text{ le nombre de succès parmi ces } n \text{ épreuves de Bernoulli.}$$

La fonction de probabilité de la variable Y est alors donnée par

$$\mathbb{P}[Y=k] = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{si } k \in \{0,1,2,...,n\} \\ 0 & \text{si } k \notin \{0,1,2,...,n\}. \end{cases}$$

Cette distribution de probabilité est appelée la loi binomiale avec paramètres n et p, ou tout simplement la *loi binomiale(n,p)*. On écrit parfois $Y \sim$ binomiale(n,p) pour signifier que la variable Y *suit* la loi binomiale(n,p). La moyenne et la variance de la loi binomiale sont données par

$$\begin{aligned} \mu &= np \\ \sigma^2 &= np(1-p) \end{aligned}$$

Si $n=1$ alors on obtient

$$\mathbb{P}[Y=k] = \begin{cases} 1-p & \text{si } k=0 \\ p & \text{si } k=1 \\ 0 & \text{si } k \notin \{0,1\}. \end{cases}$$

Ce cas spécial de la loi binomiale est appelé la loi de Bernoulli avec paramètre p.

Exemple 24. Un panier contient 10 boules rouges et 15 boules noires. On fait 8 tirages avec remise. Quelle est la probabilité d'obtenir exactement 5 boules rouges ? Quelle est l'espérance du nombre de boules rouges tirées ? Quel est l'écart-type du nombre de boules rouges tirées ?

Solution : La distribution du nombre de boules rouges tirées est la loi binomiale$(8, 2/5)$. La probabilité d'obtenir exactement 5 boules rouges est donc $\binom{8}{5} \cdot (2/5)^5 (3/5)^3 \approx 0.1239$. L'espérance est $np = 8 \cdot (2/5) \approx 3.333$ et l'écart-type est $\sqrt{np(1-p)} = \sqrt{8 \cdot (2/5) \cdot (3/5)} \approx 1.386$.

Avec le logiciel R :

- La commande `dbinom(k,n,p)` nous donne la probabilité binomiale $\binom{n}{k} p^k (1-p)^{n-k}$.
- La commande `pbinom(k,n,p)` nous donne la probabilité binomiale cumulative $\sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$.
- La commande `rbinom(m,n,p)` nous donne un échantillon aléatoire de taille m issu de la loi binomiale avec paramètres n et p.

2.6.2 La loi géométrique[2]

Considérons une séquence d'épreuves de Bernoulli indépendantes les unes des autres et ayant toutes la même probabilité de succès p. Posons

$$\begin{aligned} T &= \text{le nombre d'épreuves nécessaires pour obtenir le premier succès.} \\ N &= \text{le nombre d'échecs qui surviennent avant le premier succès.} \end{aligned}$$

[2]On peut omettre cette section, ou la laisser en exercice, si on manque de temps

Bien sûr on a $T = N + 1$. La fonction de probabilité de la variable T est donnée par

$$\mathbb{P}[T = k] = \begin{cases} (1-p)^{k-1}p & \text{si } k \in \{1, 2, 3, ...\} \\ \\ 0 & \text{si } k \notin \{1, 2, 3, ...\}. \end{cases}$$

Cette distribution de probabilité est appelée la loi géométrique sur $\{1, 2, 3, ...\}$ avec paramètres p. Nous l'avons déjà rencontrée à l'exemple 22. La moyenne et la variance de la loi géométrique sur $\{1, 2, 3, ...\}$ sont données par

$$\begin{aligned} \mu &= 1/p \\ \sigma^2 &= (1-p)/p^2 \end{aligned}$$

La fonction de probabilité de la variable N est donnée par

$$\mathbb{P}[N = k] = \begin{cases} (1-p)^{k}p & \text{si } k \in \{0, 1, 2, 3, ...\} \\ \\ 0 & \text{si } k \notin \{0, 1, 2, 3, ...\}. \end{cases}$$

Cette distribution de probabilité est appelée la loi géométrique sur $\{0, 1, 2, 3, ...\}$ avec paramètres p. La moyenne et la variance de la loi géométrique sur $\{0, 1, 2, 3, ...\}$ sont données par

$$\begin{aligned} \mu &= (1/p) - 1 = \frac{1-p}{p} \\ \sigma^2 &= (1-p)/p^2 \end{aligned}$$

Exemple 25. Un panier contient 10 boules rouges et 15 boules noires. On fait des tirages avec remise jusqu'à ce qu'on obtienne notre première boule rouge. Quelle est la probabilité qu'il faille exactement 5 tirages ? Quelle est l'espérance du nombre de tirages que ça va prendre ? Quel est l'écart-type du nombre de tirages que ça va prendre ?

Solution : La distribution du nombre de tirages est la loi géométrique sur $\{1, 2, 3, ...\}$ avec paramètre $p = 2/5$. La probabilité que ça va prendre exactement 5 tirages est donnée par $(3/5)^4 \cdot (2/5) = 162/3125 \approx 0.0518$. L'espérance est $1/p = 5/2$ et l'écart-type est $\sqrt{(1-p)/p^2} = \sqrt{(3/5)/(2/5)^2} \approx 1.936$.

Avec le logiciel R :

La loi géométrique du logiciel R est la loi géométrique sur $\{0, 1, 2, 3, ...\}$.

- La commande `dgeom(k,p)` nous donne la probabilité géométrique $(1-p)^k p$.
- La commande `pgeom(k,p)` nous donne la probabilité géométrique cumulative $\sum_{j=0}^{k}(1-p)^j p$.
- La commande `rgeom(m,p)` nous donne un échantillon aléatoire de taille m issu de la loi géométrique sur $\{0, 1, 2, 3, ...\}$ avec paramètre p.

2.6.3 La loi hypergéométrique[3]

Un panier contient a boules rouges et b boules noires. On fait n tirages sans remise à partir de ce panier. On suppose bien sûr que $0 \leq n \leq a+b$. On considére la variable aléatoire

$$Y = \text{ le nombre de boules rouges parmi les } n \text{ boules tirées.}$$

Si on avait fait des tirages avec remise, la distribution de la variable Y aurait été la loi binomiale(n,p) avec $p = a/(a+b)$. Mais puisqu'on fait des tirages sans remise, on obtient

$$\mathbb{P}[Y=k] \;=\; \begin{cases} \dfrac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}} & \text{si } k \in \{0,1,2,...,n\} \\ \\ 0 & \text{si } k \notin \{0,1,2,...,n\}. \end{cases}$$

Cette distribution s'appelle la loi hypergéométrique(a,b,n). Nous l'avons déjà rencontrée à l'exemple 21. On écrit $Y \sim$ hypergéométrique(a,b,n) pour signifier que la variable Y suit la loi hypergéométrique(a,b,n). Il faut toutefois être prudent car il n'y a pas de notation standard pour désigner la loi hypergéométrique.

NOTE : Puisqu'on fait n tirages, il est clair que le nombre de boules rouges obtenues sera un entier parmi $\{0,1,2,...,n\}$. Toutefois, ce nombre ne peut excéder a (puisqu'il y a seulement a boules rouges dans le panier) et il ne peut être inférieur à $n-b$ (puisqu'il y a seulement b boules noires dans le panier). Néanmoins, l'équation donnée ci-dessus est quand même toujours valide si on interprète correctement les coefficients binomiaux. Par exemple, si $a=3$, $b=4$ et $n=5$, alors les seules valeurs possibles de Y sont les valeurs $\{1,2,3\}$; on ne peut pas obtenir 4 rouges ou 5 rouges puisqu'il y a seulement 3 rouges dans le panier et on est certain d'obtenir au moins une rouge puisqu'on fait 5 tirages et qu'il y a seulement 4 noires dans le panier. L'équation ci-dessus est tout de même valide pour tout $k \in \{0,1,2,3,4,5\}$. Par exemple, avec $k=4$ on obtient

$$\mathbb{P}[Y=4] \;=\; \frac{\binom{a}{4}\binom{b}{n-4}}{\binom{a+b}{n}} \;=\; \frac{\binom{3}{4}\binom{4}{1}}{\binom{7}{5}} \;=\; 0$$

puisque $\binom{3}{4} = 0$ (il y a 0 façon de choisir 4 objets parmi un groupe de 3 objets !)

La moyenne et la variance de la loi hypergéométrique sont données par

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

Il est intéressant de noter que si on avait fait des tirages avec remise plutôt que sans remise, la distribution de Y aurait été la loi binomiale(n,p) avec $p = a/(a+b)$ et la moyenne et la variance auraient été

$$\begin{aligned} \mu &= np \;=\; n\,\frac{a}{a+b} \\ \sigma^2 &= np(1-p) \;=\; n\,\frac{a}{a+b}\,\frac{b}{a+b}. \end{aligned}$$

[3]On peut omettre cette section, ou la laisser en exercice, si on manque de temps

Le facteur

$$\left(1 - \frac{n-1}{a+b-1}\right)$$

qui apparaît dans la formule pour la variance de la loi hypergéométrique est parfois appelé le *facteur de correction*. Notez que si $a+b$ est beaucoup plus grand que n, alors le facteur de correction est à peu près 1 et la variance de la loi hypergéométrique est à peu près égale à la variance de la loi binomiale. Ceci correspond au fait que lorsque le nombre de boules dans le panier est très grand par rapport au nombre de tirages, il y a peu de différences entre faire des tirages sans remise et faire des tirages avec remise.

EXEMPLE 26. Un panier contient 10 boules rouges et 15 boules noires. On fait 8 tirages sans remise. Quelle est la probabilité d'obtenir exactement 5 boules rouges ? Quelle est l'espérance du nombre de boules rouges tirées ? Quel est l'écart-type du nombre de boules rouges tirées ?

SOLUTION : La distribution du nombre de boules rouges parmi les 8 boules tirées est la loi hypergéométrique$(10, 15, 8)$. La probabilité d'obtenir exactement 5 boules rouges est $\binom{10}{5}\binom{15}{3}/\binom{25}{8} \approx 0.1060$. L'espérance est $n \cdot a/(a+b) = 8 \cdot (10/25) \approx 3.3333$ et l'écart-type est

$$\sqrt{n\,\frac{a}{a+b}\,\frac{b}{a+b}\left(1 - \frac{n-1}{a+b-1}\right)} = \sqrt{8 \times \frac{10}{25} \times \frac{15}{25} \times \left(1 - \frac{7}{24}\right)} \approx 1.1662.$$

Avec le logiciel R :

- La commande `dhyper(k,a,b,n)` nous donne la probabilité hypergéométrique $\binom{a}{k}\binom{b}{n-k}/\binom{a+b}{n}$
- La commande `phyper(k,a,b,n)` nous donne la probabilité hypergéométrique cumulative $\sum_{j=0}^{k}\binom{a}{j}\binom{b}{n-j}/\binom{a+b}{n}$.
- La commande `rhyper(m,a,b,n)` nous donne un échantillon aléatoire de taille m issu de la loi hypergéométrique(n, a, b).

2.6.4 La loi de Poisson

La loi de Poisson avec paramètre λ est la distribution de probabilité discrète suivante :

$$p(k) = \begin{cases} e^{-\lambda}\,\dfrac{\lambda^k}{k!} & \text{si } k \in \{0,1,2,3,...\} \\ 0 & \text{si } k \notin \{0,1,2,3,...\}. \end{cases}$$

On écrit $X \sim \text{Poisson}(\lambda)$ pour signifier que la variable aléatoire X suit la loi de Poisson avec paramètre λ. Ici le paramètre λ est un nombre réel positif. Cette loi de probabilité est parfois appelée la *loi des événements rares*. Ceci s'explique par le fait suivant :

> Si X suit la loi binomiale(n, p) avec n grand et p petit, alors X suit à peu près la loi de Poisson(λ) avec $\lambda = np$. Autrement dit, lorsque n est très grand et p est très petit, la loi de Poisson(λ) avec $\lambda = np$ est une bonne approximation pour la loi binomiale(n, p).

Pour illustrer, supposons que $X \sim$ binomiale$(200, 0.01)$. Le tableau suivant montre que les probabilités de la loi de Poisson(2) sont presqu'égales à celles de la loi binomiale$(200, 0.01)$.

k	$\binom{200}{k}(0.01)^k(0.99)^{200-k}$	$e^{-2}\ 2^k/k!$
0	0.13398	0.13534
1	0.27066	0.27067
2	0.27203	0.27067
3	0.18136	0.18045
4	0.09022	0.09022
5	0.03572	0.03609
6	0.01173	0.01203
7	0.00328	0.00344
8	0.00080	0.00086
9	0.00017	0.00019

La moyenne et la variance de la loi de Poisson sont égales au paramètre λ :

$$\begin{aligned} \mu &= \lambda \\ \sigma^2 &= \lambda. \end{aligned}$$

Exemple 27. Supposons que 1.5% des boutures de mélèze ne s'enracinent pas. On dispose de 400 boutures. Quelle est la probabilité qu'au moins 10 de ces boutures ne s'enracinent pas ? Quelle est l'espérance du nombre de boutures qui ne s'enracineront pas ? Quel est l'écart-type du nombre de boutures qui ne s'enracineront pas ?

Solution. Posons

$$X = \text{ le nombre de boutures qui ne s'enracineront pas.}$$

En supposant que les boutures se comportent indépendamment les unes des autres, la variable aléatoire X suit la loi binomiale$(400, 0.015)$. On peut approximer cette loi binomiale par la loi de Poisson avec paramètre $\lambda = np = 400 \times 0.015 = 6$. On obtient donc les réponses approximatives suivantes :

$$\begin{aligned} \mathbb{P}[X \geq 10] &= 1 - \mathbb{P}[X \leq 9] \\ &= 1 - \sum_{k=0}^{9} \mathbb{P}[X = k] \\ &\approx 1 - \sum_{k=0}^{9} e^{-6}\ \frac{6^k}{k!} \ = \ 1 - 0.9161 = 0.0839 \\ \mu &= \lambda \ = \ 6 \\ \sigma &\approx \sqrt{\lambda} \ = \ \sqrt{6} \ \approx \ 2.4495. \end{aligned}$$

Si on fait le calcul exact avec la loi binomiale, on obtient

$$\begin{aligned}
\mathbb{P}[X \geq 10] &= 1 - \mathbb{P}[X \leq 9] \\
&= 1 - \sum_{k=0}^{9} \mathbb{P}[X = k] \\
&= 1 - \sum_{k=0}^{9} \binom{400}{k} (0.015)^k (0.985)^{400-k} \;=\; 1 - 0.9176 = 0.0824 \\
\mu &= np \;=\; 6 \\
\sigma &= \sqrt{np(1-p)} \;=\; \sqrt{6} \;\approx\; 2.4310.
\end{aligned}$$

Avec le logiciel R :

- La commande `dpois(k,λ)` nous donne la probabilité de Poisson $e^{-\lambda}\,\lambda^k/k!$
- La commande `ppois(k,λ)` nous donne la probabilité de Poisson cumulative $\sum_{j=0}^{k} e^{-\lambda}\,\lambda^j/j!$.
- La commande `rpois(m,λ)` nous donne un échantillon aléatoire de taille m issu de la loi de Poisson(λ).

2.6.5 La loi uniforme[4]

Si X est un nombre choisi au hasard et de façon uniforme sur l'intervalle $[a, b]$, alors X est une variable aléatoire continue avec densité

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon.} \end{cases}$$

Cette distribution de probabilité est appelée la loi uniforme sur l'intervalle $[a, b]$. On écrit $X \sim$ uniforme(a, b) pour signifier que la variable aléatoire X suit la loi uniforme sur l'intervalle $[a, b]$. Voici le graphe de cette densité de probabilité :

La hauteur de cette boîte rectangulaire est $1/(b-a)$ de sorte que sa surface est bel et bien égale à 1. Par symétrie, la moyenne est à mi-chemin entre a et b. On a donc

$$\mu = \frac{a+b}{2}.$$

[4]On peut omettre cette section, ou la laisser en exercice, si on manque de temps

Calculons la variance à l'aide de la formule raccourci. On suppose qu'on a X suit la loi uniforme sur l'intervalle (a, b). D'abord on obtient

$$\mathbb{E}[X^2] = \int_a^b x^2 \ \frac{1}{b-a} \ dx = \frac{1}{b-a} \ \frac{x^3}{3}\bigg|_a^b = \frac{a^2+ab+b^2}{3}.$$

On obtient donc

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

L'écart-type est donc

$$\sigma = \frac{b-a}{2\sqrt{3}}.$$

Enfin, notons que la fonction de répartition de la loi uniforme est donnée par

$$F(x) = \begin{cases} 0 & \text{si } -\infty < x < a \\ \dfrac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } b < x < \infty. \end{cases}$$

Avec le logiciel R :

- La commande `dunif(x,a,b)` nous donne la valeur de la densité uniforme(a, b) au point x. Si $x \in [a, b]$, cette valeur est $1/(b-a)$. Si $x \notin [a, b]$, cette valeur est 0.
- La commande `punif(x,a,b)` nous donne la valeur de la fonction de répartition de la loi uniforme(a, b) au point x. Si $x < a$, cette valeur est 0. Si $x > b$, cette valeur est 1. Si $x \in [a, b]$, cette valeur est $(x-a)/(b-a)$.
- La commande `qunif(`γ`,a,b)` nous donne le quantile d'ordre γ de la loi uniforme(a, b).
- La commande `runif(m,a,b)` nous donne un échantillon aléatoire de taille m issu de la loi uniforme(a, b).

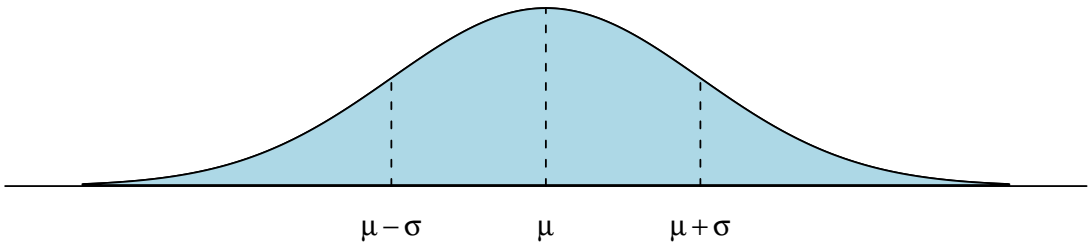
2.6.6 La loi normale

La loi normale avec moyenne μ et avec variance σ^2, dénotée $N(\mu, \sigma^2)$, est la distribution continue avec densité

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma} \ e^{-(x-\mu)^2/2\sigma^2}$$

Il s'agit de la célèbre densité de probabilité en forme de cloche. Cette densité est symétrique et centrée à μ. On écrit $X \sim N(\mu, \sigma^2)$ pour signifier que la variable aléatoire X suit la loi $N(\mu, \sigma^2)$. La loi $N(0, 1)$ est parfois appelée la loi normale *standard* ou la loi normale *centrée et réduite*. On utilise souvent la lettre Z pour dénoter une variable aléatoire qui suit la loi $N(0, 1)$. Le graphe suivant nous montre la forme de la loi $N(\mu, \sigma^2)$ et nous permet de visualiser la moyenne μ et l'écart-type σ.

La loi $N(\mu, \sigma^2)$:



Les principales propriétés de la loi normale

(a) Si $X \sim N(\mu, \sigma^2)$ et si $Y = aX + b$, alors $Y \sim N(a\mu + b, a^2\sigma^2)$.

(b) En particulier, si $X \sim N(\mu, \sigma^2)$ alors la variable $Z = (X - \mu)/\sigma$ suit la loi $N(0, 1)$.

(c) Si $X \sim N(\mu_1, \sigma_2^2)$ et $Y \sim N(\mu_1, \sigma_2^2)$, et si les variables X et Y sont indépendantes, alors $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Calcul de probabilités avec la loi normale

Pour calculer des probabilités avec la loi $N(\mu, \sigma^2)$, on se ramène à la loi $N(0, 1)$. La propriété (b) ci-dessus nous permet de faire cela. La densité de probabilité de la loi $N(0, 1)$ est dénotée $\phi(x)$. On a donc

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}.$$

La fonction de répartition de la loi $N(0, 1)$ est dénotée $\Phi(z)$. On a donc

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\, dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

L'intégrale ci-dessus ne peut être évaluée que par des méthodes numériques. Qu'importe, la fonction $\Phi(z)$ est programmée dans tous les logiciels de statistique et dans la plupart des calculatrices scientifiques. Avant l'époque des microprocesseurs, on consultait des tables de la loi $N(0, 1)$ pour trouver les différentes valeurs de la fonction $\Phi(z)$. Une telle table est reproduite à l'annexe A.1.

Exemple 28. Supposons que la distribution de la variable aléatoire X soit la loi normale avec moyenne 100 et avec écart-type 16. Calculez $\mathbb{P}[92 < X < 124]$.

Solution. À l'aide de la propriété (b), on se ramène à la loi $N(0, 1)$. On exprime ensuite la probabilité désirée en terme de la fonction $\Phi(z)$. Puis on utilise la table de la loi normale. Voici les détails :

$$\begin{aligned}
\mathbb{P}[92 < X < 124] &= \mathbb{P}\left[\frac{92-100}{16} < \frac{X-100}{16} < \frac{124-100}{16}\right] \\
&= \mathbb{P}[-0.50 < Z < 1.50] \\
&= \mathbb{P}[-0.50 < Z \leq 1.50] \\
&= \mathbb{P}[Z \leq 1.50] - \mathbb{P}[Z \leq -0.50] \\
&= \Phi(1.50) - \Phi(-0.50) \\
&= 0.9332 - 0.3085 \;=\; 0.6247.
\end{aligned}$$

Avec le logiciel R :

- La commande `dnorm(x,`μ, σ`)` nous donne la valeur de la densité $N(\mu, \sigma^2)$ à x.
- La commande `pnorm(x,`μ, σ`)` nous donne la valeur de la fonction de répartition de la loi $N(\mu, \sigma^2)$ au point x, c'est-à-dire la valeur de la fonction de répartition de la loi $N(0, 1)$ au point $(x - \mu)/\sigma$, c'est-à-dire la valeur de $\Phi((x - \mu)/\sigma)$.
- La commande `qnorm(`γ, μ, σ`)` nous donne le quantile d'ordre γ de la loi $N(\mu, \sigma^2)$.
- La commande `rnorm(m,`μ, σ`)` nous donne un échantillon aléatoire de taille m issu de la loi $N(\mu, \sigma^2)$.

Exemple 29. Supposons que la loi $N(54, 81)$ soit un bon modèle pour décrire la distribution des poids, en kg, des individus adultes dans une certaine population animale. Déterminez la proportion d'individus ayant un poids compris entre 50 kg et 65 kg. Déterminez le 95^e centile de cette distribution de poids.

Solution avec R. La distribution des poids est la loi normale avec moyenne $\mu = 54$ kg et avec écart-type $\sigma = 9$ kg. Pour obtenir la proportion d'individus ayant un poids compris entre 50 kg et 65 kg, on tape la commande `pnorm(65, 54, 9) - pnorm(50, 54, 9)` et R nous donne la réponse `0.5608`. On peut donc dire qu'environ 56% des individus de cette population ont un poids situé entre 50 kg et 65 kg. Pour obtenir le 95^e centile des poids, on tape la commande `qnorm(0.95, 54, 9)` et R nous donne la réponse `68.8037`. Le 95^e centile des poids est donc environ 68.8 kg.

Le théorème limite central

Supposons que $X_1, X_2, X_3, ..., X_n$ soient des variables aléatoires indépendantes et identiquement distribuées avec moyenne μ et variance σ^2. Posons

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Les propriétés (b) et (d) de la section 2.5 nous donnent

$$\begin{aligned}
\mathbb{E}[S_n] &= \mathbb{E}[X_1] + \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_1] \;=\; \mu + \mu + \cdots + \mu \;=\; n\mu \\
\mathbb{V}\mathrm{ar}[S_n] &= \mathbb{V}\mathrm{ar}[X_1] + \mathbb{V}\mathrm{ar}[X_2] + \cdots + \mathbb{V}\mathrm{ar}[X_n] \;=\; \sigma^2 + \sigma^2 + \cdots + \sigma^2 \;=\; n\sigma^2.
\end{aligned}$$

Le théorème limite central va un peu plus loin. Il nous dit que si n est suffisamment grand alors la variable aléatoire S_n suit à peu près la loi normale. On a alors

$$S_n \approx N(n\mu, n\sigma^2). \tag{2.20}$$

Grâce à la propriété (b) de la présente section, on peut aussi écrire ce résultat sous la forme suivante :

$$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \approx N(0,1). \tag{2.21}$$

Bref, le théorème limite central nous dit que la somme d'un grand nombre de variables aléatoires indépendantes et identiquement distribuées suit, à peu près, une loi normale. On peut aussi exprimer le théorème limite central en termes de la variable

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

de la façon suivante

$$\overline{X} \approx N(\mu, \sigma^2/n) \tag{2.22}$$

ou, de façon équivalente, sous la forme

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1). \tag{2.23}$$

Les équations (2.20) à (2.23) sont donc quatre façons différentes, mais équivalentes, d'énoncer le théorème limite central.

Exemple 30. Supposons que $X_1, X_2, X_3, ..., X_{768}$ soient des variables aléatoires indépendantes et identiquement distribuées selon la loi uniforme sur l'intervalle $[0,1]$. Rappelons que la moyenne et la variance de la loi uniforme sur l'intervalle $[0,1]$ sont respectivement $\mu = 1/2$ et $\sigma^2 = 1/12$. Que peut-on dire de la variable aléatoire $T = \sum_{j=1}^{768} X_j$? Quelle est la moyenne de la variable T ? Quel est l'écart-type de la variable T ? Que vaut la probabilité $\mathbb{P}[380 < T < 400]$?

Solution. Les propriétés de l'espérance et de la variance nous donnent

$$\begin{aligned}
\mathbb{E}[T] &= n\mu = 768 \times \frac{1}{2} = 384, \\
\mathbb{V}\mathrm{ar}[T] &= n\sigma^2 = 768 \times \frac{1}{12} = 64.
\end{aligned}$$

On a donc $\mu_T = 384$ et $\sigma_T = 8$. Le théorème limite central nous permet de conclure que $T \approx N(384, 64)$. On obtient donc

$$\begin{aligned}
\mathbb{P}[380 < T < 400] &= \mathbb{P}\left[\frac{380-384}{8} < \frac{T-384}{8} < \frac{400-384}{8}\right] \\
&\approx \mathbb{P}[-0.50 < Z < 2.00] \\
&= \mathbb{P}[-0.50 < Z \leq 2.00] \\
&= \mathbb{P}[Z \leq 2.00] - \mathbb{P}[Z \leq -0.50] \\
&= \Phi(2.00) - \Phi(-0.50) \\
&= 0.9772 - 0.3085 = 0.6687.
\end{aligned}$$

Approximation de la loi binomiale par la loi normale

Dans le cas particulier où les variables aléatoires $X_1, X_2, X_3, ...$ sont des Bernoulli(p) indépendantes les unes des autres, la distribution de la variable aléatoire $S_n = \sum_{j=1}^{n} X_j$ est la loi binomiale(n, p). On a donc, pour $k \in \{0, 1, 2, ..., n\}$

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

et, plus généralement,

$$\mathbb{P}[j \leq S_n \leq \ell] = \sum_{k=j}^{\ell} \binom{n}{k} p^k (1-p)^{n-k}.$$

Lorsque n est assez grand, le théorème limite central nous permet d'approximer cette probabilité binomiale :

$$\mathbb{P}[j \leq S_n \leq \ell] \approx \Phi\left(\frac{\ell - n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{j - n\mu}{\sqrt{n}\sigma}\right).$$

La moyenne et l'écart-type de la loi de Bernoulli étant $\mu = p$ et $\sigma = \sqrt{p(1-p)}$, cette dernière équation peut s'écrire sous la forme suivante :

$$\mathbb{P}[j \leq S_n \leq \ell] \approx \Phi\left(\frac{\ell - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{j - np}{\sqrt{np(1-p)}}\right).$$

Puisque la variable S_n est à valeurs entières, on peut améliorer cette approximation en utilisant la *correction pour la continuité* :

$$\mathbb{P}[j \leq S_n \leq \ell] \approx \Phi\left(\frac{(\ell + 1/2) - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{(j - 1/2) - np}{\sqrt{np(1-p)}}\right).$$

Exemple 31. On lance un dé 600 fois. Calculez une approximation pour la probabilité d'obtenir entre 90 et 105 fois (inclusivement) la valeur six.

Solution. Si N dénote le nombre de fois qu'on obtient la valeur six parmi nos 600 lancers, alors on a $N \sim$ binomiale$(600, 1/6)$. On veut $\mathbb{P}[90 \leq N \leq 105]$. L'équation précédente, avec $np = 100$ et $np(1-p) = 500/6$, nous donne

$$\mathbb{P}[90 \leq N \leq 105] \approx \Phi\left(\frac{(105 + 1/2) - 100}{\sqrt{500/6}}\right) - \Phi\left(\frac{(90 - 1/2) - 100}{\sqrt{500/6}}\right) \approx 0.6016.$$

La réponse exacte (obtenue avec l'aide du logiciel R) est

$$\mathbb{P}[90 \leq N \leq 105] = \sum_{k=90}^{105} \binom{600}{k} (1/6)^k (5/6)^{600-k} \approx 0.6050.$$

2.7 Le cas multidimensionnel[5]

Un vecteur aléatoire de dimension n est un vecteur

$$\mathbb{X} = (X_1, X_2, ..., X_n)$$

dont chacune des coordonnées X_j est une variable aléatoire. Nous allons examiner le cas $n = 2$. Le cas $n > 2$ se traite essentiellement de la même façon. Pour le cas $n = 2$, nous allons suivre la tradition et utiliser la notation (X, Y) plutôt que la notation (X_1, X_2).

Exemple 32. Une urne contient 7 boules bleues, 8 boules blanches et 10 boules rouges. On fait 5 tirages au hasard et sans remise à partir de cette urne et on considère les variables aléatoires suivantes :

$$\begin{aligned} X &= \text{« le nombre de boules bleues parmi les 5 boules tirées »}, \\ Y &= \text{« le nombre de boules blanches parmi les 5 boules tirées »}. \end{aligned}$$

Dans cet exemple on pourrait bien sûr étudier séparément la distribution de la variable X et la distribution de la variable Y à l'aide des outils qui ont été présentés dans les sections précédentes. Toutefois, cette approche ne nous permettrait pas de comprendre et de décrire le lien entre la variable X et la variable Y. Plutôt que de considérer les variables X et Y comme étant deux points sur la droite $\mathbb{R}$, nous allons considérer le couple (X, Y) comme étant un point dans le plan $\mathbb{R}^2$ et nous allons étudier la distribution de probabilité *conjointe* des variables X et Y, c'est-à-dire la distribution de probabilité du point aléatoire (X, Y) dans le plan $\mathbb{R}^2$. Comme dans le cas unidimensionnel, nous allons considérer séparément le cas discret et le cas absolument continu.

Le cas discret. On est dans le cas discret lorsque l'ensemble des valeurs possibles du point aléatoire (X, Y) est ou bien un ensemble fini, ou bien un ensemble infini dénombrable. Dans ce cas, nous décrivons la *distribution conjointe de X et Y* à l'aide de la *fonction de probabilité conjointe* $p(x, y)$ définie par l'équation

$$p(x, y) = \mathbb{P}[(X, Y) = (x, y)] = \mathbb{P}[(X = x) \cap (Y = y)].$$

Il est facile de voir que cette fonction satisfait les conditions suivantes :

(a) $p(x, y) \geq 0$ pour tout $(x, y) \in \mathbb{R}^2$,

(b) $\displaystyle\sum_{(x,y)\in\mathbb{R}^2} p(x, y) = 1.$

De plus, pour tout $B \subset \mathbb{R}^2$ on a

$$\mathbb{P}[(X, Y) \in B] = \sum_{(x,y)\in B} p(x, y). \tag{2.24}$$

Ces trois équations sont les analogues des équations (2.5), (2.6) et (2.7).

[5] On peut omettre cette section si on manque de temps

Revenons à notre exemple. On voit, par inspection, que les valeurs possibles du vecteur aléatoire (X, Y) sont tous les couples d'entiers (x, y) avec $x \geq 0$, $y \geq 0$ et $x + y \leq 5$. Voici, sous forme de tableau, la liste de ces valeurs possibles :

$$\begin{array}{llllll} (0,0) & (0,1) & (0,2) & (0,3) & (0,4) & (0,5) \\ (1,0) & (1,1) & (1,2) & (1,3) & (1,4) & \\ (2,0) & (2,1) & (2,2) & (2,3) & & \\ (3,0) & (3,1) & (3,2) & & & \\ (4,0) & (4,1) & & & & \\ (5,0) & & & & & \end{array}$$

Calculons la fonction de probabilité conjointe des variables X et Y. Notre urne contient en tout 25 boules et on fait 5 tirages sans remise. Il y a en tout $\binom{25}{5}$ façons différentes de tirer 5 boules à partir de cette urne. Nous sommes dans le cas équiprobable : ces $\binom{25}{5}$ résultats possibles ont tous la même probabilité. Fixons x et y, des entiers non négatifs dont la somme est plus petite ou égale à 5. Puisque l'urne contient 7 boules bleues, 8 boules blanches et 10 boules rouges, le nombre de façons différentes de choisir 5 boules de manière à en avoir x bleues et y blanches, et forcément $5-x-y$ rouges, est $\binom{7}{x}\binom{8}{y}\binom{10}{5-x-y}$. La fonction de probabilité conjointe est donc donnée par

$$p(x,y) = \begin{cases} \dfrac{\binom{7}{x}\binom{8}{y}\binom{10}{5-x-y}}{\binom{25}{5}} & \text{si } x \text{ et } y \text{ sont des entiers non-négatifs tels que } x+y \leq 5, \\ 0 & \text{sinon.} \end{cases}$$

Souvent on présente la fonction de probabilité conjointe sous forme d'un tableau. Avec notre exemple, on obtient le tableau suivant.

Y (boules blanches)

X (boules bleues)	0	1	2	3	4	5
0	0.0047	0.0316	0.0632	0.0474	0.0132	0.0011
1	0.0277	0.1265	0.1660	0.0738	0.0092	0
2	0.0474	0.1423	0.1107	0.0221	0	0
3	0.0296	0.0527	0.0184	0	0	0
4	0.0066	0.0053	0	0	0	0
5	0.0004	0	0	0	0	0

L'étudiant peut vérifier que cette fonction satisfait bel et bien les conditions (a) et (b) données plus haut. Supposons maintenant qu'on veuille calculer la probabilité d'obtenir

au moins deux boules bleues et au moins deux boules blanches. À l'aide de l'équation (2.24) on obtient

$$\begin{aligned}\mathbb{P}[(X \geq 2) \cap (Y \geq 2)] &= \mathbb{P}[(X,Y) \in \{(2,2),(3,2),(2,3)\}] \\ &= p(2,2) + p(3,2) + p(2,3) = 0.1512.\end{aligned}$$

Le cas continu. Le cas continu se traite de façon analogue. On dit que la distribution conjointe des variables X et Y est continue s'il existe une fonction $f(x,y)$ telle que

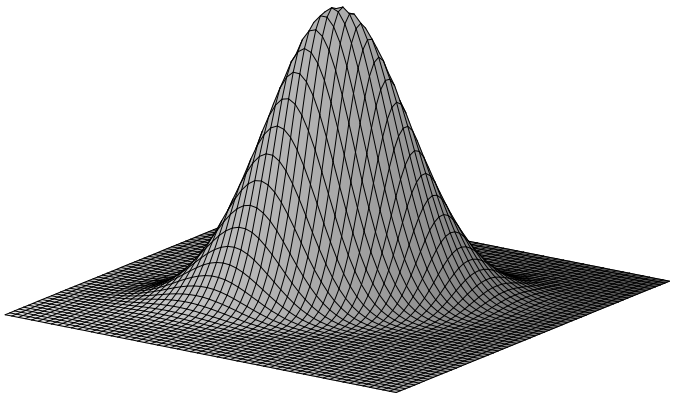
$$\mathbb{P}[(X,Y) \in B] = \int\int_B f(x,y)\,dx\,dy \tag{2.25}$$

pour tout B dans $\mathbb{R}^2$. La probabilité $\mathbb{P}[(X,Y) \in B]$ est donc donnée par le volume sous le graphe de la fonction $f(x,y)$ au dessus de la région B. Cette fonction $f(x,y)$ s'appelle la densité de probabilité conjointe des variables X et Y et elle peut toujours être choisie de façon à satisfaire les conditions suivantes :

(1) $f(x,y) \geq 0$ pour tout $(x,y) \in \mathbb{R}^2$,

(2) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$

Un exemple important de densité de probabilité conjointe est la loi normale bivariée. Voici de quoi elle a l'air :



Lois marginales

Considérons le scénario suivant : on connaît la fonction de probabilité conjointe $p(x,y)$ des variables aléatoires X et Y et on s'intéresse uniquement à la variable aléatoire X. On aimerait donc connaître la fonction de probabilité $p_X(x)$ de la variable aléatoire X. Cette fonction $p_X(x)$ peut être calculée à partir de la fonction de probabilité conjointe $p(x,y)$ de la façon suivante :

$$\begin{aligned} p_X(x) &= \mathbb{P}[X = x] \\ &= \sum_{y \in \mathbb{R}} \mathbb{P}[(X = x) \cap (Y = y)] \\ &= \sum_{y \in \mathbb{R}} p(x, y). \end{aligned}$$

Comme on est dans un contexte bidimensionnel, la distribution de X est appelée la distribution *marginale* de X et la fonction de probabilité de X est appelée fonction de probabilité *marginale* de X. De la même façon on peut calculer la fonction de probabilité marginale de Y. En résumé, les fonctions de probabilité marginales de X et de Y sont données par

$$\begin{aligned} p_X(x) &= \sum_{y \in \mathbb{R}} p(x, y), \\ p_Y(y) &= \sum_{x \in \mathbb{R}} p(x, y). \end{aligned}$$

Dans le cas où la fonction de probabilité conjointe est présentée sous forme d'un tableau, les fonctions de probabilité marginales peuvent être données dans les *marges* de ce tableau. Pour l'exemple ci-dessus, on obtient le tableau suivant. La marge de droite nous donne la fonction de probabilité marginale de X ; les probabilités indiquées dans cette marge sont les sommes des probabilités apparaissant sur les lignes correspondantes. La marge inférieure nous donne la fonction de probabilité marginale de Y ; les probabilités indiquées dans cette marge sont les sommes des probabilités apparaissant sur les colonnes correspondantes.

	0	1	2	3	4	5	$p_X(x)$
0	0.0047	0.0316	0.0632	0.0474	0.0132	0.0011	0.1613
1	0.0277	0.1265	0.1660	0.0738	0.0092	0	0.4032
2	0.0474	0.1423	0.1107	0.0221	0	0	0.3225
3	0.0296	0.0527	0.0184	0	0	0	0.1008
4	0.0066	0.0053	0	0	0	0	0.0119
5	0.0004	0	0	0	0	0	0.0004
$p_Y(y)$	0.1165	0.3584	0.3584	0.1433	0.0224	0.0011	1

Le cas continu se traite de la même façon. Si $f(x, y)$ dénote la densité conjointe des variables aléatoires X et Y, alors les densités marginales de X et de Y sont données par

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy, \\
f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx.
\end{aligned}
$$

Lois conditionnelles

Considérons à nouveau l'exemple ci-dessus. On a fait les 5 tirages et on vous apprend que parmi les 5 boules tirées il y exactement 2 boules bleues. Autrement dit, on vous informe que $X = 2$. Étant donnée cette information, quelle est la probabilité qu'on n'ait tiré aucune boule blanche ? Une boule blanche ? Deux boules blanches ? Trois boules blanches ? On peut calculer ces probabilités conditionnelles à partir de la fonction de probabilité conjointe de X et Y et de la fonction de probabilité marginale de X de la façon suivante :

$$
\mathbb{P}[Y = y | X = 2] = \frac{\mathbb{P}[(X = 2) \cap (Y = y)]}{\mathbb{P}[X = 2]} = \frac{p(2,y)}{p_X(2)}.
$$

Vue comme fonction de y, la probabilité conditionnelle $\mathbb{P}[Y = y|X = 2]$ s'appelle la *fonction de probabilité conditionnelle de Y sachant que $X = 2$* et est dénotée $p_{Y|X=2}(y)$. Plus généralement, la fonction de probabilité conditionnelle de Y sachant que $X = x$ et la fonction de probabilité conditionnelle de X sachant que $Y = y$ sont données par

$$
\begin{aligned}
p_{Y|X=x}(y) &= \frac{p(x,y)}{p_X(x)}, \\
p_{X|Y=y}(x) &= \frac{p(x,y)}{p_Y(y)}.
\end{aligned}
$$

Dans le cas continu, on a essentiellement les même équations :

$$
\begin{aligned}
f_{Y|X=x}(y) &= \frac{f(x,y)}{f_X(x)}, \\
f_{X|Y=y}(x) &= \frac{f(x,y)}{f_Y(y)}.
\end{aligned}
$$

Variables aléatoires indépendantes

On a vu à la section 2.4.4 que les événements A et B sont dits *indépendants* si on a $\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B]$. La définition suivante est donc tout à fait naturelle.

Définition. Les variables aléatoires X et Y sont dites indépendantes si pour tout choix de sous-ensembles de $\mathbb{R}$, disons C et D, les événements $X \in C$ et $Y \in D$ sont indépendants. Autrement dit, les variables aléatoires X et Y sont indépendantes si on a

$$
\mathbb{P}[(X \in C) \cap (Y \in D)] = \mathbb{P}[X \in C] \times \mathbb{P}[Y \in D]
$$

pour tout $C \subset \mathbb{R}$ et $D \subset \mathbb{R}$.

Si X et Y ne sont pas indépendantes, alors on dit qu'elles sont dépendantes. Dans l'exemple ci-dessus, les variables X et Y sont dépendantes. On peut montrer que dans le cas discret la définition précédente peut aussi s'écrire sous la forme suivante.

INDÉPENDANCE DE VARIABLES ALÉATOIRES : LE CAS DISCRET. Soient X et Y, des variables aléatoires avec fonction de probabilité conjointe $p(x, y)$. Alors X et Y sont indépendantes si et seulement si leur fonction de probabilité conjointe est égale au produit de leurs fonctions de probabilité marginales :

$$p(x, y) = p_X(x) \cdot p_Y(y) \qquad \text{pour tout } (x, y) \in \mathbb{R}^2.$$

De même, on peut montrer que dans le cas continu la définition précédente peut aussi s'écrire sous la forme suivante.

INDÉPENDANCE DE VARIABLES ALÉATOIRES : LE CAS CONTINU. Soient X et Y, des variables aléatoires avec densité de probabilité conjointe $f(x, y)$. Alors X et Y sont indépendantes si et seulement si leur densité de probabilité conjointe est égale au produit de leurs densités de probabilité marginales :

$$f(x, y) = f_X(x) \cdot f_Y(y) \qquad \text{pour tout } (x, y) \in \mathbb{R}^2.$$

Dans la pratique, c'est presque toujours le contexte qui nous dit si des variables aléatoires sont indépendantes ou non. Voici un exemple. On fait deux tirages à partir d'un panier contenant 10 boules numérotées 1 à 10 et on pose

$$\begin{aligned} X &= \text{le numéro de la boule obtenue au premier tirage} \\ Y &= \text{le numéro de la boule obtenue au deuxième tirage.} \end{aligned}$$

Si les tirages sont fait avec remise, alors X et Y sont des variables indépendantes. Si les tirages sont fait sans remise, alors X et Y sont des variables dépendantes.

La covariance

Considérons un couple de variables aléatoires, disons X et Y, avec fonction de probabilité conjointe $p(x, y)$ ou avec densité conjointe $f(x, y)$. La covariance de X et Y est définie par

$$\begin{aligned} \mathbb{C}\text{ov}[X, Y] &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \begin{cases} \displaystyle\sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} (x - \mu_X)(y - \mu_Y)\, p_{X,Y}(x, y) & \text{dans le cas discret,} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, f_{X,Y}(x, y)\, dx\, dy & \text{dans le cas continu.} \end{cases} \end{aligned}$$

INTERPRÉTATION DE LA COVARIANCE. La covariance de X et Y (on dit aussi covariance *entre* X et Y) est une mesure du degré d'association linéaire entre X et Y. Une covariance positive nous indique que plus la variable X est grande, plus la variable Y a tendance à

être grande et plus la variable X est petite, plus la variable Y a tendance à être petite. On dit alors qu'il y a une association linéaire positive entre X et Y. Avec une covariance négative, c'est l'inverse : plus la variable X est grande, plus la variable Y a tendance à être petite et plus la variable X est petite, plus la variable Y a tendance à être grande. On dit alors qu'il y a une association linéaire négative entre X et Y. Pour voir d'où vient cette interprétation, imaginez une droite verticale et une droite horizontale passant toutes les deux par le point (μ_X, μ_Y) dans le plan et considérons les 4 quadrants autour de ce point, numérotés de la manière usuelle, c'est-à-dire en partant du quadrant supérieur droit et en allant dans le sens anti-horaire. Une association linéaire positive correspond à la situation où la distribution conjointe de X et Y est plutôt concentrée sur les quadrants 1 et 3 autour de (μ_X, μ_Y). Or, pour tout (x, y) dans ces deux quadrants on a $(x-\mu_X)(y-\mu_Y) > 0$. Donc, dans le cas d'une association linéaire positive on obtient

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] > 0.$$

Une association linéaire négative correspond à la situation où la distribution conjointe de X et Y est plutôt concentrée sur les quadrants 2 et 4 autour de (μ_X, μ_Y). Or, pour tout (x, y) dans ces deux quadrants on a $(x - \mu_X)(y - \mu_Y) < 0$. Donc, dans le cas d'une association linéaire négative on obtient

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] < 0.$$

Calcul de covariance. Comme pour la variance, il existe un raccourci pour calculer la covariance :

$$\begin{aligned}
\mathbb{C}\text{ov}[X, Y] &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathbb{E}[XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mu_Y - \mu_X \mathbb{E}[Y] + \mu_X \mu_Y \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\ \mathbb{E}[Y].
\end{aligned}$$

On a donc

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\ \mathbb{E}[Y].$$

Reprenons l'exemple ci-dessus. Une urne contient 7 boules bleues, 8 boules blanches et 10 boules rouges. On fait 5 tirages au hasard et sans remise à partir de cette urne et on considère les variables aléatoires suivantes :

$$\begin{aligned}
X &= \text{« le nombre de boules bleues parmi les 5 boules tirées »} \\
Y &= \text{« le nombre de boules blanches parmi les 5 boules tirées »}.
\end{aligned}$$

À partir de la fonction de probabilité conjointe de X et Y, on obtient

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_X \sum_Y xy\, p_{X,Y}(x, y) \\
&= \sum_{k=0}^{5} \sum_{\ell=0}^{5} k\ell \frac{\binom{7}{k}\binom{8}{\ell}\binom{10}{5-k-\ell}}{\binom{25}{5}} \\
&\approx 1.408.
\end{aligned}$$

En faisant ce calcul, l'étudiant notera que seulement 10 des 36 termes

$$k\ \ell\ p_{X,Y}(k,\ell)$$

sont positifs. Les 26 autres sont nuls parce que ou bien k est nul, ou bien ℓ est nul, ou bien $p_{X,Y}(k,\ell)$ est nul. À partir des fonctions de probabilité marginales on obtient

$$\mathbb{E}[X] = 1.400 \quad \text{et} \quad \mathbb{E}[Y] = 1.600.$$

On a donc

$$\mathbb{C}\text{ov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\ \mathbb{E}[Y] \approx 1.408 - 1.4 \cdot 1.6 = -0.832.$$

Dans cet exemple, il est intuitivement clair qu'il y a une association négative entre X et Y : plus la variable X est grande, c'est-à dire plus le nombre de boules bleues est grand, plus la variable Y a tendance à être petite, c'est-à-dire plus le nombre de boules blanches a tendance à être petit. Le résultat $\mathbb{C}\text{ov}[X,Y] \approx -0.832$ est cohérent avec notre intuition.

On peut montrer que si X et Y sont des variables aléatoires indépendantes, alors on a $\mathbb{C}\text{ov}[X,Y] = 0$. L'inverse n'est pas vrai. Il est possible d'avoir $\mathbb{C}\text{ov}[X,Y] = 0$ même avec des variables X et Y dépendantes.

Le coefficient de corrélation

Pour mesurer le degré d'association linéaire entre X et Y, les statisticiens préfèrent utiliser le coefficient de corrélation. Ce dernier est dénoté $\rho_{X,Y}$ et est défini par l'équation suivante :

$$\rho_{X,Y} = \frac{\mathbb{C}\text{ov}[X,Y]}{\sigma_X\ \sigma_Y}.$$

Le coefficient de corrélation est simplement une version normalisée de la covariance. Contrairement à la covariance, dont les unités sont celles de X multipliées par celles de Y, le coefficient de corrération est un nombre pur, c'est-à-dire un nombre sans unités de mesure. Par exemple, si la variable X est exprimée en cm et la variable Y en kg, alors la covariance entre X et Y est exprimée en cm-kg. Lorsqu'on divise cette covariance par l'écart-type σ_X (exprimé en cm) et l'écart-type σ_Y (exprimé en kg), on obtient un nombre pur.

Les principales propriétés de la covariance et du coefficient de corrélation sont regroupées dans la liste suivante :

(1) $\mathbb{C}\text{ov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

(2) $\mathbb{V}\text{ar}[X+Y] = \mathbb{V}\text{ar}[X] + 2\mathbb{C}\text{ov}[X,Y] + \mathbb{V}\text{ar}[Y]$.

(3) $-1 \leq \rho_{X,Y} \leq 1$.

(4) $\rho_{X,Y} = 1$ si et seulement si $Y = aX + b$ pour un certain $a > 0$ et un certain $b \in \mathbb{R}$.

(5) $\rho_{X,Y} = -1$ si et seulement si $Y = aX + b$ pour un certain $a < 0$ et un certain $b \in \mathbb{R}$.

(6) $\mathbb{C}\text{ov}[aX+b, cY+d] = a\,c\,\mathbb{C}\text{ov}[X,Y]$ pour tout a, b, c et d.

(7) Si X et Y sont indépendantes, alors $\rho_{X,Y} = \mathbb{C}\text{ov}[X,Y] = 0$;

(8) Si X et Y sont indépendantes, alors $\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]$;

Un regard sur la suite...

Au Chapitre 1, nous avons examiné, entre autres choses, les concepts de moyenne échantillonnale $\overline{x}$, d'écart-type échantillonnal s et de coefficient de corrélation échantillonnal r. Au Chapitre 2, nous avons examiné, entre autres choses, les concepts de moyenne μ d'une variable aléatoire, d'écart-type σ d'une variable aléatoire et de coefficient de corrélation ρ entre deux variables aléatoires. Les chapitres qui vont suivre vont nous permettre de d'éclaircir les liens qui existent entre ces concepts du Chapitre 1 et ceux du Chapitre 2.

2.8 Exercices

Numéro 1. On lance un dé à deux reprises. Calculez

(a) la probabilité d'obtenir deux nombres pairs ;

(b) la probabilité que la somme des deux résultats soit un nombre pair ;

(c) la probabilité que le résultat du premier lancer soit inférieur au résultat du deuxième lancer.

Suggestion : Procédez comme à l'exemple 1 de la section 2.1.

Numéro 2. On lance une pièce de monnaie cinq fois. Calculez

(a) la probabilité d'obtenir deux faces et trois piles ;

(b) la probabilité d'obtenir cinq résultats identiques ;

(c) la probabilité d'obtenir au moins deux piles.

Suggestion : Procédez comme à l'exemple 2 de la section 2.1.

Numéro 3. On lance un dé quatre fois.

(a) Décrivez l'ensemble Ω de tous les résultats possibles.

(b) Quel est le cardinal de Ω ?

(c) Sommes-nous dans le cas équiprobable ? Expliquez.

Numéro 4. On lance un dé quatre fois.

(a) Calculez la probabilité d'obtenir 4 fois le même nombre.

(b) Calculez la probabilité d'obtenir 4 nombres différents.

(c) Calculez la probabilité d'obtenir seulement des nombres pairs.

(d) Calculez la probabilité d'obtenir aucun six.

(e) Calculez la probabilité que la somme des 4 nombres obtenus soit plus grande ou égale à 22.

Numéro 5. Parmi les propriétés 1 à 6 présentées à la section 2.2, lesquelles correspondent aux scénarios suivants ?

(a) Selon le ministère du revenu du Québec, 25% des couples mariés n'ont pas d'enfants, 22% ont une seul enfant et 17% ont deux enfants. On peut donc dire que 64% des couples mariés ont au plus 2 enfants.

(b) Á Québec, 35% des adultes lisent le journal Le Soleil, 30% lisent le Journal de Québec et 5% lisent ces deux journaux. On peut donc dire que 60% des adultes vivant à Québec lisent au moins un de ces deux journaux.

(c) L'an passé, 10% des étudiants de STT-19909 ont échoué le cours. On peut donc dire que 90% des étudiants ont réussi le cours.

Numéro 6.

(a) Avec les chiffres 0 et 1, on peut former 8 séquences de longueur 3 : 000, 001, 010, 011, 100, 101, 110, 111. Combien de séquences de longueur 30 peut-on former ?

(b) Combien de mots de longueur 100 peut-on former avec les lettres A, T, C, G ?

(c) Combien de mots de longueur n peut-on former avec un alphabet de m lettres ?

Numéro 7.

(a) Combien de séquences de longueur 30 peut-on former en utilisant 12 fois le chiffre 0 et 18 fois le chiffre 1 ?

(b) Combien de mots de longueur 30 peut-on former avec en utilisant 7 fois la lettre A, 4 fois la lettre T, 6 fois la lettre C et 13 fois la lettre G ?

Suggestion : Utilisez les coefficients binomiaux. Pour la partie (b), imaginez 30 cases (numérotées 1 à 30), choisissez les 7 cases dans lesquelles vous mettrez les A, puis choisissez parmi les 23 cases restantes les 4 cases dans lesquelles vous mettrez les T, etc.

Numéro 8. Trois filles (Annette, Bernadette et Claudette) et trois garçons (Damien, Émilien et Fabien) font la file devant un comptoir de crème glacée.

(a) Il y a combien d'arrangements possibles ?

(b) Il y a combien d'arrangements possibles si les trois filles doivent être ensemble et les trois garçons doivent être ensemble ?

(c) Il y a combien d'arrangements possibles si on exige seulement que les trois filles soient ensemble ?

Numéro 9.

(a) Avec un billet de 6/49, quelle est la probabilité d'obtenir exactement 4 bons nombres ?

(b) Vous achetez un billet de 6/49 à chaque semaine pendant 52 semaines. Quelle est la probabilité de ne jamais avoir plus que deux bons nombres ?

Numéro 10. On considère les 5 prochaines naissances qui auront lieu à l'hôpital St-Sacrement. Calculez les probabilités suivantes. Sous quelles conditions vos calculs sont-ils valides ?

(a) Quelle est la probabilité que les 5 bébés seront tous du même sexe ?

(b) Quelle est la probabilité qu'il y aura exactement 3 garçons et 2 filles ?

(c) Quelle est la probabilité qu'il y aura au moins un garçon ?

(d) Quelle est la probabilité qu'il y aura plus de filles que de garçons ?

Numéro 11. Un lot de 200 jeunes plants de tomates contient 190 plants en santé et 10 plants atteints d'une maladie asymptomatique détectable seulement à l'aide d'un test d'ADN.

(a) On choisit cinq plants au hasard. Quelle est la probabilité qu'il y ait au moins un plant malade parmi ces cinq plants ?

(b) Combien de plants doit-on choisir si on veut être sûr à 95% d'obtenir au moins un des 10 plants malades ?

Suggestion pour (b) : Refaites la partie (a) avec *cinq* remplacé par n, puis déterminez le plus petit n pour lequel la probabilité est supérieure à 0.95.

Numéro 12. Un container contient quelques millions de graines de soya. Supposons qu'exactement 20% de ces graines soient contaminées.

(a) On obtient 4 graines au hasard à partir de ce container. Quelle est la probabilité qu'aucune de ces graines ne soit contaminée ?

(b) On obtient 4 graines au hasard à partir de ce container. Quelle est la probabilité qu'exactement une de ces graines soit contaminée ?

(c) On obtient 15 graines au hasard à partir de ce container. Quelle est la probabilité qu'au moins une de ces graines soit contaminée ?

Numéro 13. Un panier contient 7 boules noires et 13 boules blanches. On fait 4 tirages sans remise. Calculez

(a) la probabilité d'obtenir, dans l'ordre, blanc, blanc, noir, blanc ;

(b) la probabilité d'obtenir quatre fois la même couleur ;

(c) la probabilité d'obtenir deux boules blanches et deux boules noires.

Numéro 14. Un panier contient 7 boules noires et 13 boules blanches. On fait 4 tirages avec remise. Calculez

(a) la probabilité d'obtenir, dans l'ordre, blanc, blanc, noir, blanc ;

(b) la probabilité d'obtenir quatre fois la même couleur ;

(c) la probabilité d'obtenir deux boules blanches et deux boules noires.

Numéro 15.

(a) Un panier contient 5 boules rouges et 4 boules blanches. On fait 3 tirages **avec** remise. Quelle est la probabilité d'obtenir la séquence *rouge-rouge-blanche* ?

(b) Un panier contient 5 boules rouges et 4 boules blanches. On fait 3 tirages **sans** remise. Quelle est la probabilité d'obtenir la séquence *rouge-rouge-blanche* ?

(c) Un panier contient 5 millions de boules rouges et 4 millions de boules blanches. On fait 3 tirages **avec** remise. Quelle est la probabilité d'obtenir la séquence *rouge-rouge-blanche* ?

(d) Un panier contient 5 millions de boules rouges et 4 millions de boules blanches. On fait 3 tirages **sans** remise. Quelle est la probabilité d'obtenir la séquence *rouge-rouge-blanche* ?

Numéro 16. Un jeu de cartes ordinaire comprend 52 cartes. Chacune de ces 52 cartes appartient à une **couleur** et possède une **valeur**. Les couleurs sont le carreau, le coeur, le trèfle et le pique. Les valeurs sont les 2, 3, 4, 5, 6, 7, 8, 9, 10, J (valet), Q (dame), K (roi) et A (as). Pour les questions qui suivent, on considère une main de poker, c'est-à-dire une combinaison de 5 cartes tirées au hasard à partir d'un jeu de 52 cartes.

(a) Calculez la probabilité d'obtenir une *paire*, c'est-à-dire une main de poker contenant en tout 4 valeurs différentes. (Il faut une paire, c'est-à-dire 2 cartes de même valeur, et les 3 autres cartes doivent être de valeurs différentes entre elles et différentes de la valeur des cartes formant la paire). Exemple d'une paire : 10 de coeur, 10 de trèfle, J de pique, 6 de carreau et 7 de coeur.

(b) Calculez la probabilité d'obtenir *deux paires*. (Les deux paires ne peuvent pas avoir la même valeur et la valeur de la cinquième cartes doit être différente des valeurs des deux paires). Exemple de deux paires : 5 de coeur, 5 de trèfle, J de pique, J de carreau et 7 de coeur.

(c) Calculez la probabilité d'obtenir un *brelan*, c'est-à-dire une main de poker contenant trois cartes de la même valeur. (Les deux autres cartes doivent être de valeurs différentes entre elles et différentes de la valeur commune aux trois premières cartes). Exemple d'un brelan : 5 de coeur, 5 de trèfle, 5 de pique, 3 de carreau et K de coeur.

(d) Calculez la probabilité d'obtenir une *main pleine*, c'est-à-dire une main de poker contenant trois cartes d'une valeur et deux cartes d'une autre valeur. Exemple d'une main pleine : 5 de coeur, 5 de trèfle, 5 de pique, 9 de coeur et 9 de pique.

(e) Calculez la probabilité d'obtenir un *carré*, c'est-à-dire une main de poker contenant quatre cartes de la même valeur (et une cinquième carte quelconque). Exemple d'un carré : 5 de coeur, 5 de trèfle, 5 de pique, 5 de carreau et 9 de pique.

Numéro 17. Deux femmes et 14 hommes sont assis au hasard sur 16 chaises formant une ligne.

(a) Quelle est la probabilité que les deux femmes soient assises une à côté de l'autre ?

(b) Quelle est la probabilité que les deux femmes occupent les deux extrémités de la ligne ?

Numéro 18. Deux femmes et 14 hommes sont assis au hasard sur 16 chaises formant un cercle.

(a) Quelle est la probabilité que les deux femmes soient assises une à côté de l'autre ?

(b) Quelle est la probabilité que les deux femmes occupent deux chaises diamétralement opposées ?

Numéro 19. Un panier contient 4 boules rouges et 6 boules noires. Dans chacun des cas suivants, obtenez l'ensemble des valeurs possibles, la distribution, la moyenne et l'écart-type de la variable aléatoire X.

(a) On fait 7 tirages avec remise et on pose $X =$ « *le nombre de boules rouges parmi les 7 tirages* ».

(b) On fait 7 tirages sans remise et on pose $X =$ « *le nombre de boules rouges parmi les 7 tirages* ».

NUMÉRO 20. Un panier contient 4 boules rouges et 6 boules noires. Dans chacun des cas suivants, obtenez l'ensemble des valeurs possibles, la distribution, la moyenne et l'écart-type de la variable aléatoire X.

(a) On fait des tirages avec remise jusqu'à ce qu'on obtienne une boule rouge et on pose $X =$ « *le nombre de tirages nécessaires* ».

(b) On fait des tirages sans remise jusqu'à ce qu'on obtienne une boule rouge et on pose $X =$ « *le nombre de tirages nécessaires* ».

NUMÉRO 21. Un panier contient 5 boules rouges et 5 boules noires. Dans chacun des cas suivants, obtenez l'ensemble des valeurs possibles, la distribution, la moyenne et l'écart-type de la variable aléatoire X.

(a) On fait des tirages avec remise jusqu'à ce qu'on ait obtenu une boule rouge et une boule blanche. On pose $X =$ « *le nombre de tirages nécessaires* ».

(b) On fait des tirages sans remise jusqu'à ce qu'on ait obtenu une boule rouge et une boule blanche. On pose $X =$ « *le nombre de tirages nécessaires* ».

NUMÉRO 22.

(a) Un panier contient 5 boules rouges et 4 boules blanches. On fait 3 tirages **avec** remise. Déterminez la distribution de la variable aléatoire $X =$ *le nombre de boules rouges parmi les 3 tirages*.

(b) Un panier contient 5 boules rouges et 4 boules blanches. On fait 3 tirages **sans** remise. Déterminez la distribution de la variable aléatoire $X =$ *le nombre de boules rouges parmi les 3 tirages*.

NUMÉRO 23. Voici la densité de probabilité d'une certaine variable aléatoire X :

$$f(x) = \begin{cases} h \times (1-(x-2)^2) & \text{si } 1 \leq x \leq 3 \\ 0 & \text{sinon.} \end{cases}$$

(a) Déterminez la valeur de la constante positive h.
Indice : Il faut que la surface sous la courbe soit égale à 1.

(b) Dessinez le graphe de cette densité.

(c) Calculez $\mathbb{P}[X \geq 5/2]$.

NUMÉRO 24. Voici la densité de probabilité d'une certaine variable aléatoire Y :

$$f(y) = \begin{cases} 0 & \text{si } y < 0 \\ y/25 & \text{si } 0 \leq y \leq 5 \\ (10-y)/25 & \text{si } 5 < y \leq 10 \\ 0 & \text{si } y > 10. \end{cases}$$

(a) Dessinez le graphe de cette densité.

(b) Calculez le 30^e centile de cette distribution. Autrement dit, trouvez la valeur y_* qui est telle que $\mathbb{P}[Y \leq y_*] = 0.30$.

(c) Calculez $\mathbb{P}[3 < Y < 5]$.

NUMÉRO 25. Supposons que $Z \sim N(0, 1)$.

(a) Obtenez $\mathbb{P}[-2/5 < Z < 6/5]$.

(b) Obtenez $\mathbb{P}[Z > -3/5]$.

(c) Obtenez le 90^e centile de cette distribution. Autrement dit, trouvez la valeur z_* qui est telle que $\mathbb{P}[Z \leq z_*] = 0.90$.

NUMÉRO 26. Supposons que $X \sim N(45, 25)$.

(a) Obtenez $\mathbb{P}[46 < X < 52]$.

(b) Obtenez $\mathbb{P}[X > 53]$.

(c) Obtenez le troisième quartile de cette distribution.

NUMÉRO 27. La loi normale avec moyenne 75 kg et écart-type 8 kg est un bon modèle pour décrire la distribution des poids des individus d'une certaine population. On choisit 16 individus au hasard à partir de cette population et on considère la variable $N =$ « *le nombre d'individus pesant plus de 83 kg parmi les 16 choisis* ».

(a) Obtenez $\mathbb{P}[N \geq 4]$.

(b) Obtenez l'espérance de N.

(c) Obtenez l'écart-type de N.

NUMÉRO 28. Un panier contient 4 boules rouges et 6 boules noires. On fait 200 tirages avec remise et on considère la variable aléatoire $X =$ « *le nombre de fois qu'on obtient une boule rouge* ».

(a) Quelle est la distribution exacte de la variable X ?

(b) À l'aide du théorème limite central, calculez une approximation pour $\mathbb{P}[85 \leq X \leq 95]$.

(c) Comparez la réponse obtenue à la partie (b) avec la réponse exacte que vous obtiendrez avec le logiciel R.

NUMÉRO 29. La loi de Poisson avec moyenne 3 est un bon modèle pour le nombre de fautes d'orthographe que Sophie fait lorsqu'elle écrit une page de texte.

(a) Calculez la probabilité que Sophie ne fasse aucune faute dans la prochaine page qu'elle écrira.

(b) Calculez la probabilité que Sophie fasse au moins 3 fautes dans la prochaine page qu'elle écrira.

(c) Calculez la probabilité que, parmi les 6 prochaines pages que Sophie écrira, il y en aura au moins trois qui contiendront au moins 3 fautes.

Numéro 30. Dans un champ récemment ensemancé, 15% des jeunes plants sont atteints d'une maladie génétique. On choisit 20 plants au hasard et on effectue, sur chacun de ces 20 plants, un test qui nous permet de détecter si le plant est malade ou non.

(a) Quelle est la distribution du nombre de plants malades parmi les 20 plants choisis ?
(b) Dessinez le graphe de cette distribution.
(c) Quelle est la probabilité qu'il n'y ait aucun plant malade parmi nos 20 plants ?
(d) Quelle est la probabilité qu'il y ait exactement 3 plants malades parmi nos 20 plants ?
(e) Quelle est la probabilité qu'il y ait au moins 6 plants malades parmi nos 20 plants ?
(f) Quelle est l'espérance du nombre de plants malades parmi nos 20 plants ?
(g) Quel est l'écart-type du nombre de plants malades parmi nos 20 plants ?

Numéro 31. On suppose que la loi normale avec moyenne 16 cm et avec écart-type 2 cm est un bon modèle pour décrire la distribution des hauteurs des plants dont il est question au numéro précédent. Dessinez le graphe de cette loi normale. Sur ce graphe, illustrez les réponses à chacune des questions suivantes.

(a) Quel pourcentage des plants ont une hauteur inférieure à 18 cm ?
(b) Quel pourcentage des plants ont une hauteur entre 15 et 18 cm ?
(c) Quel pourcentage des plants ont une hauteur supérieure à 20 cm ?
(d) Quel est le 80^e centile des hauteurs ? Autrement dit, trouvez la hauteur h qui est telle que 80% des plants sont de hauteur inférieure à h

Numéro 32. On choisit 30 plants au hasard à partir du champ dont il est question au numéro précédent. On dénote par N le nombre de plants dont la hauteur excède 18 cm.

(a) Quelle est la distribution de la variable aléatoire N ?
(b) Quelle est l'espérance de la variable aléatoire N ?
(c) Quel est l'écart-type de la variable aléatoire N ?
(d) Calculez $\mathbb{P}[4 \leq N \leq 9]$.

Numéro 33. Avec l'aide du logiciel R, examinez le graphe de la fonction de probabilité de la loi binomiale(20, p) pour chacune des valeurs suivantes de p : 0.05, 0.10, 0.15, 0.20, 0.25,..., 0.90, 0.95.

Numéro 34. On considère une certaine population de citrouilles. Le poids moyen de ces citrouilles est 4.25 kg. On sait que 33% de ces citrouilles pèsent plus que 4.47 kg. En supposant que la loi normale est un bon modèle pour décrire la distribution des poids des citrouilles, trouvez l'écart-type de cette loi normale.

Numéro 35. Avec l'aide du logiciel R, examinez le graphe de la fonction de probabilité de la loi de Poisson(λ) pour chacune des valeurs suivantes de λ : 2.0, 4.5, 9.0, 20.0.

Numéro 36. Lorsque n est grand et p est petit, la loi binomiale(n, p) ressemble beaucoup à la loi de Poisson(λ) avec $\lambda = np$. Avec l'aide du logiciel R, vérifiez cette affirmation dans le cas de la loi binomiale(30,0.05).

Numéro 37.

(a) Supposons que la loi de Poisson avec moyenne 3 soit un bon modèle pour décrire la distribution du nombre de limaces présentes sur un plant de tomate. On choisit un plan de tomate au hasard.

(i) Quelle est la probabilité qu'il y ait exactement 2 limaces sur ce plant ?

(ii) Quelle est la probabilité qu'il y ait au moins 4 limaces sur ce plant ?

(iii) Dessinez le graphe de la loi de Poisson avec moyenne 3 et illustrez sur ce graphe les réponses obtenues aux questions (i) et (ii) ci-dessus.

(b) On choisit 8 plants de tomates au hasard.

(i) Quelle est la probabilité d'obtenir au moins un plan de tomate avec au moins 4 limaces ?

(ii) Quelle est l'espérance du nombre total de limaces sur ces 8 plants de tomate ?

(iii) Quel est l'écart-type du nombre total de limaces sur ces 8 plants de tomate ?

Numéro 38. On considère 6 plants de tomate. On écrit X_j pour dénoter la production, mesurée en kg, du plant de tomate numéro j pour l'été 2009. On suppose que $X_1, X_2, ..., X_6$ sont des variables aléatoires indépendantes et identiquement distribuées, avec espérance 6.0 kg et avec écart-type 1.0 kg. Posons $T = X_1 + X_2 + \cdots + X_6$. Donc, T dénote la production totale pour ces 6 plants de tomates. Calculez l'espérance et l'écart-type de la variable aléatoire T.

Numéro 39. Dans chacun des scénarios suivants, déterminez si le coefficient de corrélation entre X et Y est positif ou négatif. Aucun calcul n'est nécessaire.

(a) On lance un dé trois fois et on considère les variables

$$\begin{aligned} X &= \text{le plus petit des trois résultats} \\ Y &= \text{le plus grand des trois résultats.} \end{aligned}$$

(b) On fait 6 tirages sans remise à partir d'un panier contenant 2 boules bleues, 2 boules blanches, 3 boules rouges et 3 boules noires. On considère

$$\begin{aligned} X &= \text{le nombre de boules rouges parmi les 6 tirages} \\ Y &= \text{le nombre de boules noires parmi les 6 tirages.} \end{aligned}$$

(c) On fait 6 tirages avec remise à partir d'un panier contenant 2 boules bleues, 2 boules blanches, 3 boules rouges et 3 boules noires. On considère

$$\begin{aligned} X &= \text{le nombre de boules rouges parmi les 6 tirages} \\ Y &= \text{le nombre de boules noires parmi les 6 tirages.} \end{aligned}$$

(d) On fait 3 tirages sans remise à partir d'un panier contenant 2 boules bleues et 3 boules rouges et on fait 3 tirages avec remise à partir d'un panier contenant 2 boules blanches et 3 boules noires. On considère

$$\begin{aligned} X &= \text{le nombre de boules rouges parmi les 6 tirages} \\ Y &= \text{le nombre de boules noires parmi les 6 tirages.} \end{aligned}$$

Numéro 40. Obtenez et présentez sous forme de tableau la distribution conjointe des variables aléatoires X et Y de la partie (b) du numéro précédent.

Numéro 41. On considère des variables aléatoires discrètes X et Y. Les valeurs possibles de X sont les entiers $0, 1$ et 2. Les valeurs possibles de Y sont les entiers $0, 1, 2$ et 3. La fonction de probabilité conjointe de ces deux variables est donnée par le tableau suivant :

	0	1	2	3
0	1/50	3/50	10/50	6/50
1	2/50	6/50	6/50	2/50
2	7/50	4/50	2/50	1/50

(a) Calculez $\mathbb{P}[Y \leq 2]$.

(b) Calculez $\mathbb{P}[X + Y \leq 2]$.

(c) Calculez la fonction de probabilité marginale de X et tracez son graphe.

(d) Calculez la fonction de probabilité marginale de Y et tracez son graphe.

(e) Obtenez la fonction de probabilité conditionnelle de Y sachant que $X = 2$.

(f) Calculez la covariance de X et Y.

(g) Calculez le coefficient de corrélation entre X et Y.

(h) Les variables X et Y sont-elles indépendantes ?

Numéro 42. Voici les résultats des loteries 6/49 et Québec 49 du 30 janvier 2008 :

Lotto 6/49 :	01	20	21	23	40	41
Lotto Québec 49 :	01	20	21	29	40	47

Les tirages de ces deux loteries ont lieu tous les mercredis et tous le samedis.

(a) Quelle est la probabilité que ces deux loteries auront exactement quatre nombres en commun lors de la prochaine journée de tirages ?

(b) Quelle est la probabilité que ces deux loteries auront exactement quatre nombres en commun au moins une fois au cours des sept prochaines années ?

Numéro 43.

(a) Un épicier reçoit une livraison de sacs de carottes de 1 kg. On choisit 25 sacs au hasard et on mesure les poids exacts de ces sacs. On pose $N =$ *le nombre de sacs dont le poids est inférier à 0.95 kg.* En supposant que douze pour cent des sacs de cette livraison pèsent moins que 0.95 kg, obtenez, avec R-Commander, les probabilités suivantes : $\mathbb{P}[N = 4]$, $\mathbb{P}[N \leq 5]$ et $\mathbb{P}[N \geq 9]$. Obtenez également le graphe de la distribution de la variable aléatoire N.

(b) Ce même épicier reçoit une livraison de sacs de pommes de 2 kg. En supposant que la loi normale avec moyenne 2.10 kg et avec écart-type 0.07 kg soit un bon modèle pour décrire la distribution des poids exacts de ces sacs, obtenez

(i) la proportion de sacs pesant moins de 2.00 kg,

(ii) la proportion de sacs pesant plus que 2.10 kg,

(iii) le 90^e centile de cette distribution,

(iv) le graphe de cette distribution.

Chapitre 3

Estimation et tests d'hypothèses : Problèmes à un échantillon

3.1 Théorie générale de l'estimation

3.1.1 Population et variable statistique

On considère une certaine population et on s'intéresse à une certaine variable statistique, c'est-à-dire une caractéristique dont la valeur numérique varie d'un membre de la population à l'autre. Les membres de la population, ou *individus*, peuvent être des êtres humains, des animaux, des objets, etc. Dans la plupart des exemples que nous considérons, les variables d'intérêt sont des variables quantitatives.

On suppose que la distribution de la variable d'intérêt est décrite par une certaine fonction de probabilité discrète ou une certaine densité de probabilité. Cette distribution de probabilité, appelée la distribution théorique, est souvent connue à une ou deux constantes près. Ces constantes sont appelées les paramètres du modèle et notre objectif est d'estimer ces paramètres.

Exemple 1 : La variable d'intérêt est une variable dichotomique. La distribution théorique est la loi de Bernoulli(p) et la proportion théorique p est inconnue.

Exemple 2 : La variable d'intérêt est une variable quantitative de type continu et la distribution théorique est la loi $N(\mu, \sigma^2)$. La moyenne théorique μ et la variance théorique σ^2 sont inconnues.

3.1.2 Échantillon aléatoire et statistique

Pour estimer le ou les paramètre(s) inconnu(s) (p, μ, σ^2,...), nous utilisons les observations, c'est-à-dire les données, que nous avons obtenues. Ces observations sont dénotées $x_1, x_2, x_3, ..., x_n$. Nous supposons que ces nombres sont les valeurs observées, c'est-à-dire les *réalisations*, de n variables aléatoires indépendantes et identiquement distribuées disons $X_1, X_2, X_3, ..., X_n$. Dans le jargon statistique, on dit que les variables aléatoires X_1, X_2, $X_3, ..., X_n$ constituent un échantillon aléatoire de taille n issu d'une population. L'expression *échantillon aléatoire* est également utilisée pour les valeurs observées $x_1, x_2, ..., x_n$

des variables aléatoires X_1, X_2,..., X_n. On utilise donc la notation $X_1, X_2, ..., X_n$ pour désigner l'échantillon aléatoire qui n'a pas encore été observé et la notation $x_1, x_2, ..., x_n$ pour désigner l'échantillon aléatoire obtenu. Les quantités $X_1, X_2, ..., X_n$ sont des variables aléatoires alors que les quantités $x_1, x_2, ..., x_n$ sont les valeurs observées de ces variables aléatoires.

Étant donné un échantillon aléatoire $X_1, X_2, ..., X_n$, on appelle *statistique* toute variable aléatoire qui peut être calculée à partir des variables aléatoires $X_1, X_2, ..., X_n$, disons $V = g(X_1, X_2, ..., X_n)$. Si $x_1, x_2, ..., x_n$ dénote l'échantillon observé, alors $v = g(x_1, x_2, ..., x_n)$ dénote la valeur observée de la statistique V. Voici quelques exemples de statistiques qu'on a déjà rencontrées :

$$\text{La moyenne échantillonnale} \quad \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

$$\text{La variance échantillonnale} \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

$$\text{La médiane échantillonnale} \quad Q_2 = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ est impair,} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ est pair.} \end{cases}$$

Pour ce dernier exemple, nous avons utilisé la convention suivante :

$$\begin{aligned} X_{(1)} &= \text{la plus petite valeur parmi les valeurs } X_1, X_2, ..., X_n, \\ X_{(2)} &= \text{la deuxième plus petite valeur parmi les valeurs } X_1, X_2, ..., X_n, \\ X_{(3)} &= \text{la troisième plus petite valeur parmi les valeurs } X_1, X_2, ..., X_n, \\ &\vdots \\ X_{(n)} &= \text{la plus grande valeur parmi les valeurs } X_1, X_2, ..., X_n. \end{aligned}$$

3.1.3 Estimateur et estimation

Un estimateur d'un paramètre inconnu, disons le paramètre θ, est une statistique, disons la statistique $V = g(X_1, X_2, ..., X_n)$, qu'on utilise pour estimer ce paramètre. Si on obtient les données $x_1, x_2, ..., x_n$, alors le nombre $v = g(x_1, x_2, ..., x_n)$ s'appelle une estimation du paramètre θ.

Exemple 3 : Si $X_1, X_2, ..., X_n$ est un échantillon aléatoire provenant de la loi $N(\mu, \sigma^2)$, alors $\overline{X}$ est un estimateur de la moyenne théorique μ et la valeur observée $\overline{x}$ est une estimation de la moyenne théorique μ.

Exemple 4 : Si $X_1, X_2, ..., X_n$ est un échantillon aléatoire provenant de la loi $N(\mu, \sigma^2)$, alors S^2 est un estimateur de la variance théorique σ^2 et la valeur observée s^2 est une estimation de la variance théorique σ^2.

Exemple 5 : Si $\hat{p}$ est la proportion échantillonnale de succès calculée à partir d'un échantillon aléatoire de taille n provenant d'une population dichotomique avec proba-

bilité de succès p inconnue, alors $\hat{p}$ est un estimateur du paramètre p et la proportion échantillonnale observée $\hat{p}_{obs}$, qu'on dénote habituellement $\hat{p}$ malgré le risque d'ambiguïté, est une estimation de la proportion théorique p.

3.1.4 Biais et erreur type

Un estimateur, disons V, du paramètre θ est dit *sans biais* si on a $\mathbb{E}[V] = \theta$. Cette condition est très désirable. Elle signifie que notre estimateur *vise juste*, dans le sens que si on utilisait cet estimateur un très grand nombre de fois alors, à la longue, la moyenne de nos estimations serait (à peu près) égale à la vraie valeur du paramètre à estimer. Parfois on obtiendrait des estimations supérieures à la valeur du paramètre, parfois on obtiendrait des estimations inférieures à la valeur du paramètre, mais la moyenne de toutes ces estimations serait (à peu près) égale au paramètre θ.

Toujours avec la notation du paragraphe précédent, imaginez qu'on obtienne nos observations, disons $x_1, x_2, ..., x_n$, et imaginez qu'on calcule la valeur observée v de notre estimateur V. On obtient $v = 7.42$ grammes. Notre estimation du paramètre θ est donc 7.42 grammes. Quelle mesure de précision peut-on associer à cette estimation ? La mesure de précision qu'on utilise est appelée *l'erreur type*. Pour un estimateur sans biais, le seul cas que nous considérerons, cette erreur type est simplement l'écart-type σ_V de notre estimateur V. Les trois sections suivantes vont nous permettre de bien comprendre les concepts de biais et d'erreur type.

3.2 L'estimation d'une moyenne

On considère une certaine population et on s'intéresse à une certaine variable statistique quantitative. La moyenne théorique de cette variable statistique, qu'on appelle aussi la moyenne de la population, est dénotée μ. La variance théorique de cette variable statistique, qu'on appelle aussi la variance de la population, est dénotée σ^2. On s'intéresse ici à l'estimation de la moyenne théorique μ.

L'estimateur usuel de la moyenne théorique μ est la moyenne échantillonnale $\overline{X}$ calculée à partir de notre échantillon aléatoire $X_1, X_2, ..., X_n$ de la façon suivante :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

En utilisant les propriétés de l'espérance et de la variance énoncées à la section 2.5, on obtient facilement les résultats suivants :

$$\mu_{\overline{X}} = \mathbb{E}[\overline{X}] = \mu, \tag{3.1}$$

$$\sigma^2_{\overline{X}} = \mathbb{V}\text{ar}[\overline{X}] = \frac{\sigma^2}{n}, \tag{3.2}$$

$$\sigma_{\overline{X}} = \sqrt{\mathbb{V}\text{ar}[\overline{X}]} = \frac{\sigma}{\sqrt{n}}. \tag{3.3}$$

Le résultat (3.1) nous dit que la moyenne échantillonnale $\overline{X}$ est un estimateur sans biais pour la moyenne théorique μ. Le résultat (3.3) nous permet de calculer l'erreur type associée à la moyenne échantillonnale.

Exemple 6. On veut estimer le poids moyen des rats de laboratoire à la naissance. On obtient un échantillon aléatoire de taille $n = 68$. Autrement dit, on mesure le poids à la naissance de 68 rats de laboratoire. On imagine que ces 68 rats ont été choisis au hasard à partir de la population de tous les rats de laboratoire de l'espèce sous considération. On calcule notre moyenne échantillonnale et on obtient $\overline{x} = 127.35$ grammes. On calcule notre écart-type échantillonnal et on obtient $s = 6.41$ grammes. Notre estimation pour la moyenne μ de la population est donc 127.35 grammes. L'erreur type associée à cette estimation est calculée de la façon suivante :

$$\text{erreur type associée à } \overline{x} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{6.41}{\sqrt{68}} \approx 0.78 \text{ gramme.}$$

Résultat important : Lorsqu'on utilise la moyenne échantillonnale $\overline{x}$ pour estimer la moyenne théorique μ, l'erreur type associée à notre estimation est calculée à l'aide de l'équation suivante :

$$\text{erreur type } = \frac{s}{\sqrt{n}}.$$

Interprétation de l'erreur type dans l'exemple 6 : À l'exemple 6, notre estimation de la moyenne théorique μ était la moyenne échantillonnale $\overline{x} = 127.35$ grammes et notre erreur type était $s/\sqrt{n} = 0.78$ grammes. On interprète cette erreur type de la façon suivante. Imaginez que 3000 biologistes, indépendamment les uns des autres, obtiennent chacun un échantillon aléatoire de taille $n = 68$. Chaque biologiste calcule la moyenne de ses 68 observations. Chaque biologiste obtient un $\overline{x}$ différent. Certaines de ces moyennes échantillonnales $\overline{x}$ sont proches de la vraie moyenne théorique μ. D'autres sont loin de la vraie moyenne théorique μ. L'erreur type $s/\sqrt{n} = 0.78$ gramme est une estimation de la distance typique entre ces 3000 estimations et la vraie moyenne théorique μ. En pratique on obtient un seul échantillon de taille $n = 68$, donc une seule estimation $\overline{x}$. On ne sait pas si le $\overline{x}$ qu'on a obtenu est proche de μ ou loin de μ. Tout ce qu'on sait, c'est que notre $\overline{x}$ a été obtenu avec une *recette* qui, si on l'utilisait un très grand nombre de fois, nous donnerait des estimations $\overline{x}$ qui seraient en moyenne à une distance environ 0.78 de la vraie moyenne théorique μ.

3.3 L'estimation d'une variance

Comme dans la section précédente, on considère une certaine population et on s'intéresse à une certaine variable statistique quantitative. La moyenne théorique de cette variable statistique, qu'on appelle aussi la moyenne de la population, est dénotée μ. La variance théorique de cette variable statistique, qu'on appelle aussi la variance de la population, est dénotée σ^2. On s'intéresse ici à l'estimation de la variance théorique σ^2.

L'estimateur usuel de la variance théorique σ^2 est la variance échantillonnale S^2 calculée à partir de notre échantillon aléatoire $X_1, X_2, ..., X_n$ de la façon suivante :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}.$$

En utilisant les propriétés de l'espérance et de la variance, on peut montrer que

$$\mu_{S^2} = \mathbb{E}[S^2] = \sigma^2, \tag{3.4}$$

$$\sigma^2_{S^2} = \mathbb{V}\text{ar}[S^2] = \frac{2\sigma^4}{n-1}, \tag{3.5}$$

$$\sigma_{S^2} = \sqrt{\mathbb{V}\text{ar}[S^2]} = \frac{\sqrt{2}\ \sigma^2}{\sqrt{n-1}}. \tag{3.6}$$

Le résultat (3.4) nous dit que la variance échantillonnale S^2 est un estimateur sans biais pour la variance théorique σ^2. C'est ce résultat qui explique pourquoi on divise par $n-1$ dans la définition de la variance échantillonnale. Si on avait divisé par n, on aurait obtenu un estimateur biaisé. Le résultat (3.6) nous permet de calculer l'erreur type associée à la variance échantillonnale. Attention : les résultats (3.1) à (3.4) sont toujours vrais ; les résultats (3.5) et (3.6) sont valides seulement dans le cas où la loi normale est un bon modèle pour la population à partir de laquelle on échantillonne.

Exemple 7. Reprenons l'exemple 6 et estimons la variance théorique σ^2. Notre estimation pour la variance théorique σ^2 est la variance échantillonnale $s^2 = (6.41)^2$ grammes^2 = 41.09 grammes^2. Il est raisonnable de supposer que la loi normale est un bon modèle pour décrire la distribution des poids des rats de laboratoire à la naissance. En pratique, il suffirait de dessiner l'histogramme des 68 poids et de constater sa forme de cloche à peu près symétrique. L'erreur type associée à notre estimation est donc

$$\text{erreur type associée à } s^2 = \sigma_{S^2} = \frac{\sqrt{2}\ \sigma^2}{\sqrt{n-1}} \approx \frac{\sqrt{2}\ s^2}{\sqrt{n-1}} = \frac{\sqrt{2}\ 41.09}{\sqrt{68-1}} \approx 7.10 \text{ grammes}^2.$$

Résultat important : Lorsqu'on utilise la variance échantillonnale s^2 pour estimer la variance théorique σ^2 d'une distribution normale, l'erreur type associée à notre estimation est calculée à l'aide de l'équation suivante :

$$\text{Erreur type } = \frac{\sqrt{2}\ s^2}{\sqrt{n-1}} = \sqrt{\frac{2}{n-1}}\ s^2.$$

3.4 L'estimation d'une proportion

On considère une certaine population et on s'intéresse à une certaine variable statistique dichotomique c'est-à-dire une variable avec seulement deux valeurs possibles. Sans perte de généralité on peut supposer que ces deux valeurs possibles sont la valeur 0 et la valeur 1. La proportion de 1 dans la population est dénotée p. La proportion de 0 est donc $1-p$.

À partir de cette population, on obtient un échantillon aléatoire de taille n. On détermine le nombre de 1 et le nombre de 0 dans notre échantillon et on pose

$$\hat{p} = \text{ la proportion de 1 dans l'échantillon.}$$

La statistique $\hat{p}$ s'appelle la proportion échantillonnale. On utilise cette proportion échantillonnale pour estimer la proportion théorique p. En utilisant les propriétés de l'espérance et de la variance et les propriétés de la loi binomiale, on montre facilement que

$$\mu_{\hat{p}} = \mathbb{E}[\hat{p}] = p, \tag{3.7}$$

$$\sigma^2_{\hat{p}} = \mathbb{V}\text{ar}[\hat{p}] = \frac{p(1-p)}{n}, \tag{3.8}$$

$$\sigma_{\hat{p}} = \sqrt{\mathbb{V}\text{ar}[\hat{p}]} = \sqrt{\frac{p(1-p)}{n}}. \tag{3.9}$$

Le résultat (3.7) nous dit que la proportion échantillonnale $\hat{p}$ est un estimateur sans biais pour la proportion théorique p. Le résultat (3.9) nous permet de calculer l'erreur type associée à la proportion échantillonnale $\hat{p}$.

Exemple 8. Reprenons l'exemple 6. Dénotons par p la proportion des rats de laboratoire qui sont porteurs du virus XB-27 à la naissance. Nous désirons estimer p. Parmi les 68 rats que nous avons observés, il y en a 23 qui sont porteurs du virus XB-27. Les 45 autres rats n'ont pas le virus. Notre estimation de la proportion théorique p est donc notre proportion échantillonnale $\hat{p}_{obs} = 23/68 \approx 0.338$, ou 33.8%. L'erreur type associée à cette estimation est

$$\text{erreur type } = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}} = \sqrt{\frac{(0.338)(1-0.338)}{68}} \approx 0.057.$$

Remarque au sujet de la notation. À la section 3.2, on a fait la distinction entre la moyenne échantillonnale $\overline{X}$ (une variable aléatoire) et la moyenne échantillonnale observée $\overline{x}$ (un nombre). De même, à la section 3.3, on a fait la distinction entre la variance échantillonnale S^2 (une variable aléatoire) et la variance échantillonnale observée s^2 (un nombre). Dans la présente section, nous avons utilisé la notation $\hat{p}$ pour dénoter la proportion échantillonnale (une variable aléatoire) et la notation $\hat{p}_{obs}$ pour dénoter la proportion échantillonnale observée (un nombre). Dans ce qui suit, nous laissons tomber la notation $\hat{p}_{obs}$ et, comme la majorité des auteurs, nous utilisons la notation $\hat{p}$ pour dénoter aussi bien la variable aléatoire $\hat{p}$ que la valeur observée de la variable aléatoire $\hat{p}$. La plupart du temps, le contexte nous dit s'il faut interpréter $\hat{p}$ comme étant une variable aléatoire (pas encore observée) ou la valeur observée de cette variable aléatoire.

Résultat important : Lorsqu'on utilise une proportion échantillonnale $\hat{p}$ pour estimer une proportion théorique p, l'erreur type associée à notre estimation est calculée à l'aide de l'équation suivante :

$$\text{erreur type } = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

REMARQUE. Les résultats (3.7), (3.8) et (3.9) sont des cas particuliers des résultats (3.1), (3.2) et (3.3). En effet, si $X_1, X_2, ..., X_n$ dénote l'échantillon aléatoire issu de notre population dichotomique, avec $X_i = 1$ si la i^e observation est un succès et $X_i = 0$ si c'est un échec, alors on a

$$\hat{p} = \frac{\text{nombre de 1 dans l'échantillon}}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \overline{X}.$$

Autrement dit, la proportion échantillonnale $\hat{p}$ n'est rien d'autre qu'un bon vieux $\overline{X}$. Dans le présent scénario, la distribution théorique de notre variable est la loi de Bernoulli(p). La moyenne théorique est donc $\mu = p$ et la variance théorique est $\sigma^2 = p(1-p)$. Les équations (3.7), (3.8) et (3.9) sont simplement les équations (3.1), (3.2) et (3.3) avec $\overline{X}$ remplacé par $\hat{p}$, μ remplacé par p et σ^2 remplacé par $p(1-p)$.

3.5 Intervalle de confiance pour une moyenne

3.5.1 Introduction

Considèrons le scénario de la section 3.2 et supposons que la loi normale soit un bon modèle pour décrire la distribution de la variable d'intérêt. Nous avons vu à la section 3.2 que l'estimation de μ est la moyenne échantillonnale $\overline{x}$ et que l'erreur type associée à cette estimation est donnée par $s/\sqrt{n}$. Le résultat suivant nous permet d'obtenir un intervalle de confiance pour la moyenne théorique μ.

THÉORÈME 3.1 : Si $X_1, X_2, ..., X_n$ sont des variables aléatoires indépendantes et identiquement distribuées, avec distribution $N(\mu, \sigma^2)$, et si $\overline{X}$ et S^2 dénotent la moyenne échantillonnale et la variance échantillonnale, alors on a

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{3.10}$$

où t_{n-1} dénote la loi de Student avec $n-1$ degrés de liberté.

3.5.2 La loi de Student

La loi de Student avec k degrés de liberté est la loi de probabilité avec densité donnée par

$$f(t) = \frac{\Gamma\left((k+1)/2\right)}{\Gamma\left(k/2\right)\sqrt{k\pi}} \; \frac{1}{\left(1 + (t^2/k)\right)^{(k+1)/2}}.$$

Heureusement, nous n'aurons jamais à utiliser cette formule compliquée. Le paramètre k est un entier positif. C'est un paramètre qu'on appelle le *nombre de degrés de liberté*. La justification pour cette terminologie viendra plus tard. La loi de Student avec k degrés de liberté est souvent dénotée t_k. On écrit donc $T \sim t_k$ pour signifier que la distribution de la variable aléatoire T est la loi de Student avec k degrés de liberté.

PROPRIÉTÉS ÉLÉMENTAIRES DE LA LOI DE STUDENT :

(i) La densité de la loi t_k est symétrique, centrée à 0 et en forme de cloche.

(ii) Si $T \sim t_k$ avec $k > 2$, alors $\mathbb{E}[T] = 0$ et $\mathbb{V}\text{ar}[T] = k/(k-2)$.

(iii) Plus k est grand, plus la loi t_k ressemble à la loi $N(0,1)$.

Le théorème 3.1 est la principale raison d'être de la loi de Student.

3.5.3 Les quantiles de la loi de Student

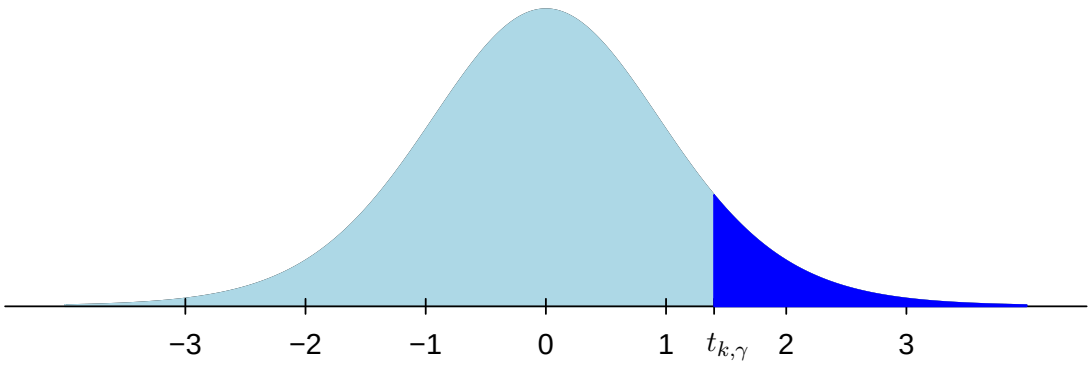
On écrit $t_{k,\gamma}$ pour dénoter le quantile d'ordre $1-\gamma$ de la loi de Student avec k degrés de liberté. Donc si $T \sim t_k$, alors on a

$$\mathbb{P}[T \leq t_{k,\gamma}] = 1 - \gamma$$

ou, de façon équivalente,

$$\mathbb{P}[T > t_{k,\gamma}] = \gamma.$$

Voici le graphique de la loi de Student avec k degrés de liberté. La surface à gauche de $t_{k,\gamma}$, ombragée pâle, est égale à $1-\gamma$. La surface à droite de $t_{k,\gamma}$, ombragée foncée, est égale à γ.



La loi de Student étant symétrique par rapport à 0, on a aussi

$$\mathbb{P}[T < -t_{k,\gamma}] = \mathbb{P}[T > t_{k,\gamma}] = \gamma.$$

Notons également que si $0 < \alpha < 1$, alors on obtient

$$\mathbb{P}[-t_{k,\alpha/2} < T < t_{k,\alpha/2}] = 1 - \left(\mathbb{P}[T < -t_{k,\alpha/2}] + \mathbb{P}[T > t_{k,\alpha/2}]\right) = 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1 - \alpha.$$

Pour obtenir les quantiles de la loi de Student, on utilise une table de la loi de Student, un logiciel de statistique, ou même une calculatrice scientifique. Par exemple, le quantile d'ordre 85% de la loi de Student avec 12 degrés de liberté est $t_{12,0.15} = 1.083$. On trouve cette valeur à la ligne $k = 12$ et la colonne $\gamma = 0.15$ dans la table de la loi de Student présentée à l'Annexe A.2. On peut aussi obtenir ce quantile avec le logiciel R. On tape la commande `qt(0.85, 12)` et R nous retourne la valeur 1.083211.

3.5.4 L'intervalle de confiance pour μ

Fixons α, par exemple $\alpha = 0.05$ ou $\alpha = 0.01$. À partir du résultat (3.10), on obtient

$$\mathbb{P}\left[-t_{n-1,\alpha/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right] = 1 - \alpha.$$

Lorsqu'on isole le paramètre μ dans les inégalités ci-dessus, on obtient

$$\mathbb{P}\left[\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

c'est-à-dire

$$\mathbb{P}\left[\mu \in \left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\,,\, \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)\right] = 1 - \alpha. \tag{3.11}$$

On conclut donc que la probabilité que l'intervalle aléatoire

$$\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\,,\, \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

contienne la moyenne théorique μ est $1 - \alpha$. On obtient nos observations $x_1, x_2, ..., x_n$, on calcule notre moyenne échantillonnale $\overline{x}$ et notre écart-type échantillonnal s. On obtient ainsi l'intervalle

$$\left(\overline{x} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\ \overline{x} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) \tag{3.12}$$

Cet intervalle s'appelle l'intervalle de confiance de niveau $1 - \alpha$ pour la moyenne μ.

Exemple 9. Reprenons l'exemple 6 de la section 3.2 et calculons un intervalle de confiance de niveau 95% pour la moyenne théorique μ. On a $n = 68$, $\overline{x} = 127.35$ grammes et $s = 6.41$ grammes. Avec niveau de confiance $1 - \alpha = 95\%$, on a $\alpha = 5\%$. Dans la table de la loi de Student on trouve $t_{n-1,\alpha/2} = t_{67,0.05} \approx 2.00$. On insère tout ça dans l'intervalle (3.12) et on obtient l'intervalle $(125.80, 128.90)$.

Interprétation : Dans l'exemple 6, nous avons obtenu un échantillon aléatoire de taille $n = 68$. À partir de cet échantillon aléatoire, nous avons obtenu $\overline{x} = 127.35$ grammes et $s = 6.41$ grammes. Nous avons ensuite inséré ces valeurs dans l'intervalle (3.12) et nous avons obtenu l'intervalle $(125.80, 128.90)$. Imaginez 3000 biologistes qui refont cette expérience indépendamment les uns des autres. Chaque biologiste obtient un échantillon de taille $n = 68$. Chaque biologiste obtient un $\overline{x}$ et un s. Chaque biologiste insère ces valeurs dans l'intervalle (3.12). Certains biologistes obtiennent un intervalle qui contient la moyenne théorique μ et d'autres obtiennent un intervalle qui ne contient pas la moyenne μ. En vertu de l'équation (3.11), on s'attend à ce qu'environ 95% de ces intervalles contiennent μ. Dans la pratique, nous avons un seul échantillon aléatoire, donc un seul $\overline{x}$ et un seul s. Nous avons obtenu $\overline{x} = 127.35$ et $s = 6.41$, nous avons inséré ces valeurs dans l'équation (3.12) et nous avons obtenu l'intervalle $(125.80, 128.90)$. Nous sommes confiant à 95% que la vraie moyenne théorique μ se situe quelque part dans l'intervalle $(125.80, 128.90)$ parce que nous avons utilisé une méthode qui 95 fois sur 100 donne un intervalle contenant μ.

3.6 Intervalle de confiance pour une variance

3.6.1 Introduction

On considère le scénario de la section 3.2 et on suppose que la loi normale est un bon modèle pour décrire la distribution de la variable d'intérêt. Nous avons vu à la section 3.3 que l'estimation de σ^2 est la variance échantillonnale s^2 et que l'erreur type associée à cette estimation est donnée par $\sqrt{2}\ s^2/\sqrt{n-1}$. Le résultat suivant nous permet d'obtenir un intervalle de confiance pour la variance théorique σ^2.

THÉORÈME 3.2 : Si $X_1, X_2, ..., X_n$ sont des variables aléatoires indépendantes et identiquement distribuées, avec distribution $N(\mu, \sigma^2)$, et si S^2 dénote la variance échantillonnale, alors on a

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{3.13}$$

où χ^2_{n-1} dénote la loi du khi-deux avec $n-1$ degrés de liberté.

3.6.2 La loi du khi-deux

La loi du khi-deux avec k degrés de liberté est la loi de probabilité avec densité donnée par l'équation suivante :

$$f(u) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)}\ u^{(k/2)-1}\mathrm{e}^{-u/2} & \text{si } u \geq 0, \\ 0 & \text{si } u < 0. \end{cases}$$

Comme dans le cas de la loi de Student, nous n'aurons jamais à utiliser cette formule compliquée. Comme dans le cas de la loi de Student, le paramètre k est un entier positif. On écrit $U \sim \chi^2(k)$ pour signifier que la distribution de la variable aléatoire U est la loi du khi-deux avec k degrés de liberté. Le résultat suivant est la principale raison d'être de la loi du khi-deux :

THÉORÈME 3.3 : Si $Z_1, Z_2, ..., Z_k$ sont des variables aléatoires indépendantes et identiquement distribuées, avec distribution $N(0, 1)$, alors on a

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2_k. \tag{3.14}$$

Le théorème 3.3 ressemble beaucoup au théorème 3.2. Examinons de plus près l'énoncé du théorème 3.2. Puisque $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, la statistique $(n-1)S^2/\sigma^2$ qui apparaît dans l'énoncé du théorème 3.2 peut s'écrire de la façon suivante :

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

Les variables aléatoires $(X_i - \overline{X})/\sigma$ sont presque indépendantes et suivent presque la loi $N(0, 1)$. Il suffirait de remplacer le $\overline{X}$ par μ pour qu'elles soient des variables indépendantes

qui suivent toutes la loi $N(0,1)$. Si on remplace $\overline{X}$ par μ, le théorème 3.3 nous donne alors

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n.$$

Le théorème 3.2 est valide parce qu'il est possible de montrer que la somme $\sum_{i=1}^{n}((X_i - \overline{X})/\sigma)^2$ peut s'écrire comme une somme de $n-1$ carrés de variables aléatoires indépendantes les unes des autres et qui suivent toutes la loi $N(0,1)$. D'ailleurs, chaque fois qu'on rencontre une statistique qui suit une loi du khi-deux avec k degrés de liberté, on peut être assuré que cette statistique peut s'écrire comme une somme de k carrés de variables aléatoires $N(0,1)$ indépendantes les unes des autres.

Propriétés élémentaires de la loi du khi-deux :

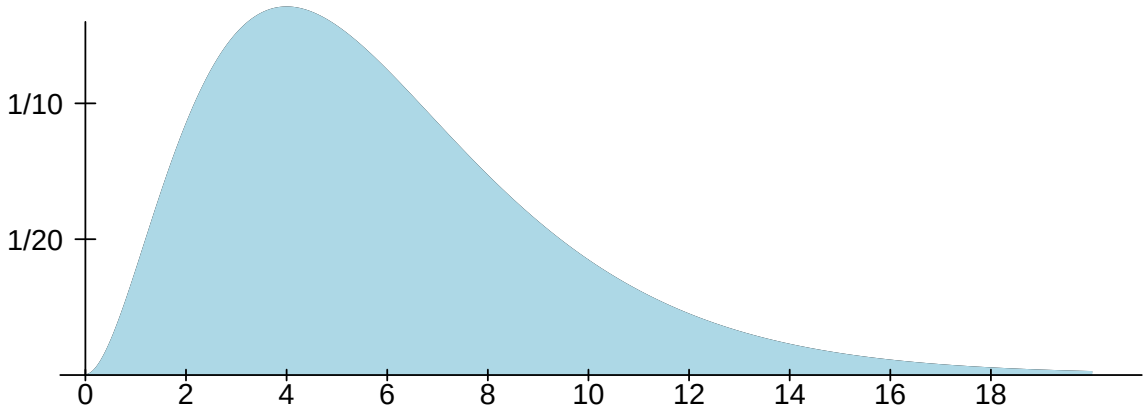
(i) L'espérance de la loi $\chi^2(k)$ est égale à k.

(ii) La variance de la loi $\chi^2(k)$ est égale à $2k$.

(iii) Si U et V sont des variables aléatoires indépendantes et si $U \sim \chi^2(k)$ et $V \sim \chi^2(\ell)$ alors $U + V \sim \chi^2(k + \ell)$.

(iv) Si $k > 2$, la densité de la loi $\chi^2(k)$ est en forme de cloche asymétrique étirée vers la droite.

Le graphe ci-dessous illustre la loi du khi-deux avec 6 degrés de liberté.



3.6.3 Les quantiles de la loi du khi-deux

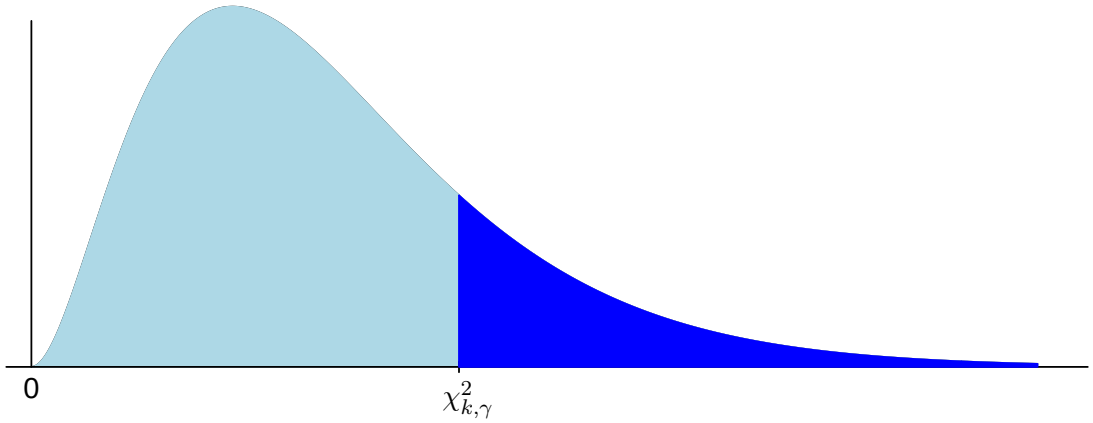
On écrit $\chi^2_{k,\gamma}$ pour dénoter le quantile d'ordre $1-\gamma$ de la loi du khi-deux avec k degrés de liberté. Donc si $U \sim \chi^2_k$, alors on a

$$\mathbb{P}[U \leq \chi^2_{k,\gamma}] = 1 - \gamma$$

ou, de façon équivalente,

$$\mathbb{P}[U > \chi^2_{k,\gamma}] = \gamma.$$

Voici le graphique de la loi du khi-deux avec k degrés de liberté. La surface à gauche de $\chi^2_{k,\gamma}$, ombragée pâle, est égale à $1-\gamma$. La surface à droite de $\chi^2_{k,\gamma}$, ombragée foncée, est égale à γ.



Pour obtenir les quantiles de la loi du khi-deux, on utilise une table de la loi du khi-deux, un logiciel de statistique, ou même une calculatrice scientifique. Par exemple, le quantile d'ordre 90% de la loi du khi-deux avec 12 degrés de liberté est $\chi^2_{12,0.10} = 18.55$. On trouve cette valeur dans la table de la loi du khi-deux à la ligne $k = 12$ et la colonne $\gamma = 0.10$. On peut aussi obtenir ce quantile avec le logiciel R. On tape la commande `qchisq(0.90,12)` et R nous retourne la valeur 18.54935.

À l'aide de la table de la loi du khi-deux présentée à l'Annexe A.3, l'étudiant devrait pouvoir vérifier les affirmations suivantes :

1. Si $U \sim \chi^2_{12}$, alors $\mathbb{P}[U > 18.55] = 0.10$.
2. Si $U \sim \chi^2_{12}$, alors $\mathbb{P}[4.40 < U < 23.34] = 0.95$.
3. Si $U \sim \chi^2_{27}$, alors $0.025 < \mathbb{P}[U > 42] < 0.050$.
4. Le quantile d'ordre 90% de la loi χ^2_8 est 13.36.
5. La médiane de la loi χ^2_{30} est 29.34.
6. Le premier centile de la loi χ^2_{40} est $\chi^2_{40,0.99} = 22.16$.
7. Le dixième centile de la loi χ^2_{30} est $\chi^2_{30,0.9} = 20.60$.

3.6.4 L'intervalle de confiance pour σ^2

Fixons α, par exemple $\alpha = 0.05$ ou $\alpha = 0.01$. À partir du résultat (3.13), on obtient

$$\mathbb{P}\left[\chi^2_{n-1,1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\frac{\alpha}{2}}\right] = 1-\alpha.$$

Lorsqu'on isole le paramètre σ^2 dans les inégalités ci-dessus, on obtient

$$\mathbb{P}\left[\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right] = 1-\alpha$$

c'est-à-dire

$$\mathbb{P}\left[\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}},\ \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)\right] = 1-\alpha. \tag{3.15}$$

On peut donc conclure que la probabilité que l'intervalle aléatoire

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}},\ \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)$$

contienne la variance théorique σ^2 est $1-\alpha$. On obtient nos observations $x_1, x_2, ..., x_n$, on calcule notre variance échantillonnale s^2. On obtient ainsi l'intervalle

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}},\ \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right). \tag{3.16}$$

Cet intervalle s'appelle l'intervalle de confiance de niveau $1-\alpha$ pour la variance σ^2.

Exemple 10. Reprenons l'exemple 7 et calculons un intervalle de confiance de niveau 95% pour la variance théorique σ^2. On a $n = 68$ et $s^2 = (6.41)^2 = 41.09$ grammes^2. Avec niveau de confiance $1-\alpha = 95\%$, on a $\alpha = 5\%$. Avec le logiciel R, on trouve $\chi^2_{n-1,\alpha/2} = \chi^2_{67,0.025} = 91.519$ et $\chi^2_{n-1,1-\alpha/2} = \chi^2_{67,0.975} = 46.261$. On insère tout ça dans l'intervalle (3.16) et on obtient l'intervalle $(30.08, 59.51)$. Il s'agit de l'intervalle de confiance de niveau 95% pour la variance théorique σ^2. L'intervalle de confiance de niveau 95% pour l'écart-type théorique σ est donc donné par $(\sqrt{30.08}, \sqrt{59.51}) = (5.48, 7.71)$.

Interprétation : On est confiant à 95% que l'intervalle $(5.48, 7.71)$ contient σ parce que cet intervalle a été obtenu en utilisant une méthode qui 95 fois sur 100 donne un intervalle qui contient σ.

Important : L'intervalle (3.16) est un intervalle de confiance pour la variance théorique σ^2. Si on veut un intervalle de confiance pour l'écart-type théorique σ, il suffit de prendre

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}},\ \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}}\right)$$

3.7 Intervalle de confiance pour une proportion

Avec la notation et les hypothèses de la section 3.4, on a le résultat suivant. Ce résultat peut être déduit à partir du fameux théorème limite central.

THÉORÈME 3.4 : Si $X_1, X_2, ..., X_n$ sont des variables aléatoires indépendantes et identiquement distribuées, avec distribution Bernoulli(p) et si $\hat{p}$ dénote la proportion échantillonnale, alors, lorsque n est suffisamment grand,

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1). \tag{3.17}$$

À partir du résultat (3.17) et en procédant comme à la section 3.5.4, il est facile de montrer que l'intervalle suivant est un intervalle de confiance de niveau $1-\alpha$ pour la proportion théorique p :

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right). \tag{3.18}$$

Cet intervalle est valide à condition que n soit suffisamment grand, par exemple $n = 30$ ou $n = 50$. Dans l'intervalle (3.18), le $z_{\alpha/2}$ dénote le quantile d'ordre $1 - \alpha/2$ de la loi $N(0,1)$. Autrement dit, $z_{\alpha/2}$ est la valeur qui est telle que

$$\mathbb{P}[Z \leq z_{\alpha/2}] = 1 - \alpha/2$$

où Z est une variable aléatoire qui suit la loi $N(0,1)$. De façon équivalente, $z_{\alpha/2}$ est la valeur qui est telle que

$$\mathbb{P}[Z > z_{\alpha/2}] = \alpha/2.$$

Autrement dit, la surface à droite de la valeur $z_{\alpha/2}$ sous la densité de la loi $N(0,1)$ est égale à $\alpha/2$.

EXEMPLE 11. Reprenons l'exemple 8 de la section 3.4 et calculons un intervalle de confiance de niveau 95% pour la proportion théorique p. On a $n = 68$ et $\hat{p} = 23/68 = 0.338$. La table de la loi $N(0,1)$ nous donne $z_{\alpha/2} = z_{0.025} = 1.96$. On insère tout ça dans l'intervalle (3.18) et on obtient l'intervalle $(0.226, 0.451)$. On interprète cet intervalle de la même façon qu'on a interprété l'intervalle de confiance pour une moyenne et l'intervalle de confiance pour une variance ou un écart-type.

3.8 Calcul de taille d'échantillon

Il est parfois possible de déterminer à l'avance la taille d'échantillon nécessaire pour que l'intervalle de confiance soit de telle ou telle longueur désirée. Nous considérons ici le cas de l'intervalle de confiance pour une proportion c'est-à-dire l'intervalle (3.18). La longueur de cet intervalle est

$$L = 2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

La demi-longueur est donc

$$d = \frac{L}{2} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Cette demi-longueur d représente la distance maximale entre $\hat{p}$ et p lorsque l'intervalle de confiance contient le paramètre p. Notez que l'intervalle de confiance peut s'écrire de la façon suivante :

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \pm d.$$

Supposons qu'on fixe d à l'avance et qu'on aimerait déterminer la taille d'échantillon n qui fera en sorte que la demi-longueur de notre intervalle de confiance sera précisément ce d qu'on s'est fixé d'avance. Pour trouver n, il suffit de résoudre l'équation

$$d = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Lorsqu'on résoud cette équation, on obtient

$$n = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{d^2}. \tag{3.19}$$

Mais on ne connaît pas $\hat{p}$. On n'a pas encore obtenu nos données puisqu'on est en train d'essayer de déterminer quelle taille d'échantillon on va utiliser ! Pour s'en sortir, on utilise le fait que $\hat{p}(1-\hat{p}) \leq 1/4$. (Pour comprendre cela, dessinez le graphe de la fonction $f(x) = x(1-x)$. Vous allez constater que cette fonction atteint son maximum à $x = 1/2$ et vaut alors $1/4$. On a donc $f(x) \leq 1/4$ pour tout x, c'est-à-dire $x(1-x) \leq 1/4$ pour tout x.) Pour être certain d'avoir un n suffisamment grand, il suffit donc de prendre

$$n \approx \frac{z_{\alpha/2}^2(1/4)}{d^2} = \frac{z_{\alpha/2}^2}{4d^2} = \left(\frac{z_{\alpha/2}}{2d}\right)^2 = \left(\frac{z_{\alpha/2}}{L}\right)^2. \tag{3.20}$$

Exemple 12. Lorsqu'on estime une proportion p, quelle taille d'échantillon n nous assure que notre intervalle de confiance de niveau 95% aura une erreur d d'au plus 0.02 ? Autrement dit, quelle taille d'échantillon n nous assure que la longueur L de notre intervalle de confiance de niveau 95% sera au plus 0.04 ?

Solution. On prend

$$n = \left(\frac{z_{\alpha/2}}{2d}\right)^2$$

avec $\alpha = 0.05$ et avec $d = 0.02$. La table de la loi normale nous donne $z_{\alpha/2} = z_{0.025} = 1.96$ et on obtient

$$n = \left(\frac{z_{\alpha/2}}{2d}\right)^2 = \left(\frac{1.96}{2 \cdot 0.02}\right)^2 = \left(\frac{1.96}{0.04}\right)^2 = 2401.$$

Exemple 13. Entendu à la radio : *D'après un sondage réalisé par la Société Radio-Canada, 57% des Canadiens sont contre la peine de mort. L'erreur est 3 pour cent 19 fois sur 20.* Interprétez cet énoncé et déterminez la taille de l'échantillon utilisé.

SOLUTION. Le 3% représente l'erreur $d = 0.03$. Le 19 fois sur 20 représente le niveau de confiance $19/20 = 95/100 = 95\%$. La taille de l'échantillon était donc

$$n = \left(\frac{z_{\alpha/2}}{2d}\right)^2 = \left(\frac{1.96}{2 \cdot 0.03}\right)^2 = \left(\frac{1.96}{0.06}\right)^2 = 1067.$$

3.9 Tests d'hypothèses : un exemple illustratif

Supposons que la loi normale avec moyenne égale à 12.50 cm et avec écart-type égal à 1.50 cm soit un bon modèle pour la distribution des tailles d'un certain type de plants de tomates trois semaines après germination. Un biologiste propose un nouveau type d'engrais. L'effet principal de ce nouveau type d'engrais serait d'accélérer la croissance des plants de tomates. Afin de voir si son nouvel engrais produit l'effet désiré, le biologiste va l'utiliser sur 25 plants de tomate. Notre biologiste mesurera ensuite les tailles de ces 25 plants trois semaines après germination, puis calculera la moyenne de ces 25 tailles. Si cette moyenne échantillonnale est proche de 12.50 cm, il conclura que son nouvel engrais n'est pas meilleur que l'ancien. Mais si la taille moyenne de ces 25 plants est suffisamment supérieure à 12.50 cm, notre biologiste conclura que la nouvelle moyenne théorique est effectivement supérieure à 12.50 cm. Il pourra ensuite faire une étude plus approfondie afin d'estimer avec précision la moyenne théorique des tailles de plants de tomates soumis au nouvel engrais.

Pour simplifier le problème, nous allons supposer que le nouveau type d'engrais n'affecte pas la forme de la distribution des tailles. Autrement dit, on suppose que la loi normale avec écart-type $\sigma = 1.50$ cm est un bon modèle pour décrire la distribution des tailles, trois semaines après germination, des plants de tomates soumis au nouvel engrais. Toutefois, on admet la possibilité qu'avec le nouvel engrais la moyenne théorique ait une valeur plus grande que 12.50 cm. Dénotons par μ la moyenne théorique des tailles, trois semaines après germination, des plants de tomates soumis au nouvel engrais. La question que notre biologiste se pose est la suivante : est-ce que cette moyenne théorique μ est égale à 12.50 cm ou est-ce qu'elle est plus grande que 12.50 cm ? La première possibilité, $\mu = 12.50$ cm, représente le statu quo et est appelée *l'hypothèse nulle*. La deuxième possibilité, $\mu > 12.50$ cm, représente le changement et est appelée *l'hypothèse alternative*. On écrit H_0 pour l'hypothèse nulle et H_1 pour l'hypothèse alternative. On a donc

$$H_0 : \mu = 12.50 \text{ cm},$$
$$H_1 : \mu > 12.50 \text{ cm}.$$

Notre biologiste devra prendre une décision. Il devra ou bien *accepter* H_0, ou bien *rejeter* H_0 en faveur de H_1. Pour prendre sa décision, il utilisera une *règle de décision* basée sur les observations qu'il aura obtenues. Ici, les observations sont les 25 tailles qu'il obtiendra. Dénotons par $X_1, X_2, ..., X_{25}$ ces 25 tailles et par $\overline{X}$ la moyenne de ces 25 tailles. La règle de décision que notre biologiste utilisera aura la forme suivante :

On rejette H_0 si $\overline{X}$ est suffisamment grand.

Autrement dit, la règle de décision sera de la forme

$$\text{On rejette } H_0 \text{ si } \overline{X} \geq c,$$

où c est une constante appropriée. Il reste à voir comment choisir la constante c.

3.10 Tests d'hypothèses : théorie générale

3.10.1 Le modèle et les hypothèses

L'exemple de la section précédente est du type suivant. On considère une distribution de probabilité avec un certain paramètre, disons le paramètre θ, dont la valeur est mise en doute. On s'intéresse à deux hypothèses concernant ce paramètre θ : l'hypothèse nulle qu'on note H_0 et qui représente le statu quo et l'hypothèse alternative qu'on note H_1 et qui représente une nouvelle théorie, une nouvelle réalité, etc. Ces hypothèses sont exprimées en termes du paramètre θ de notre modèle. L'hypothèse nulle est de la forme

$$H_0 : \theta = \theta_o$$

alors que l'hypothèse alternative prend habituellement l'une ou l'autre des trois formes suivantes :

(i) $H_1 : \theta > \theta_o$,

(ii) $H_1 : \theta < \theta_o$,

(iii) $H_1 : \theta \neq \theta_o$.

Ces trois scénarios sont appelés, dans l'ordre, *hypothèse alternative unilatérale à droite, hypothèse alternative unilatérale à gauche* et *hypothèse alternative bilatérale*. La valeur θ_o est appelée *la valeur spécifiée par l'hypothèse nulle*. Le but du statisticien est de prendre une décision : ou bien on accepte H_0, ou bien on rejette H_0 en faveur de H_1. Notre décision dépendra des données que nous observerons.

Dans l'exemple de la section précédente, la moyenne μ joue le rôle du paramètre θ, la distribution de probabilité est la loi $N(\mu, 2.25)$ et la valeur spécifiée par l'hypothèse nulle est $\theta_o = \mu_o = 12.50$ cm. L'hypothèse alternative est unilatérale à droite.

3.10.2 Les données

Comme d'habitude, nos données seront dénotées $x_1, x_2, ..., x_n$ et seront interprétées comme étant la réalisation d'un échantillon aléatoire de taille n, disons $X_1, X_2, ..., X_n$, issu d'une certaine distribution qui dépend du paramètre θ. Dans l'exemple de la section précédente, on dispose d'un échantillon aléatoire de taille $n = 25$ issu de la loi $N(\mu, 2.25)$.

3.10.3 La règle de décision

Dans la plupart des exemples qu'on rencontre en pratique, la règle de décision est de la forme

$$\text{on rejette } H_0 \text{ si } V \geq c.$$

Ici V est une statistique calculée à partir de nos observations $X_1, X_2, ..., X_n$. Cette statistique V est alors appelée la *statistique du test*. L'ensemble des valeurs de V pour lesquelles

on est amené à rejeter H_0 est parfois appelé la *région de rejet*. De même, l'ensemble des valeurs de V pour lesquelles on est amené à accepter H_0 est appelé la *région d'acceptation*. Dans le présent exemple, la région de rejet est l'intervalle $[c, \infty)$ alors que la région d'acceptation est l'intervalle $(-\infty, c)$. Le choix de la région de rejet, c'est-à-dire le choix de la constante c, est, jusqu'à un certain point, arbitraire. L'approche classique consiste à choisir cette constante c de façon à ce que la probabilité d'erreur de première espèce soit égale à telle ou telle valeur désirée.

3.10.4 Erreur de première espèce et erreur de deuxième espèce

Lorsque la règle de décision est appliquée et qu'une décision est prise, il peut arriver qu'on commette une erreur. Autrement dit, il peut arriver qu'on prenne la mauvaise décision. On distingue deux types d'erreur. Si on rejette l'hypothèse nulle H_0 alors que cette hypothèse est vraie, alors on dit qu'une *erreur de première espèce* a été commise. La probabilité d'erreur de première espèce est habituellement dénotée α. Si on accepte l'hypothèse nulle H_0 alors que cette hypothèse est fausse, alors on dit qu'une *erreur de deuxième espèce* a été commise. La probabilité d'erreur de deuxième espèce est souvent dénotée β.

La probabilité d'erreur de première espèce associée à une règle de décision, aussi appelée le *seuil de notre règle de décision* ou le *seuil de notre test*, est habituellement facile à calculer. Dans l'exemple de la section 3.9, on obtient

$$\begin{aligned}
\alpha &= \mathbb{P}_{H_0}[\text{ On rejette } H_0\ ] \\
&= \mathbb{P}_{H_0}[\overline{X} \geq c] \\
&= \mathbb{P}_{H_0}\left[\frac{\overline{X} - 12.50}{(1.50)/\sqrt{25}} \geq \frac{c - 12.50}{(1.50)/\sqrt{25}}\right] \\
&= \mathbb{P}\left[Z \geq \frac{c - 12.50}{(1.50)/\sqrt{25}}\right] \\
&= \text{ surface à droite de } \frac{c - 12.50}{(1.50)/\sqrt{25}} \text{ sous la } N(0,1).
\end{aligned}$$

Si la constante c est spécifiée, alors on peut calculer α. Inversement, si on fixe α d'avance, alors on peut trouver la valeur c qui fera que notre probabilité d'erreur de première espèce sera égale à ce α qu'on s'est fixé d'avance.

Contrairement à l'hypothèse nulle H_0, l'hypothèse alternative H_1 ne spécifie pas la valeur du paramètre. Dans l'exemple de la section 3.9, l'hypothèse H_1 nous dit simplement que $\mu > 12.50$ sans spécifier la valeur de μ. La probabilité d'erreur de deuxième espèce dépend de cette valeur μ. À titre illustratif, calculons la probabilité d'erreur de deuxième espèce dans le cas où μ serait égal à 13.75 cm. On obtient alors

$$
\begin{aligned}
\beta &= \beta(13.75) \\
&= \mathbb{P}_{\{\mu=13.75\}}[\text{ On accepte } H_0\ ] \\
&= \mathbb{P}_{\{\mu=13.75\}}[\overline{X} < c] \\
&= \mathbb{P}_{\{\mu=13.75\}}\left[\frac{\overline{X}-13.75}{(1.50)/\sqrt{25}} < \frac{c-13.75}{(1.50)/\sqrt{25}}\right] \\
&= \mathbb{P}\left[Z < \frac{c-13.75}{(1.50)/\sqrt{25}}\right] \\
&= \text{ surface à gauche de } \frac{c-13.75}{(1.50)/\sqrt{25}} \text{ sous la } N(0,1).
\end{aligned}
$$

Si la constante c est spécifiée, alors on peut calculer $\beta = \beta(\mu)$ pour n'importe quelle valeur de la moyenne μ.

3.10.5 L'approche classique : test de seuil α

L'approche classique consiste à fixer à l'avance une valeur α, habituellement $\alpha = 0.05$ ou $\alpha = 0.01$, et à choisir notre règle de décision de façon à ce que notre probabilité d'erreur de première espèce soit égale à cette valeur α. Puisque l'hypothèse nulle représente le statu quo, cette approche reflète une attitude conservatrice. On veut que la probabilité d'erreur de première espèce soit petite.

Reprenons l'exemple de la section 3.9 sous une forme plus générale, plus abstraite. On considère une distribution $N(\mu, \sigma^2)$, avec variance σ^2 connue. On veut tester l'hypothèse nulle $H_0 : \mu = \mu_o$ contre l'hypothèse alternative $H_1 : \mu > \mu_o$. On dispose d'un échantillon aléatoire de taille n, disons $X_1, X_2, ..., X_n$, issu de cette distribution $N(\mu, \sigma^2)$. Notre règle de décision prend alors la forme

$$\text{on rejette } H_0 \text{ si } \overline{X} \geq c.$$

Cette règle de décision peut aussi s'écrire sous la forme

$$\text{on rejette } H_0 \text{ si } \frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} \geq c',$$

avec $c' = \frac{c-\mu_o}{\sigma/\sqrt{n}}$. Imaginez qu'on se fixe un α, par exemple $\alpha = 0.05$ ou $\alpha = 0.01$, et qu'on veuille déterminer la constante c', ou de façon équivalente la constante c, pour faire en sorte que notre probabilité d'erreur de première espèce soit précisément ce α qu'on s'est fixé d'avance. Rien de plus facile ! On veut

$$\mathbb{P}_{H_0}[\text{ on rejette } H_o] = \alpha$$

c'est-à-dire

$$\mathbb{P}_{H_0}[\overline{X} \geq c] = \alpha$$

c'est-à-dire

$$\mathbb{P}_{H_0}\left[\frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} \geq c'\right] = \alpha. \tag{3.21}$$

La notation $\mathbb{P}_{H_0}$ signifie qu'on calcule une probabilité en supposant que l'hypothèse nulle H_0 est vraie. Or on sait que si H_0 est vraie alors notre statistique de test

$$Z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}}$$

suit la loi $N(0,1)$. Donc, pour que l'équation (3.21) soit vraie, il faut prendre $c' = z_\alpha$ (et donc $c = \mu_o + z_\alpha \sigma/\sqrt{n}$). Notre règle de décision au seuil α est donc donnée par

$$\text{on rejette } H_0 \text{ si } Z \geq z_\alpha \tag{3.22}$$

ou, de façon équivalente,

$$\text{on rejette } H_0 \text{ si } \overline{X} \geq \mu_o + z_\alpha \tfrac{\sigma}{\sqrt{n}}.$$

3.10.6 Le seuil observé ou *p-value*

Considérons à nouveau l'exemple où l'on dispose d'un échantillon aléatoire de taille n issu d'une loi $N(\mu, \sigma^2)$ avec variance σ^2 connue et où l'on veut tester au seuil α l'hypothèse $H_0 : \mu = \mu_o$ contre l'alternative unilatérale $H_1 : \mu > \mu_o$. La règle de décision basée sur un échantillon aléatoire de taille n est alors donnée par l'équation (3.22). Pour fixer les idées, supposons que $\sigma = 3$, $\mu_o = 20$, $n = 25$ et $\alpha = 0.05$. On a donc

$$H_0 : \mu = 20,$$
$$H_1 : \mu > 20.$$

On obtient nos 25 observations, on calcule notre moyenne échantillonnale et on obtient $\overline{x} = 21.45$. La valeur observée de notre statistique de test est donc

$$Z_{obs} = \frac{\overline{x} - \mu_o}{\sigma/\sqrt{n}} = \frac{21.45 - 20.00}{3/\sqrt{25}} = 2.417.$$

Au seuil $\alpha = 5\%$, la valeur critique de notre règle de décision est $z_{0.05} = 1.645$. La valeur observée de notre statistique de test excède cette valeur critique. Donc, au seuil 5%, on rejette H_0.

Il est important d'être conscient du fait que notre règle de décision dépend du seuil α qu'on s'est fixé. Si au lieu de prendre $\alpha = 0.05$ on avait pris $\alpha = 0.005$, alors notre valeur critique aurait été $z_{0.005} = 2.576$ et l'hypothèse nulle H_0 n'aurait pas été rejetée. Le choix du seuil α dépend de la nature du problème sous considération et des éventuelles conséquences d'une erreur de première espèce. Les valeurs $\alpha = 0.05$ et $\alpha = 0.01$ sont les valeurs les plus utilisées dans la pratique.

Même avec un choix judicieux d'un seuil α, il n'est pas souhaitable de simplement rapporter qu'on a obtenu une statistique Z_{obs} supérieure à la valeur critique z_α ou une statistique Z_{obs} inférieure à la valeur critique z_α. Dans l'exemple ci-dessus, si on observe $\overline{x} = 21.00$, alors on obtient $Z_{obs} = 1.667$ et, au seuil 5%, on rejette H_0 de justesse alors que si on observe $\overline{x} = 22.50$, alors on obtient $Z_{obs} = 4.167$ et, au seuil 5%, on rejette H_0 sans hésitation. Une façon de mesurer le degré de plausibilité de l'hypothèse nulle une fois

que notre statistique de test a été observée est de calculer la probabilité d'obtenir une statistique de test aussi extrême que celle qu'on vient d'observer si on recommençait notre expérience et si l'hypothèse H_0 était vraie. Cette probabilité s'appelle le *seuil observé* ou le *p-value*. Certains auteurs l'appellent aussi la *valeur significative observée*. Dans notre exemple, si on observe $\overline{x} = 21.00$, alors on obtient $Z_{obs} = 1.667$ et notre *p-value* est

$$p\text{-}value = \mathbb{P}[Z \geq 1.667] = 0.0478$$

tandis que si on observe $\overline{x} = 22.50$, alors on obtient $Z_{obs} = 4.167$ et notre *p-value* est

$$p\text{-}value = \mathbb{P}[Z \geq 4.167] = 0.000015.$$

Voici une définition générale du *p-value*. On considère un problème de test d'hypothèse avec, par exemple, $H_0 : \theta = \theta_o$ et $H_1 : \theta > \theta_o$, et avec règle de décision « on rejette H_0 si $V \geq c$ ». Notre statistique de test V est calculée à partir d'un échantillon aléatoire $X_1, X_2, ..., X_n$ issu d'une distribution avec paramètre θ. On observe nos données $x_1, x_2, ..., x_n$, on calcule notre statistique de test et on obtient la valeur V_{obs}. Le *p-value* est la probabilité d'obtenir une statistique de test V aussi extrême que celle qu'on vient d'obtenir si on recommençait notre expérience, c'est-à-dire si on obtenait un nouvel échantillon de taille n, et si l'hypothèse H_0 était vraie. Autrement dit,

$$p\text{-}value = \mathbb{P}_{H_0}[V \geq V_{obs}].$$

Il est important de noter que le *p-value* est toujours la probabilité d'obtenir, sous H_0, une statistique de test *aussi extrême* que celle qu'on vient d'obtenir. Ici, « *aussi extrême* » veut toujours dire « *aussi extrême dans la direction de l'alternative* ». Les exemples des trois prochaines sections vont nous permettre de clarifier cet important concept de *p-value*.

Lien entre le *p-value* et le seuil α : Supposons qu'on se soit fixé un certain seuil α. Notre règle de décision pourrait alors s'énoncer de la façon suivante :

on rejette H_0 si notre *p-value* est plus petit ou égal à notre seuil α.

3.11 Tests sur une moyenne

On considère une certaine population. On s'intéresse à une certaine variable, disons la variable X. On suppose que la loi $N(\mu, \sigma^2)$ est un bon modèle pour décrire la distribution de cette variable. Contrairement à l'exemple de la section précédente, on ne connaît pas la variance théorique σ^2. L'hypothèse nulle prend la forme

$$H_0 : \mu = \mu_o.$$

L'hypothèse alternative sera toujours l'une des trois hypothèses suivantes :

$$H_1 : \mu > \mu_o, \qquad H_1 : \mu < \mu_o, \qquad H_1 : \mu \neq \mu_o.$$

À la section 3.10, nous avons vu que dans le cas où la variance théorique σ^2 était connue, on pouvait baser notre règle de décision sur la statistique

$$Z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}}.$$

Pour déterminer notre région de rejet, nous avons utilisé le fait que si H_o est vraie alors cette statistique de test suit la loi $N(0,1)$. Dans le scénario qui nous intéresse présentement, la variance théorique σ^2 est inconnue. Notre règle de décision sera donc basée sur la statistique

$$T = \frac{\overline{X} - \mu_o}{S/\sqrt{n}}. \tag{3.23}$$

Pour déterminer notre région de rejet, nous utiliserons le fait que si H_0 est vraie alors notre statistique T suit la loi de Student à $n-1$ degrés de liberté, une conséquence du théorème 3.1.

3.11.1 Le cas $H_1 : \mu > \mu_o$

En procédant comme à la section 3.10, on arrive à la conclusion que la règle de décision au seuil α est la suivante :

$$\text{on rejette } H_0 \text{ si } T \geq t_{n-1,\alpha} \tag{3.24}$$

où T est la statistique de test donnée par l'équation (3.23). Dénotons par T_{obs} la valeur observée de notre statistique de test. Autrement dit,

$$T_{obs} = \frac{\overline{x} - \mu_o}{s/\sqrt{n}}.$$

Le *p-value* est alors obtenu de la façon suivante :

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[T \geq T_{obs}] \\ &= \text{ surface à droite de } T_{obs} \text{ sous la densité } t_{n-1}. \end{aligned}$$

3.11.2 Le cas $H_1 : \mu < \mu_o$

En s'inspirant du cas précédent et en exploitant le fait que la loi de Student est symétrique, on arrive à la conclusion que la règle de décision au seuil α est la suivante :

$$\text{on rejette } H_0 \text{ si } T \leq -t_{n-1,\alpha} \tag{3.25}$$

où T est la statistique de test donnée par l'équation (3.23). Le *p-value* est obtenu de la façon suivante :

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[T \leq T_{obs}] \\ &= \text{ surface à gauche de } T_{obs} \text{ sous la densité } t_{n-1}. \end{aligned}$$

3.11.3 Le cas $H_1 : \mu \neq \mu_o$

Le cas bilatéral peut être traité de la façon suivante. Si on obtient un $\overline{X}$ proche de μ_o, alors on accepte H_o ; si on obtient un $\overline{X}$ loin de μ_o, alors on rejette H_o en faveur de H_1. Ce raisonnement suggère la règle de décision suivante :

$$\text{on rejette } H_0 \text{ si la distance } |\overline{X} - \mu_o| \text{ est suffisamment grande.}$$

Cette règle de décision est équivalente à la suivante :

$$\text{on rejette } H_0 \text{ si } \left| \frac{\overline{X} - \mu_o}{S/\sqrt{n}} \right| \text{ est suffisamment grand,}$$

c'est-à-dire

$$\text{on rejette } H_0 \text{ si } |T| \geq c$$

où T est donnée par l'équation (3.23). Si on veut que le seuil soit α, il faut prendre $c = t_{n-1,\alpha/2}$. La règle de décision est donc

$$\text{on rejette } H_0 \text{ si } |T| \geq t_{n-1,\alpha/2}.$$

Le calcul du *p-value* est un peu plus délicat que dans le cas unilatéral. Il faut se rappeler que le *p-value* représente la probabilité, sous H_0, d'obtenir une statistique de test *aussi extrême* que celle qu'on vient d'obtenir. Dans le présent scénario, on rejette H_0 si notre statistique de test T est loin de 0. Le *p-value* est donc donné par

$$\textit{p-value} \quad = \quad \begin{cases} \text{2 fois la surface à gauche de } T_{obs} \text{ sous la } t_{n-1} & \text{si } T_{obs} < 0 \\ \text{2 fois la surface à droite de } T_{obs} \text{ sous la } t_{n-1} & \text{si } T_{obs} \geq 0 \end{cases}$$

Exemple 14 : Aux États-Unis, le taux de cholestérol moyen chez les hommes de 40 ans est de 190 mg/dl. [Remarque : Au Canada, on mesure le taux de cholestérol sanguin en millimoles par litre (mmol/l) ; aux États-Unis on le mesure en milligrammes par décilitre (mg/dl) ; la conversion est 38.7 mg/dl = 1 mmol/l]. La loi normale est un bon modèle pour ce type de variable. La *Chicago Vegetarian Association* affirme sur son site web que les végétariens ont, en moyenne, un taux de cholestérol inférieur à la moyenne nationale. Nous allons mettre cette théorie à l'épreuve. Nous allons obtenir un échantillon aléatoire de 45 Américains végétariens de 40 ans et nous allons mesurer leur taux de cholestérol sanguin.

(a) Énoncez clairement les hypothèses ainsi que la règle de décision au seuil 1%.

(b) On mesure nos 45 taux de cholestérol. La moyenne est 178.4 et l'écart-type est 22.5. Quelle est notre décision au seuil 1% ? Quel est notre *p-value* ?

Solution : Considérons d'abord la partie (a). On doit tester $H_0 : \mu = 190$ vs $H_1 : \mu < 190$. Ici μ dénote la moyenne théorique pour la distribution des taux de cholestérol des Américains végétariens de 40 ans. L'hypothèse $H_0 : \mu = 190$ représente le statu quo. L'hypothèse $H_1 : \mu < 190$ représente l'affirmation de la *Chicago Vegetarian Association*. Nous sommes conservateurs. Nous sommes sceptiques. Nous rejetterons $H_0 : \mu = 190$ en faveur de $H_1 : \mu < 190$ seulement si nos observations sont très convaincantes. La règle de décision est

$$\text{on rejette } H_0 \text{ si } T \leq -t_{n-1,\alpha} \tag{3.26}$$

où T est la statistique de test donnée par l'équation (3.23). Ici on a

$$T = \frac{\overline{X} - 190}{S/\sqrt{45}}$$

et

$$t_{n-1,\alpha} = t_{44,0.01} = 2.4141.$$

Ce quantile d'ordre 99% de la loi de Student à 44 degrés de liberté a été obtenu à l'aide du logiciel R avec la commande `qt(0.99, 44)`. Avec la table de la loi de Student, on obtient

$$2.403 = t_{50,0.01} < t_{44,0.01} < t_{40,0.01} = 2.423.$$

Si on fait une interpolation linéaire entre ces deux valeurs, on obtient $t_{44,0.01} \approx 2.415$.

Considérons maintenant la partie (b). On obtient

$$T_{obs} = \frac{\overline{x} - 190}{s/\sqrt{45}} = \frac{178.4 - 190}{22.5/\sqrt{45}} = -3.458.$$

Cette valeur étant inférieure à notre valeur critique $-t_{44,0.01} = -2.4141$, on conclut qu'au seuil 1% on doit rejeter H_0 en faveur de H_1. Notre *p-value* est donné par

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[T \leq T_{obs}] \\ &= \text{ surface à gauche de } T_{obs} \text{ sous la densité } t_{n-1} \\ &= \text{ surface à gauche de } -3.458 \text{ sous la densité } t_{44} \\ &= 0.00061. \end{aligned}$$

Cette probabilité a été obtenue avec la commande `pt(-3.458,44)` dans le logiciel R. Voici l'interprétation de ce *p-value* : si le taux de cholestérol moyen des Américains végétariens de 40 ans était le même que celui de la population de tous les Américains de 40 ans, alors la probabilité d'obtenir une statistique T aussi extrême que celle qu'on vient d'obtenir aurait été seulement 0.00061. On ne croit pas aux miracles ! On rejette H_0 vivement !

Avec la table de la loi de Student, le mieux qu'on puisse faire est de conclure que le *p-value* est quelque part entre 0.0005 et 0.001. On arrive à cette conclusion en interpolant entre les lignes $k = 40$ et $k = 50$ de la table. Pour ces deux lignes, on note que la surface à droite de 3.458 sous la densité t_{44} est quelque part entre 0.005 et 0.001. Il en est donc de même pour la loi de Student à 44 degrés de liberté. La loi de Student étant symétrique, on conclut que la surface à gauche de -3.458 sous la densité t_{44} est quelque part entre 0.0005 et 0.001.

3.12 Tests sur une variance [1]

On suppose le même scénario d'échantillonnage qu'à la section précédente. On considère une certaine population. On s'intéresse à une certaine variable, disons la variable X. On suppose que la loi $N(\mu, \sigma^2)$ est un bon modèle pour décrire la distribution de cette variable. Cette fois-ci c'est la valeur de la variance théorique qui est mise en doute. L'hypothèse nulle prend la forme

$$H_0 : \sigma^2 = \sigma_o^2.$$

[1] On peut omettre cette section, ou la laisser en exercice, si on manque de temps.

L'hypothèse alternative sera toujours l'une des trois hypothèses suivantes :

$$H_1 : \sigma^2 > \sigma_o^2, \qquad H_1 : \sigma^2 < \sigma_o^2, \qquad H_1 : \sigma^2 \neq \sigma_o^2.$$

Puisque notre test d'hypothèse porte sur la variance théorique σ^2, il est naturel d'utiliser une règle de décision basée sur la variance échantillonnale S^2. De façon équivalente, on peut baser notre règle de décision sur la statistique

$$U = \frac{(n-1)\ S^2}{\sigma_0^2}. \tag{3.27}$$

Si H_0 est vraie, c'est-à-dire si $\sigma^2 = \sigma_o^2$, alors on sait (grâce au théorème 3.2) que cette statistique U suit la loi du khi-deux avec $n-1$ degrés de liberté.

3.12.1 Le cas $H_1 : \sigma^2 > \sigma_o^2$

Si l'hypothèse alternative est $H_1 : \sigma^2 > \sigma_o^2$, alors il est raisonnable de rejeter H_0 seulement si on observe une variance échantillonnale S^2 qui est beaucoup plus grande que la valeur σ_o^2 proposée par l'hypothèse nulle H_0. Autrement dit, notre règle de décision devrait être de la forme

$$\text{on rejette } H_0 \text{ si } S^2 \text{ est beaucoup plus grand que } \sigma_o^2.$$

Cette règle de décision peut aussi s'écrire sous la forme

$$\text{on rejette } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_o^2} \geq c$$

c'est-à-dire

$$\text{on rejette } H_0 \text{ si } U \geq c$$

où U est la statistique définie par l'équation (3.27) et où c est une constante convenablement choisie. Or on sait que si H_0 est vraie alors on a $U \sim \chi^2_{n-1}$. Il s'ensuit que pour avoir un seuil α il faut prendre $c = \chi^2_{n-1,\alpha}$. La règle de décision au seuil α est donc

$$\text{on rejette } H_0 \text{ si } U \geq \chi^2_{n-1,\alpha}.$$

Si U_{obs} dénote la valeur observée de notre statistique de test, c'est-à-dire si

$$U_{obs} = \frac{(n-1)\ s^2}{\sigma_o^2},$$

alors le *p-value* est calculé de la façon suivante :

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[U \geq U_{obs}] \\ &= \text{ surface à droite de } U_{obs} \text{ sous la densité } \chi^2_{n-1}. \end{aligned}$$

3.12.2 Le cas $H_1 : \sigma^2 < \sigma_o^2$

En raisonnant comme dans le cas précédent, on arrive aux résultats suivants. Pour tester $H_0 : \sigma^2 = \sigma_o^2$ contre $H_1 : \sigma^2 < \sigma_o^2$ au seuil α, on utilise la règle de décision suivante :

$$\text{on rejette } H_0 \text{ si } U \leq \chi^2_{n-1,1-\alpha}.$$

Une fois le U_{obs} obtenu, on calcule notre *p-value* de la façon suivante :

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[U \leq U_{obs}] \\ &= \text{ surface à gauche de } U_{obs} \text{ sous la densité } \chi^2_{n-1}. \end{aligned}$$

3.12.3 Le cas $H_1 : \sigma^2 \neq \sigma_o^2$

En s'inspirant des sections 3.11.3, 3.12.1 et 3.12.2, on arrive aux résultats suivants. Pour tester $H_0 : \sigma^2 = \sigma_o^2$ contre $H_1 : \sigma^2 \neq \sigma_o^2$ au seuil α, on utilise la règle de décision suivante :

$$\text{on rejette } H_0 \text{ si } U \leq \chi^2_{n-1,1-\frac{\alpha}{2}} \text{ ou si } U \geq \chi^2_{n-1,\frac{\alpha}{2}}.$$

Une fois le U_{obs} obtenu, on calcule notre *p-value* de la façon suivante :

$$\begin{aligned} \textit{p-value} &= \begin{cases} 2 \times \mathbb{P}_{H_0}[U \leq U_{obs}] & \text{si } U_{obs} \leq \text{ médiane de la } \chi^2_{n-1} \\ 2 \times \mathbb{P}_{H_0}[U \geq U_{obs}] & \text{si } U_{obs} \geq \text{ médiane de la } \chi^2_{n-1} \end{cases} \\ &= \begin{cases} 2 \text{ fois la surface à gauche de } U_{obs} \text{ sous la } \chi^2_{n-1} & \text{si } U_{obs} \leq \text{ médiane de } \chi^2_{n-1} \\ 2 \text{ fois la surface à droite de } U_{obs} \text{ sous la } \chi^2_{n-1} & \text{si } U_{obs} \geq \text{ médiane de } \chi^2_{n-1} \end{cases} \end{aligned}$$

Exemple 15 : On suppose que la loi normale est un bon modèle pour décrire la distribution des quotients intellectuels dans une population adulte. Les psychologues nous disent que l'écart-type de cette loi normale est égal à 16. Ceci est équivalent à dire que la variance est 256. Pour mettre cette théorie à l'épreuve, nous allons mesurer les quotients intellectuels de 25 adultes et nous allons ensuite comparer notre variance échantillonnale avec cette valeur 256 proposée par les psychologues.

(a) Énoncez clairement les hypothèses ainsi que la règle de décision au seuil 1%.

(b) On réalise notre expérience. Notre écart-type échantillonnal est $s = 19.35$. Au seuil 1%, quelle est notre décision ? Calculez le *p-value*.

Solution : Considérons d'abord la partie (a). On veut tester $H_0 : \sigma^2 = 256$ vs $H_1 : \sigma^2 \neq 256$. La règle de décision est

$$\text{on rejette } H_0 \text{ si } U \leq \chi^2_{n-1,1-\frac{\alpha}{2}} \text{ ou si } U \geq \chi^2_{n-1,\frac{\alpha}{2}}.$$

Ici on a

$$U = \frac{(n-1)S^2}{\sigma_o^2} = \frac{24\ S^2}{256}$$

et

$$\chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{24,0.995} = 9.886 \qquad \text{et} \qquad \chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{24,0.005} = 45.559.$$

Ces quantiles ont été obtenus avec le logiciel R. Pour y arriver, on a utilisé les commandes `qchisq(0.005,24)` et `qchisq(0.995,24)`. On peut aussi obtenir ces quantiles à partir de la table de la loi du khi-deux. Par exemple, en allant à la ligne $k = 24$ et la colonne $\gamma = 0.005$ de la table du khi-deux, on trouve la valeur $\chi^2_{24,0.005} = 45.56$.

Considérons maintenant la partie (b). On a obtenu $s = 19.35$. La valeur observée de notre statistique de test est donc

$$U_{obs} = \frac{24 \times (19.35)^2}{256} = 35.102.$$

On ne rejette pas H_0. Notre *p-value* est

$$\begin{aligned} \textit{p-value} &= \text{2 fois la surface à droite de } U_{obs} \text{ sous la } \chi^2_{n-1} \\ &= \text{2 fois la surface à droite de 35.102 sous la } \chi^2_{24} \\ &= 0.134. \end{aligned}$$

Pour obtenir ce *p-value*, on a utilisé la commande `2*(1-pchisq(35.102,24))` dans le logiciel R. Avec la table de la loi du khi-deux, le mieux qu'on puisse faire est de conclure que la surface à droite 35.102 sous la densité de la loi du khi-deux à 24 degrés de liberté est quelque part entre 0.05 et 0.10 et que le *p-value* est donc quelque part entre 0.10 et 0.20.

3.13 Tests sur une proportion

On veut tester l'hypothèse nulle

$$H_0 : p = p_o.$$

contre l'une des hypothèses alternatives suivantes :

$$H_1 : p > p_o, \qquad H_1 : p < p_o, \qquad H_1 : p \neq p_o.$$

Ici p représente une certaine proportion théorique. La valeur p_o spécifiée par l'hypothèse nulle nous est donnée et représente le statu quo. Notre règle de décision sera basée sur une proportion échantillonnale $\hat{p}$ calculée à partir d'un échantillon aléatoire de taille n issu de la population sous considération.

3.13.1 Le cas $H_1 : p > p_o$

Si l'hypothèse alternative est $H_1 : p > p_o$, alors il est raisonnable de rejeter H_0 seulement si on observe une proportion échantillonnale $\hat{p}$ qui est beaucoup plus grande que la valeur p_o proposée par l'hypothèse nulle H_0. Autrement dit, notre règle de décision devrait être de la forme

$$\text{on rejette } H_0 \text{ si } \hat{p} \text{ est beaucoup plus grand que } p_o.$$

Cette règle de décision peut aussi s'écrire sous la forme

$$\text{on rejette } H_0 \text{ si } \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \text{ est beaucoup plus grand que } 0$$

c'est-à-dire

$$\text{on rejette } H_0 \text{ si } \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \geq c$$

où c est une constante convenablement choisie. Grâce au théorème limite central, on sait que si H_0 est vraie et si n est suffisamment grand alors la statistique de test

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \tag{3.28}$$

suit à peu près la loi $N(0,1)$. La règle de décision au seuil α est donc

$$\text{on rejette } H_0 \text{ si } Z \geq z_\alpha$$

où Z est la statistique donnée par l'équation (3.28). Le *p-value* est alors donné par

$$\textit{p-value} = \mathbb{P}_{H_0}[Z \geq Z_{obs}] = \text{ surface à droite de } Z_{obs} \text{ sous la } N(0,1)$$

où Z_{obs} est la valeur observée de notre statistique Z.

3.13.2 Le cas $H_1 : p < p_o$

Si l'hypothèse alternative est $H_1 : p < p_o$, alors la règle de décision est

$$\text{on rejette } H_0 \text{ si } Z \leq -z_\alpha$$

où Z est la statistique donnée par l'équation (3.28). Le *p-value* est alors donné par

$$\textit{p-value} = \mathbb{P}_{H_0}[Z \leq Z_{obs}] = \text{ surface à gauche de } Z_{obs} \text{ sous la } N(0,1)$$

où Z_{obs} est la valeur observée de notre statistique Z.

3.13.3 Le cas $H_1 : p \neq p_o$

Si l'hypothèse alternative est $H_1 : p \neq p_o$, alors la règle de décision est

$$\text{on rejette } H_0 \text{ si } |Z| \geq z_{\alpha/2}$$

où Z est la statistique donnée par l'équation (3.28). Le *p-value* est alors donné par

$$\textit{p-value} \quad = \quad \begin{cases} 2 \text{ fois la surface à gauche de } Z_{obs} \text{ sous la } N(0,1) & \text{si } Z_{obs} < 0 \\ 2 \text{ fois la surface à droite de } Z_{obs} \text{ sous la } N(0,1) & \text{si } Z_{obs} \geq 0 \end{cases}$$

où Z_{obs} est la valeur observée de notre statistique Z.

Exemple 16 : Il est bien connu que le taux de germination des graines de gazon distribuées par la compagnie Després est de 80%. Selon un chercheur d'Agriculture Canada, on obtient un meilleur taux de germination si on met nos graines de gazon au congélateur durant la nuit qui précède l'ensemencement. Nous allons mettre cette théorie à l'épreuve.

(a) Énoncez les hypothèses appropriées.

(b) Pour mettre à l'épreuve la théorie du chercheur d'Agriculture Canada, nous allons réaliser l'expérience suivante. Nous allons obtenir 120 graines de gazon de la compagnie Després, nous allons mettre ces graines au congélateur pendant toute une nuit, puis le lendemain matin nous allons semer ces 120 graines dans un environnement typique. Au bout de trois semaines, nous allons compter combien de graines, parmi ces 120 graines, sont devenues des brins d'herbe. À partir de notre proportion échantillonnale, nous allons prendre une décision : ou bien on rejette H_0 en faveur de H_1, ou bien on conclut qu'il n'y a pas lieu de rejeter H_0. Énoncez clairement la règle de décision au seuil 1%.

(c) Notre expérience est maintenant terminée. Parmi nos 120 graines, il y en a 103 qui ont germé. Les 17 autres n'ont rien donné. Que doit-on conclure ?

(d) Calculez et interprétez le *p-value*.

SOLUTION : Considérons d'abord la partie (a). Les hypothèses appropriées sont $H_0 : p = 0.80$ vs $H_1 : p > 0.80$. Ici p dénote le taux de germination, c'est-à-dire la proportion théorique des graines qui vont devenir des brins d'herbe, pour la population des graines de gazon qui passent une nuit au congélateur avant d'être ensemencées. Le statu quo est que cette proportion p est la même que pour les graines qui ne passent pas une nuit au congélateur. Autrement dit, le statu quo nous dit que $p = 0.80$. L'hypothèse alternative est la théorie du chercheur d'Agriculture Canada, c'est-à-dire $p > 0.80$.

Considérons maintenant la partie (b). Puisqu'on fait un test unilatéral à droite, notre règle de décision est la suivante :

$$\text{on rejette } H_0 \text{ si } Z \geq z_\alpha$$

où Z est la statistique donnée par l'équation (3.28). Notre statistique Z est donc

$$Z = \frac{\hat{p} - 0.80}{\sqrt{\frac{0.80(1-0.80)}{120}}}$$

avec

$$\hat{p} = \frac{\text{nombre de graines qui germent}}{120}.$$

Notre valeur critique z_α est, d'après la table de la loi normale,

$$z_\alpha = z_{0.01} = 2.326.$$

Rappelons que cette valeur critique $z_{0.01}$ est le 99^e centile de la loi $N(0, 1)$. Nous pouvons obtenir ce quantile avec la commande `qnorm(0.99)` dans le logiciel R.

À la partie (c), notre proportion échantillonnale est $p = 103/120 = 0.8583$. La valeur observée de notre statistique de test est donc

$$Z_{obs} = \frac{0.8583 - 0.80}{\sqrt{\frac{0.80(1-0.80)}{120}}} = 1.60.$$

Conclusion : au seuil 1% il n'y a pas lieu de rejeter H_0. Voici le *p-value* :

$$p\text{-}value = \mathbb{P}_{H_0}[Z \geq 1.60] = 0.0548.$$

Il est tentant de raisonner de la façon suivante : *On a fait une expérience avec 120 graines de gazon. On a obtenu un taux de germination de 103/120, c'est-à-dire 85.83 pour cent. C'est presque 86%. C'est bien meilleur que 80%. Cette expérience prouve que la théorie du chercheur d'Agriculture Canada est bonne.* Voici la réponse du statisticien. Rappelez-vous que le statisticien est très conservateur. Il n'aime pas les nouvelles théories ! Il n'est pas facile à convaincre. Il raisonne donc comme suit : *Se pourrait-il que la nuit au congélateur n'ait aucun effet sur le taux de germination ? Se pourrait-il que le taux observé de 85.83% parmi nos 120 graines soit dû tout simplement à la chance ? Si la nuit au congélateur n'avait aucun effet sur les graines de gazon, quelle serait la probabilité qu'au moins 103 graines germent dans une expérience avec 120 graines ? Autrement dit, quelle serait la probabilité d'obtenir une statistique Z plus grande ou égale à 1.60 ? Réponse : 0.0548 (le p-value qu'on a obtenu). Bref, avec un taux théorique de 80%, la probabilité d'obtenir une statistique Z aussi extrême que celle qu'on a obtenu dans notre expérience est environ 5.5%. Bref, avec un taux de germination théorique de 80%, il n'est pas très surprenant que 103 graines germent parmi un groupe de 120 graines.*

REMARQUE : La démarche utilisée dans la présente section est valide à condition que la taille d'échantillon n soit suffisamment grande. En effet, pour arriver à notre valeur critique z_α ou pour calculer notre *p-value*, nous utilisons le théorème limite central pour notre proportion échantillonnale $\hat{p}$. Ce théorème nous dit que si n est *suffisamment grand* (et si H_0 est vraie), alors on a

$$\frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_0)}{n}}} \approx N(0,1). \tag{3.29}$$

Le *p-value* qu'on obtient est donc approximatif puisqu'il est basé sur l'approximation (3.29). Il est possible de calculer la valeur exacte du *p-value* en travaillant avec la loi binomiale. Reprenons l'exemple des graines de gazon. Dénotons par M le nombre de graines qui germent parmi les 120 graines de notre expérience. Si H_0 est vraie, alors on a

$$M \sim \text{binomiale}(120, 0.80).$$

Nous avons obtenu $M_{obs} = 103$. La valeur exacte de notre *p-value* est donc

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[M \geq M_{obs}] \\ &= \mathbb{P}_{H_0}[M \geq 103] \\ &= \sum_{j=103}^{120} \mathbb{P}_{H_0}[M = j] \\ &= \sum_{j=103}^{120} \binom{120}{j} (0.80)^j (0.20)^{120-j} \;=\; 0.065. \end{aligned}$$

On a obtenu cette probabilité binomiale avec la commande `1 - pbinom(102, 120, 0.80)` dans le logiciel R.

3.14 Graphe quantile-quantile et test de Shapiro et Wilk

Pour plusieurs des problèmes que nous avons étudiés jusqu'à maintenant, nous faisions l'hypothèse que la loi normale était un bon modèle pour décrire la distribution de la variable d'intérêt. Parfois c'est l'expérience qui nous dit que cette hypothèse est raisonnable. Dans la plupart des cas, on se contente d'examiner l'histogramme de nos données. Si on observe une forme de cloche symétrique, on conclut que l'hypothèse de normalité est raisonnable. Cette approche est très subjective. Il existe des méthodes plus objectives pour décider si l'hypothèse de normalité est raisonnable. Dans la présente section nous allons examiner un graphe appelé le graphe quantile-quantile gaussien (en anglais : *normal Q-Q plot*) et un test appelé le test de normalité de Shapiro et Wilk.

L'idée principale.

On dispose d'un échantillon aléatoire de taille n, disons $y_1, y_2, y_3, ..., y_n$. Calculons notre moyenne échantillonnale $\overline{y}$ et notre écart-type échantillonnal s et, pour $k = 1, 2, 3, ..., n$, posons

$$z_k = \frac{y_k - \overline{y}}{s}.$$

Notez que si l'histogramme des y_k ressemble à une loi normale, alors l'histogramme des z_k ressemble à une loi $N(0, 1)$. Plaçons ces n observations standardisées en ordre croissant, de la plus petite valeur à la plus grande valeur. La plus petite observation standardisée est dénotée $z_{(1)}$, la deuxième plus petite observation standardisée est dénotée $z_{(2)}$, la troisième plus petite observation standardisée est dénotée $z_{(3)}$, etc. On a donc $\{z_1, z_2, ..., z_n\} = \{z_{(1)}, z_{(2)}, ..., z_{(n)}\}$ et $z_{(1)} < z_{(2)} < z_{(3)} < \cdots < z_{(n)}$. Les valeurs $z_{(1)}, z_{(2)}, ..., z_{(n)}$ sont, grosso modo, les quantiles échantillonnaux d'ordre $1/(n+1), 2/(n+1), 3/(n+1)$,..., $n/(n+1)$ pour l'échantillon $z_1, z_2, ..., z_n$. Maintenant, dénotons par $u_1 < u_2 < u_3 < \cdots < u_n$ les quantiles théoriques d'ordre $1/(n+1), 2/(n+1), 3/(n+1)$,..., $n/(n+1)$ pour la loi $N(0, 1)$, c'est-à-dire les n nombres qui sont tels que sous la densité $N(0, 1)$ on a

$$\begin{aligned}
\text{surface à gauche de } u_1 &= \tfrac{1}{n+1} \\
\text{surface entre } u_1 \text{ et } u_2 &= \tfrac{1}{n+1} \\
\cdots &\quad \cdots \\
\text{surface entre } u_{n-1} \text{ et } u_n &= \tfrac{1}{n+1} \\
\text{surface à droite de } u_n &= \tfrac{1}{n+1}.
\end{aligned}$$

On peut montrer que si notre échantillon aléatoire provient d'une distribution normale, alors le graphe des n points $(u_1, z_{(1)}), (u_2, z_{(2)})$,..., $(u_n, z_{(n)})$ sera grosso modo une droite de pente 1 passant par l'origine. Si le graphe a plutôt une forme de « S allongé » ou une forme de « banane », alors il faut conclure que notre échantillon ne provient pas d'une distribution normale.

Le graphe quantile-quantile gaussien.

Le graphe quantile-quantile gaussien *(normal Q-Q plot)* est simplement le graphique des n points $(u_1, z_{(1)}), (u_2, z_{(2)})$,..., $(u_n, z_{(n)})$, sauf que, pour des raisons qui ne seront pas

présentées ici, au lieu de prendre les valeurs u_k définies ci-dessus, on prend plutôt les valeurs u_k qui sont telles que

$$\begin{aligned}
\text{surface à gauche de } u_1 &= \tfrac{5/8}{n+(1/4)} \\
\text{surface entre } u_1 \text{ et } u_2 &= \tfrac{1}{n+(1/4)} \\
\cdots &\quad \cdots \\
\text{surface entre } u_{n-1} \text{ et } u_n &= \tfrac{1}{n+(1/4)} \\
\text{surface à droite de } u_n &= \tfrac{5/8}{n+(1/4)}.
\end{aligned}$$

Notez que u_k est simplement la valeur qui est telle que la surface à gauche de u_k sous la loi $N(0,1)$ est $(k-(3/8))/(n+(1/4))$. Autrement dit, u_k est la valeur telle que

$$\mathbb{P}[Z \leq u_k] = \frac{k-(3/8)}{n+(1/4)}.$$

Dans R, on obtient ce quantile u_k avec la commande `qnorm((k-(3/8))/(n+(1/4))`. Si la distribution à partir de laquelle provient notre échantillon est bel et bien une distribution normale, les n points $(u_k, z_{(k)})$ forment à peu près une droite. Cette droite est parfois appelée la *droite de Henry*. Si notre échantillon provient d'une distribution très différente de la loi normale, les n points de notre *normal Q-Q plot* n'auront pas l'air d'une droite. Voici deux exemples illustratifs.

Exemple 17 : On obtient un échantillon de taille 13 issu d'une certaine population :

69.5 73.5 73.0 57.3 67.2 78.3 76.0 74.0 59.1 72.5 82.8 52.8 61.7

Tracez le graphe quantile-quantile gaussien et commentez.
Solution : D'abord on obtient $\overline{y} = 69.054$ et $s = 8.914$. Puis on calcule les $z_k = (y_k - \overline{y})/s$ et on les place en ordre croissant. Voici les valeurs $z_{(1)}, z_{(2)}, z_{(3)}, ..., z_{(13)}$ obtenues :

$-1.82 \quad -1.32 \quad -1.12 \quad -0.82 \quad -0.21 \quad 0.05 \quad 0.39 \quad 0.44 \quad 0.50 \quad 0.55 \quad 0.78 \quad 1.04 \quad 1.54$

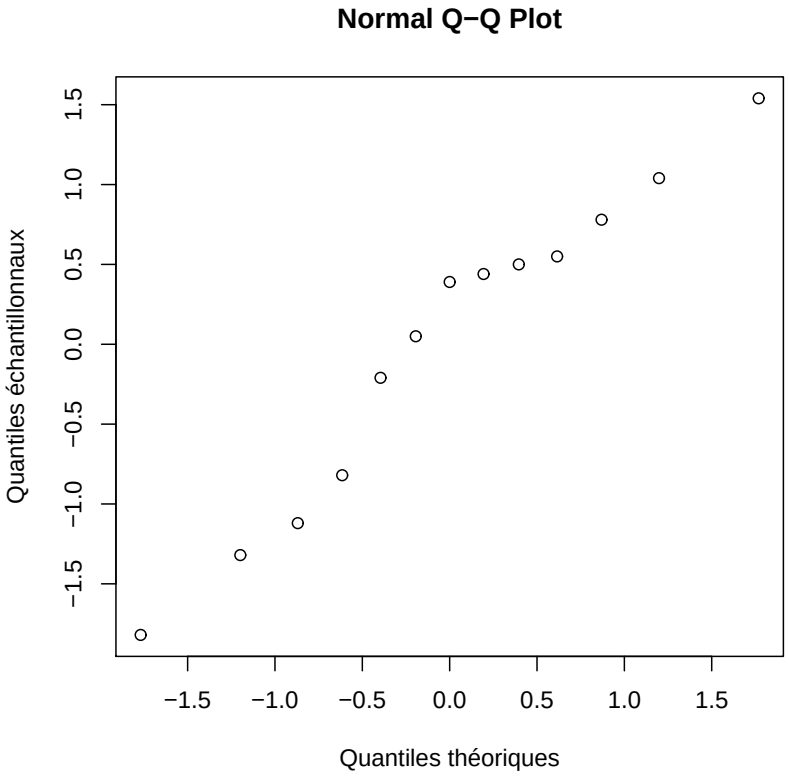
On complète ensuite le tableau suivant :

k	$z_{(k)}$	$\frac{k-(3/8)}{n+(1/4)}$	u_k
1	-1.82	0.0472	-1.67
2	-1.32	0.1226	-1.16
3	-1.12	0.1981	-0.85
4	-0.82	0.2736	-0.60
5	-0.21	0.3491	-0.39
6	0.05	0.4245	-0.19
7	0.39	0.5000	0.00
8	0.44	0.5755	0.19
9	0.50	0.6509	0.39
10	0.55	0.7264	0.60
11	0.78	0.8019	0.85
12	1.04	0.8774	1.16
13	1.54	0.9528	1.67

Par exemple, pour la ligne numéro 11, on a $k = 11$, donc $z_{(k)} = z_{(11)} = 0.78$. On obtient aussi $(k-(3/8))/(n+(1/4)) = (11-(3/8))/(13+(1/4)) = 0.8019$. Enfin, la valeur $u_k = u_{11}$ est simplement le quantile d'ordre 0.8019 de la loi $N(0,1)$, c'est-à-dire la valeur à gauche de laquelle la surface sous la $N(0,1)$ est 0.8019. Grâce ła commande `qnorm(0.8019)` on obtient $u_{11} = 0.85$. Le *normal Q-Q plot* est simplement le graphe des 13 points $(u_k, z_{(k)})$. On peut le faire à la main. On peut aussi le faire avec le logiciel R en utilisant la commande

```
qqnorm(c(69.5,73.5,73.0,57.3,67.2,78.3,76.0,74.0,59.1,72.5,82.8,52.8,61.7))
```

Voici le *normal Q-Q plot* produit par le logiciel R :



Ce graphe est suffisamment linéaire pour conclure qu'il n'y a pas lieu de rejeter l'hypothèse de normalité.

Exemple 18 : Voici un échantillon de taille 25 issu d'une certaine population. Tracez le *normal Q-Q plot* et commentez.

31.08	12.24	2.58	6.62	0.64	12.38	0.95	3.57	7.48
2.00	6.50	16.48	4.95	12.77	12.10	0.45	6.17	
2.24	0.91	0.02	0.62	10.80	2.01	3.86	9.66	

Solution : On procède comme à l'exemple 17 :

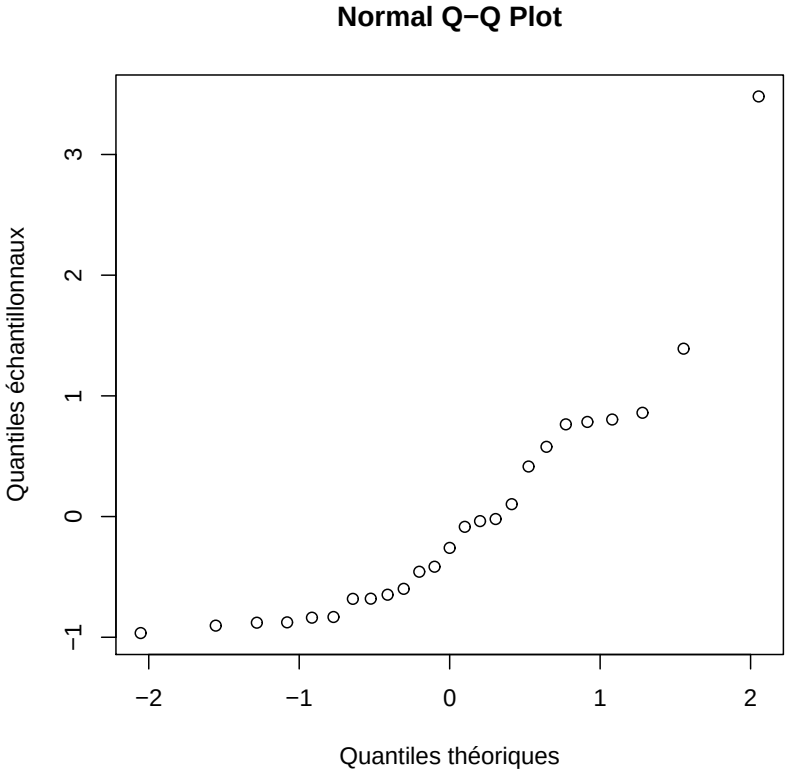
1. On calcule $\overline{y}$ et s.

2. On calcule les $z_k = (y_k - \overline{y})/s$ et on les place en ordre croissant.
3. On complète un tableau comme celui de l'exemple 17.
4. On trace le graphe des 25 points $(u_k, z_{(k)})$.

Avec le logiciel R, il suffit de taper les commandes suivantes :

```
z <- c(31.08, 12.24, 2.58, 6.62,..., 9.66)
qqnorm(z, xlab="Quantiles théoriques", ylab="Quantiles échantillonnaux")
```

Le logiciel R nous donne alors le graphe suivant :



Ce graphe est en forme de banane. Il suggère qu'on devrait rejeter l'hypothèse de normalité.

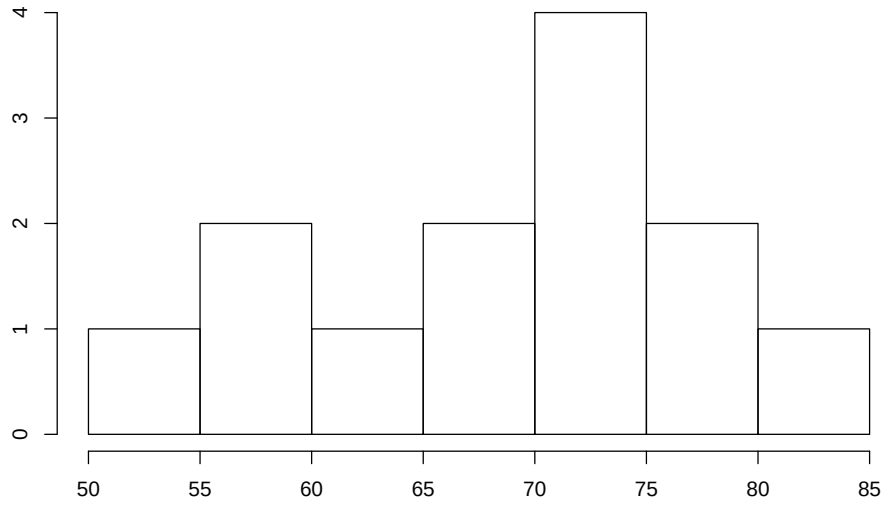
Remarque. Les Q-Q plots ci-dessus sont dessinés à partir des n points $(u_k, z_{(k)})$. On peut aussi les dessiner à partir des n points $(u_k, y_{(k)})$. Ceci a pour effet de changer l'échelle verticale du graphe. La forme du graphe demeure inchangée et le diagnostic demeure donc le même.

Le test de Shapiro et Wilk

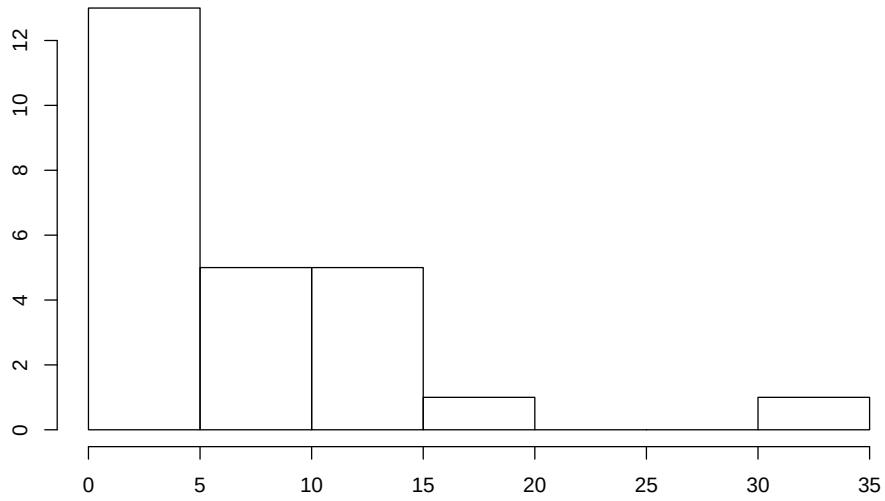
Utiliser le *normal Q-Q plot*, ça fait très savant mais dans le fond ça demeure très subjectif et ce n'est guère mieux que d'examiner l'histogramme. L'histogramme des 13 observations de l'exemple 17 est présenté au haut de la page suivante. Cet histogramme est suffisant pour nous assurer qu'il n'y a pas lieu de rejeter l'hypothèse de normalité. Nous n'avons pas vraiment besoin du *normal Q-Q plot*!

Il faut bien sûr garder à l'esprit que cet histogramme est basé sur seulement 13 observations. Avec un échantillon de si petite taille, il ne faut pas s'attendre à obtenir un histogramme parfaitement symétrique, même si l'échantillon provient réellement d'une loi normale. Pour vous en convaincre, expérimentez avec le logiciel R. Tapez à plusieurs reprises la commande `hist(rnorm(13))` et examinez les histogrammes qui en résultent. Répétez ensuite l'expérience avec $n = 100$, puis avec $n = 1000$. La commande `hist(rnorm(`n`))` produit l'histogramme d'un échantillon aléatoire de taille n issu de la loi $N(0, 1)$.

Pour l'exemple 17, voici l'histogramme des 13 observations :



Pour l'exemple 18, voici l'histogramme des 25 observations :



Cet histogramme est suffisant pour nous convaincre que la loi normale n'est pas un bon modèle. Nous n'avons pas vraiment besoin du *normal Q-Q plot*!

Ce dont nous avons besoin, c'est une règle de décision objective. Plutôt que d'examiner un graphe (histogramme, *normal Q-Q plot*, diagramme en boîte ou autre graphe non présenté ici) on voudrait une procédure basée sur une statistique numérique et une façon de calculer un *p-value* associé à cette statistique. Le *normal Q-Q plot* nous suggère une façon de faire cela ! Voici l'idée. Il suffit de calculer le coefficient de corrélation r pour les n points $(u_k, z_{(k)})$, ou, de façon équivalente, les n points $(u_k, y_{(k)})$. Puisque les séquences $(u_k; k = 1, 2, ..., n)$ et $(y_{(k)}; k = 1, 2, ..., n)$ sont en ordre croissant, ce coefficient de corrélation r sera toujours positif. S'il est proche de 1, alors on peut conclure que le *normal Q-Q plot* ressemble beaucoup à une droite et donc on accepte l'hypothèse de normalité. Si ce r est loin de 1, alors on conclut que le *normal Q-Q plot* ne ressemble pas beaucoup à une droite et donc on rejette l'hypothèse de normalité. En pratique, on calcule une certaine statistique appelée la statistique de Shapiro et Wilk. Dans le logiciel R, cette statistique est dénotée W. La façon précise de calculer W est compliquée. L'important est de savoir que ce W est à peu près égal au carré du coefficient de corrélation r ci-dessus. Cette statistique W prend toujours une valeur entre 0 et 1. Si l'hypothèse de normalité est satisfaite, W aura tendance à prendre des valeurs très proche de 1. Si l'hypothèse de normalité n'est pas satisfaite, W aura tendance à prendre des valeurs un peu moins grandes.

La règle de décision est donc

on rejette l'hypothèse de normalité si le W de Shapiro et Wilk est trop petit.

La commande `shapiro.test` du logiciel R nous donne la valeur observée de la statistique W ainsi que le *p-value* associé à cette valeur observée. Pour illustrer le test de Shapiro et Wilk, reprenons nos deux exemples.

Retour à l'exemple 17 :

On crée d'abord notre fichier de données :

```
donnee-17 <- c(69.5, 73.5, 73.0, 57.3, 67.2, 78.3, 76.0, 74.0, 59.1,
72.5, 82.8, 52.8, 61.7)
```

Puis on tape la commande

```
shapiro.test(donnee-17)
```

et R nous répond ceci :

```
Shapiro-Wilk normality test
data : donnee-17
W = 0.9511, p-value = 0.6155
```

Conclusion : Le *p-value* est très grand ! Il n'y a pas lieu de rejeter H_0 !

Retour à l'exemple 18 :

On crée d'abord notre fichier de données :

```
donnee-18 <- c(31.08, 12.24, 2.58, 6.62, 0.64, 12.38, 0.95, 3.57, 7.48,
2.00, 6.50, 16.48, 4.95, 12.77, 12.10, 0.45, 6.17, 2.24, 0.91, 0.02,
0.62, 10.80, 2.01, 3.86, 9.66)
```

Puis on tape la commande

```
shapiro.test(donnee-18)
```

et R nous répond ceci :

```
Shapiro-Wilk normality test
data : donnee-18
W = 0.8145, p-value = 0.0004
```

Conclusion : Le *p-value* est très petit ! On rejette l'hypothèse de normalité H_0 !

3.15 Exercices

NUMÉRO 1. Supposons que la loi normale avec moyenne μ et variance σ^2 soit un bon modèle pour décrire la distribution d'une certaine variable d'intérêt, disons la variable X, dans une certaine population. Supposons qu'on obtienne un échantillon aléatoire de taille 7 à partir de cette population, disons $X_1, X_2, ..., X_7$. On pose

$$A = \frac{\overline{X} - \mu}{\sigma/\sqrt{7}}, \qquad B = \frac{\overline{X} - \mu}{S/\sqrt{7}}, \qquad C = \frac{6S^2}{\sigma^2}.$$

(a) Quelle est la distribution de la variable aléatoire A ?

(b) Quelle est la distribution de la variable aléatoire B ?

(c) Quelle est la distribution de la variable aléatoire C ?

(d) Que vaut $\mathbb{E}[A]$? $\mathbb{E}[B]$? $\mathbb{E}[C]$? $\mathbb{V}\mathrm{ar}[A]$? $\mathbb{V}\mathrm{ar}[B]$? $\mathbb{V}\mathrm{ar}[C]$?

(e) Obtenez $\mathbb{P}[1.04 < A \leq 1.65]$, $\mathbb{P}[-1.34 < A \leq 0.35]$ et $\mathbb{P}[A > 1.25]$.

(f) Obtenez $\mathbb{P}[0.718 < B \leq 1.44]$, $\mathbb{P}[B \leq -0.906]$ et $\mathbb{P}[B > 3.143]$.

(g) Obtenez $\mathbb{P}[5.35 < C \leq 10.65]$, $\mathbb{P}[C \leq 1.64]$ et $\mathbb{P}[C > 2.20]$.

(h) Dessinez le graphe de la densité de la variable A et indiquez sur ce graphe la probabilité $\mathbb{P}[1.04 < A \leq 1.65]$.

(i) Dessinez le graphe de la densité de la variable B et indiquez sur ce graphe la probabilité $\mathbb{P}[0.718 < B \leq 1.44]$.

(j) Dessinez le graphe de la densité de la variable C et indiquez sur ce graphe la probabilité $\mathbb{P}[C > 2.20]$.

NUMÉRO 2. On a mesuré les poids de 16 kiwis provenant d'une ferme de Bay of Plenty en Nouvelle-Zélande. Voici ces 16 poids, en grammes :

65.06	71.44	67.93	69.02	67.28	62.34	66.23	64.16
68.56	70.45	64.91	69.90	65.52	66.75	68.54	67.90

On suppose que la loi normale avec moyenne μ et variance σ^2 est un bon modèle pour décrire la distribution des poids des kiwis de Bay of Plenty.

(a) Estimez la moyenne théorique μ à partir de ces données.

(b) Quelle est l'erreur type associée à l'estimation précédente ?

(c) Estimez la variance théorique σ^2 à partir de ces données.

(d) Quelle est l'erreur type associée à l'estimation précédente ?

(e) Obtenez un intervalle de confiance de niveau 95% pour la moyenne μ.

(f) Obtenez un intervalle de confiance de niveau 90% pour la variance σ^2.

(g) Obtenez un intervalle de confiance de niveau 90% pour l'écart-type σ.

(h) Tracez le diagramme en boîte, l'histogramme et le *normal Q-Q plot* de ces 20 observations. Est-ce qu'il vous semble raisonnable de supposer que la loi normale est un bon modèle pour la population des poids des kiwis de Bay of Plenty ?

(i) Avec le logiciel R, faites le test de normalité de Shapiro et Wilk. Quelles sont vos conclusions ?

NUMÉRO 3. On désire estimer la proportion de chevreuils d'Anticosti qui sont atteints du syndrome MO-22. Nous allons obtenir un échantillon aléatoire de chevreuils d'Anticosti et nous allons déterminer combien parmi eux sont atteints de ce syndrome. Nous allons ensuite calculer un intervalle de confiance de niveau 90% pour la proportion théorique p des chevreuils d'Anticosti atteints de ce syndrome.

(a) Quelle doit être la taille de l'échantillon si on veut être assuré que l'erreur de notre intervalle de confiance (c'est-à-dire la demi-longueur) soit au plus 0.03 ?

(b) On obtient un échantillon de 57 chevreuils. Parmi ces 57 chevreuils, il y en a 23 qui sont atteints du syndrome MO-22. Estimez la proportion p.

(c) Quelle est l'erreur type associée à l'estimation précédente ?

(d) Obtenez un intervalle de confiance de niveau 90% pour la proportion p.

NUMÉRO 4. Le fichier `Serpents.xls` (disponible sur le site web du cours) contient les poids en grammes (colonne 1) et les longueurs en centimètres (colonne 2) de 147 serpents nouveau-nés. On suppose que ces 147 serpents constituent un échantillon aléatoire de taille $n = 147$ issu de la population de tous les serpents nouveau-nés de cette espèce. On suppose que la loi normale avec moyenne μ_1 et variance σ_1^2 est un bon modèle pour décrire la distribution des poids dans cette population et que la loi normale avec moyenne μ_2 et variance σ_2^2 est un bon modèle pour décrire la distribution des longueurs dans cette population.

(a) Obtenez un intervalle de confiance de niveau 95% pour μ_1.

(b) Obtenez un intervalle de confiance de niveau 95% pour σ_1.

(c) L'hypothèse de normalité de la distribution des poids est-elle raisonnable ? Justifiez votre réponse.

(d) Obtenez un intervalle de confiance de niveau 95% pour μ_2.

(e) Obtenez un intervalle de confiance de niveau 95% pour σ_2.

(f) L'hypothèse de normalité de la distribution des longueurs est-elle raisonnable ? Justifiez votre réponse.

NUMÉRO 5. Voici 16 observations obtenues à partir d'une certaine population :

24.82	21.04	10.14	15.87	12.36	27.35	15.02	23.80
9.91	20.54	12.04	15.50	12.89	22.78	23.61	18.77

Que peut-on dire au sujet de la population ? Plus précisément, répondez aux questions suivantes :

(a) La loi normale est-elle un modèle approprié ?

(b) Que peut-on dire au sujet de la moyenne de cette population ?

(c) Que peut-on dire au sujet de l'écart-type de cette population ?

NUMÉRO 6. On obtient un échantillon aléatoire de taille n. On suppose que la loi $N(\mu, \sigma^2)$ est un bon modèle pour la population à partir de laquelle on échantillonne. Pour chaque item de la liste de gauche, associez un item de la liste de droite.

1. l'échantillon aléatoire (pas encore observé)	(a) μ
2. l'échantillon aléatoire (observé)	(b) σ^2
3. la moyenne échantillonnale (pas encore observée)	(c) $N(0,1)$
4. la distribution de la moyenne échantillonnale	(d) χ^2_{n-1}
5. la moyenne échantillonnale (observée)	(e) σ^2/n
6. la variance échantillonnale (pas encore observée)	(f) $\sigma/\sqrt{n}$
7. la variance échantillonnale (observée)	(g) $\overline{X}$
8. l'écart-type échantillonnal (pas encore observé)	(h) $2\sigma^4/(n-1)$
9. l'écart-type échantillonnal (observé)	(i) s
10. l'espérance de la moyenne échantillonnale	(j) $x_1, x_2, ..., x_n$
11. la variance de la moyenne échantillonnale	(k) $\sqrt{2}\ \sigma^2/\sqrt{n-1}$
12. l'écart-type de la moyenne échantillonnale	(l) $N(\mu, \sigma^2)$
13. l'espérance de la variance échantillonnale	(m) $X_1, X_2, ..., X_n$
14. la variance de la variance échantillonnale	(n) S^2
15. l'écart-type de la variance échantillonnale	(o) S
16. (la valeur théorique de) l'erreur type associée à $\overline{x}$	(p) p
17. (l'estimation de) l'erreur type associée à $\overline{x}$	(q) $\overline{x}$
18. (la valeur théorique de) l'erreur type associée à s^2	(r) r
19. (l'estimation de) l'erreur type associée à s^2	(s) $Q2$
20. la distribution de $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	(t) $\sqrt{2}\ s^2/\sqrt{n-1}$
21. la distribution de $\frac{\overline{X}-\mu}{S/\sqrt{n}}$	(u) $s/\sqrt{n}$
22. la distribution de $\frac{(n-1)S^2}{\sigma^2}$	(v) $N(\mu, \sigma^2/n)$
23. la moyenne de la population	(w) t_{n-1}
24. la variance de la population	(x) S
25. l'écart-type de la population	(y) s^2
26. la distribution de la population	(z) σ
27. la médiane de la population	
28. la médiane échantillonnale	

NUMÉRO 7. Le fichier `Oiseaux.xls` (disponible sur le site web du cours) contient les informations suivantes au sujet de 49 moineaux :

Colonne 1 : un code d'identification.

Colonne 2 : la variable dichotomique `survie` (oui ou non).

Colonne 3 : la variable `longueur` (du bout de la queue jusqu'à la tête en mm).

Colonne 4 : la variable `etendue` (du bout d'une aile à l'autre en mm).

Colonne 5 : la variable `tete` (longueur de la tête en mm).

Colonne 6 : la variable `humerus` (longueur de l'humérus en mm).

Colonne 7 : la variable `sternum` (longueur du sternum en mm).

Nous allons supposer que ces 49 moineaux constituent un échantillon aléatoire issu d'une certaine population de moineaux. La variable dichotomique (colonne 2) est la réponse à un certain traitement : ou bien le moineau survit au traitement, ou bien il meurt. Les cinq autres variables (colonnes 3 à 7) sont des variables quantitatives de type continu. Dans le présent exercice, nous examinons les six variables individuellement. Les liens entre ces différentes variables seront examinés plus tard.

(a) Dénotons par p la proportion théorique de survie. Voici deux façons équivalentes d'interpréter cette proportion p. Si toute la population de moineaux était soumise au traitement, la proportion de moineaux qui survivraient au traitement serait p. Si on choisit un moineau au hasard à partir de cette population et si on lui administre le traitement, la probabilité qu'il survive est p. On veut tester $H_0 : p = 1/2$ vs $H_1 : p < 1/2$.

(i) Énoncez clairement la règle de décision au seuil 5%.

(ii) Avec les données du fichier `Oiseaux.xls`, est-ce qu'on accepte ou est-ce qu'on rejette H_0 au seuil 5% ?

(iii) D'après l'approximation par la loi normale, quel est le *p-value* ?

(iv) Obtenez la valeur exacte du *p-value* à l'aide de la distribution binomiale.

(b) On se demande si la loi normale est un bon modèle pour décrire les distributions des cinq variables quantitatives. Pour chacune de ces cinq variables,

(i) dessinez le diagramme en boîte (boxplot) ;

(ii) dessinez l'histogramme ;

(iii) dessinez le graphe quantile-quantile gaussien (*normal Q-Q plot*) ;

(iv) performez le test de normalité de Shapiro et Wilk.

Quelles sont vos conclusions ?

(c) Obtenez un intervalle de confiance de niveau 95% pour la moyenne théorique de la variable `etendue`.

(d) Obtenez un intervalle de confiance de niveau 95% pour l'écart-type théorique de la variable `etendue`.

(e) Pour la variable `tete`, testez $H_0 : \mu = 31$ mm vs $H_1 : \mu \neq 31$ mm. Quel est votre *p-value* ?

(f) Pour la variable `humerus`, testez $H_0 : \sigma = 1/2$ mm vs $H_1 : \sigma > 1/2$ mm. Quel est votre *p-value* ?

Numéro 8. On considère un champ récemment ensemancé. On suppose que la loi normale est un bon modèle pour décrire la hauteur des plants dans ce champ. Certains plants sont

malades. On obtient un échantillon aléatoire de 36 plants. Parmi ces 36 plants, 5 sont malades et 31 sont en santé. Voici les 36 hauteurs, en cm :

68.6	66.8	66.7	65.8	68.4	55.7	59.0	62.3	60.1
75.6	67.0	74.1	61.3	66.6	57.5	64.4	72.1	65.0
53.1	58.7	63.2	66.9	64.8	57.6	68.9	60.7	64.7
63.5	62.4	59.6	64.3	67.9	68.9	72.1	62.4	65.5

Ces données sont disponibles sur le site web du cours. Il s'agit du fichier `Plants.xls`.

(a) Obtenez une estimation de la hauteur moyenne des plants de ce champ.

(b) Calculez l'erreur type associée à l'estimation obtenue à la partie (a).

(c) Obtenez un intervalle de confiance de niveau 95% pour la hauteur moyenne des plants de ce champ.

(d) Obtenez une estimation pour l'écart-type des hauteurs des plants dans ce champ.

(e) Obtenez un intervalle de confiance de niveau 95% pour l'écart-type des hauteurs des plants dans ce champ.

(f) Obtenez une estimation pour la proportion des plants malades dans ce champ.

(g) Calculez l'erreur type associée à l'estimation obtenue à la partie (f).

(h) Obtenez un intervalle de confiance de niveau 95% pour la proportion de plants malades dans ce champ.

(i) À la partie (h), quelle taille d'échantillon aurait-il fallu utiliser pour être sûr d'obtenir un intervalle de longueur au plus 0.04 ?

(j) Dessinez l'histogramme des 36 hauteurs. Si vous le faites à la main, choisissez des classes raisonnables.

(k) Dessinez le diagramme en boîte des 36 hauteurs.

Numéro 9.

(a) Obtenez le quantile d'ordre 0.90 de la loi normale standard.

(b) Obtenez le quantile d'ordre 0.90 de la loi normale de moyenne 50 et d'écart-type 6.

(c) Obtenez le quantile d'ordre 0.90 de la loi de Student avec 20 degrés de liberté.

(d) Obtenez le quantile d'ordre 0.90 de la loi du khi-deux avec 20 degrés de liberté.

(e) Pour chacune des 4 questions ci-dessus, dessinez avec soin le graphe de la distribution et indiquez sur ce graphe la quantile que vous avez obtenu.

Remarque : « Quantile d'ordre 0.90 » et « 90^e centile » sont deux expression qui veulent dire la même chose.

Numéro 10. Un certain génotype est supposée être présent chez 20% des plants d'une certaine population. On soupçonne que la proportion réelle, disons p, est supérieure à cette valeur théorique de 20%. On veut tester $H_0 : p = 0.20$ contre $H_1 : p > 0.20$. On obtient un échantillon aléatoire de 25 plants et on détermine le nombre N de plants qui ont le génotype en question. On utilise la règle de décision suivante :

$$\text{On rejette } H_0 \text{ si } N \geq 10.$$

(a) Quel est le seuil exact de cette règle de décision ?

(b) On réalise l'expérience et on obtient $N_{\text{obs}} = 13$. Quel est le *p-value* ?

(c) Au seuil 1%, est-ce qu'on accepte ou est-ce qu'on rejette H_0 ?

(d) Complétez la phrase suivante : Si l'hypothèse nulle avait été vraie, on se serait attendu à ce que N soit environ _____, plus ou moins environ _____.

NUMÉRO 11. On suppose que la loi normale $N(\mu, \sigma^2)$ est un bon modèle pour décrire la distribution de la quantité d'eau absorbée par un plant de tomate durant la première semaine du mois d'août (sous des conditions expérimentales contrôlées). On veut tester $H_0 : \mu = 11$ contre $H_1 : \mu \neq 11$, au seuil 5%. On réalise l'expérience suivante : à l'aide d'un dispositif complexe, on mesure la quantité d'eau absorbée par n plants choisis au hasard. À partir des résultats de cette expérience, on doit prendre une décision : accepter ou rejeter H_0.

(a) Énoncez clairement la règle de décision au seuil 5%.

On mesure les quantités d'eau absorbée par 15 plants. Voici les résultats :

10.98 10.86 8.93 8.48 9.65 9.84 11.19 12.31 12.63 9.09 8.24 9.99 7.35 10.80 9.35

(b) Calculez la valeur observée de votre statistique de test.

(c) Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette H_0 ?

(d) À l'aide de la table appropriée, que peut-on dire du *p-value* ?

(e) Refaites les partie (b), (c) et (d) avec R-Commander.

(f) La loi normale est-elle un bon modèle ? Avec R-Commander, examinez l'histogramme, le boxplot et le graphe quantile-quantile. Puis faites le test de Shapiro et Wilk.

NUMÉRO 12. Voici un échantillon de taille 17 obtenu à partir d'une certaine population :

18.22	20.83	26.88	21.42	19.36	35.67	20.40	20.39	23.89
23.21	18.43	21.57	22.91	24.54	21.62	31.33	17.70	

Est-ce que la loi normale est un bon modèle pour cette population ? Avec R-Commander, on fait le test de Shapiro et Wilk. On obtient les résultats suivants :

$$W_{\text{obs}} = 0.8465 \qquad \text{et} \qquad \textit{p-value} = 0.009487.$$

Le *p-value* étant plus petit que 1%, on rejette l'hypothèse de normalité.

(a) On examine l'histogramme, le boxplot et le graphe quantile-quantile. On constate la présence de deux observations aberrantes. Après un examen attentif, on constate que les observations 31.33 et 35.67 ont été mal enregistrées ! Les bonnes valeurs sont 21.33 et 25.67. Si on refait notre test de Shapiro et Wilk avec 31.33 et 35.67 remplacées par 21.33 et 25.67, la valeur de notre statistique de Shapiro et Wilk sera-t-elle plus petite ou plus grande que 0.8465 ? Notre *p-value* sera-t-il plus petit ou plus grand que 0.009487 ? Répondez à ces deux questions sans utiliser R-Commander.

(b) Avec R-Commander, analysez ces données, d'abord avec 31.33 et 35.67, puis avec 21.33 et 25.67. Examinez l'histogramme, le boxplot, le graphe quantile-quantile. Faites le test de Shapiro et Wilk.

Numéro 13. Un biologiste doit réaliser une expérience de grande envergure dans le but de comparer plusieurs variétés de blé. Pour son expérience, il utilise des milliers de pots de terres noires riches en divers nutriments. Pour les fins de cette étude, il est important que le concentration de potassium ne varie peu d'un pot à l'autre. Le biologiste a déterminé que l'écart-type ne doit pas dépasser la valeur 3.0. La compagnie qui lui vend ces pots de terre l'assure que l'écart-type des teneurs en potassium ne dépasse pas 3.0. On obtient un échantillon aléatoire de 21 pots et on mesure la teneur en potassium pour chacun de ces pots. Voici les résultats :

51.6	48.8	56.9	57.2	53.3	55.8	54.7
51.7	60.2	51.2	56.9	54.9	56.3	55.9
50.3	52.1	56.7	62.1	62.0	52.8	54.7

Un simple calcul nous donne l'écart-type échantillonnal $s = 3.59$. C'est plus grand que la valeur 3.0. Le biologiste devrait-il s'inquiéter ?

(a) Faites un test approprié. Quelles sont les hypothèses H_0 et H_1 ?

(b) Quelle est la règle de décision au seuil 5% ?

(c) Avec les données ci-dessus, quel est le *p-value* ?

(d) La procédure utilisée ci-dessus est valide à condition que la loi normale soit un bon modèle pour décrire la distribution des teneurs en potassium. L'hypothèse de normalité vous semble-t-elle raisonnable ? Expliquez.

Numéro 14. On a capturé 265 poissons dans le Chenal du Moine, près de Sorel. On a mesuré la concentration de mercure dans le foie de ces poissons. Ces concentrations, mesurées en ppm, se trouvent dans la première colonne du fichier `STT-1920-chap-3-no-14.xls`. Parmi ces 265 poissons, il y avait des perchaudes, des barbottes, des dorés, des brochets et des achigans. Le nom de l'espèce apparaît dans la deuxième colonne du fichier.

(a) Représentez la distribution des espèces présentes dans l'échantillon à l'aide d'un diagramme en pointes de tarte. Quelles sont les espèces les plus présentes dans notre échantillon ? Quelles sont les espèces les moins présentes dans notre échantillon ? Est-il raisonnable de tirer des conclusions du genre « *Dans le chenal du Moine, il y a plus de perchaudes que d'achigans* » ? Expliquez.

(b) À l'aide de diagrammes en boîtes juxtaposés, comparez les concentrations de mercure des différentes espèces de poisson. Quelles sont vos conclusions ?

(c) Calculez, pour chacune des 5 espèces de poisson, un intervalle de confiance de niveau 95% pour le concentration moyenne de mercure. Expliquez en quelques mots comment ces intervalles doivent être interprétés. Par exemple, pour les perchaudes on obtient l'intervalle $[86.845, 92.695]$ pour la concentration de mercure moyenne de la population des perchaudes. Quelle population de perchaudes ? Autrement dit, quelle est la population de référence ?

(d) Les longueurs des cinq intervalles obtenus à la partie (c) ne sont pas toutes égales. En général, quelles sont les deux facteurs qui influencent la longueur de l'intervalle de confiance de niveau 95% pour une moyenne ?

Chapitre 4

Estimation et tests d'hypothèses : Problèmes à deux échantillons

4.1 Introduction

Récapitulons ce que nous avons fait uu chapitre 3. Nous avons considéré les problèmes à un échantillon. Le scénario était toujours le même :

- On considère une certaine population.
- On s'intéresse à une certaine variable statistique, disons la variable X.
- On s'intéresse à un certain paramètre de la distribution de la variable X, disons le paramètre θ.
- À partir de cette population, on obtient un échantillon aléatoire de taille n, disons $X_1, X_2, ..., X_n$.

Les trois exemples de paramètres θ que nous avons étudiés en détails sont les suivants : la moyenne μ, la variance σ^2, une proportion p. Nous avons étudié les deux principaux problèmes d'inférence statistique : l'estimation d'un paramètre et les tests d'hypothèses sur un paramètre. Pour les problèmes d'estimation, nous avons vu

- comment estimer le paramètre ;
- comment calculer l'erreur type associée à une estimation du paramètre ;
- comment calculer un intervalle de confiance pour le paramètre.

Pour les problèmes de tests d'hypothèses, nous avons vu

- comment choisir une bonne règle de décision ;
- comment calculer un *p-value*.

Dans le contexte des problèmes d'estimation, les principaux concepts étudiés étaient les suivants : population, variable d'intérêt, distribution de la variable d'intérêt, paramètre, échantillon aléatoire, statistique, estimateur, estimation, erreur type, intervalle de confiance. Dans le contexte des problèmes de tests d'hypothèses, les principaux concepts étudiés étaient les suivants : population, variable d'intérêt, distribution de la variable d'intérêt, paramètre, échantillon aléatoire, hypothèse nulle, hypothèse alternative, statistique de test, règle de décision, erreur de première espèce, erreur de deuxième espèce, seuil, *p-value*.

Nous allons maintenant considérer les problèmes à deux échantillons. Commençons par un exemple simple. On veut comparer deux types d'engrais pour plants de tomates, disons l'engrais A et l'engrais B. La variable qui nous intéresse est la taille (hauteur) du plant de tomates trois semaines après germination. Nous allons faire l'expérience suivante. Nous disposons de 40 plants de tomates qui viennent tout juste de germer. Nous allons utiliser l'engrais A sur 20 plants et l'engrais B sur les 20 autres. Au bout de trois semaines, nous allons mesurer les tailles de nos 40 plants de tomates. Écrivons μ_A pour la moyenne théorique des tailles des plants de tomates soumis à l'engrais A et μ_B pour la moyenne théorique des tailles de ceux soumis à l'engrais B. Nous allons considérer les deux problèmes suivants.

Problème d'estimation. Comment estime-t-on la différence $\mu_A - \mu_B$? C'est facile : il suffit de prendre comme estimation la différence des moyennes échantillonnales, c'est-à-dire $\overline{x}_A - \overline{x}_B$. Mais, alors, quelle est l'erreur type associée à cette estimation ? Comment calcule-t-on un intervalle de confiance pour $\mu_A - \mu_B$?

Problème de test d'hypothèse. Imaginez qu'on veuille tester $H_0 : \mu_A = \mu_B$ contre l'alternative $H_1 : \mu_A > \mu_B$ (ou peut-être $H_1 : \mu_A < \mu_B$, ou peut-être même $H_1 : \mu_A \neq \mu_B$). Quelle sera notre règle de décision ? Comment calculera-t-on notre *p-value* ?

Dans les pages qui suivent, nous allons examiner plusieurs scénarios de problèmes à deux échantillons. Le problème qui nous intéresse le plus est le problème de la comparaison de deux moyennes. Mais avant de s'attaquer aux moyennes, examinons brièvement le problème de la comparaison de deux variances.

4.2 Comparaison de deux variances

4.2.1 Introduction

On suppose ici que

- $X_{1,1}, X_{1,2}, X_{1,3}, ..., X_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma_1^2)$.
- $X_{2,1}, X_{2,2}, X_{2,3}, ..., X_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma_2^2)$.
- Ces deux échantillons sont indépendants l'un de l'autre.
- Les paramètres μ_1, μ_2, σ_1^2 et σ_2^2 sont inconnus.

Notre objectif est de comparer les variances théoriques σ_1^2 et σ_2^2. Plus précisément, nous aimerions tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'une ou l'autre des hypothèses alternatives usuelles (unilatérale à droite $H_1 : \sigma_1^2 > \sigma_2^2$, unilatérale à gauche $H_1 : \sigma_1^2 < \sigma_2^2$ ou bilatérale $H_1 : \sigma_1^2 \neq \sigma_2^2$). Dans certains cas, nous aimerions plutôt estimer le rapport σ_1^2/σ_2^2, calculer l'erreur type associée à notre estimation, calculer un intervalle de confiance pour σ_1^2/σ_2^2. Avant d'aller plus loin, il faut se familiariser avec une loi de probabilité importante. Il s'agit de la *loi de Fisher*, aussi appelée la *loi de Fisher et Snedecor*.

4.2.2 La loi de Fisher

Supposons que

(i) $U \sim \chi^2_k$ (autrement dit, U suit la loi du khi-deux avec k degrés de liberté).

(ii) $V \sim \chi^2_\ell$ (autrement dit, V suit la loi du khi-deux avec ℓ degrés de liberté).

(iii) Les variables aléatoires U et V sont indépendantes.

Alors la distribution de la variable aléatoire $R = \frac{U/k}{V/\ell}$ s'appelle la loi F de Fisher avec k et ℓ degrés de liberté. Cette loi sera dénotée $F_{k,\ell}$. La densité de cette loi de probabilité est donnée par l'équation suivante :

$$f(r) = \begin{cases} \dfrac{\Gamma\left((k+\ell)/2\right)(k/\ell)^{k/2}}{\Gamma\left(k/2\right)\Gamma\left(\ell/2\right)} \dfrac{r^{(k/2)-1}}{\left(1+kr/\ell)\right)^{(k+\ell)/2}} & \text{si } r \geq 0, \\ 0 & \text{si } r < 0. \end{cases}$$

Heureusement, nous n'aurons jamais à utiliser cette formule très complexe. Pour nous, l'important est de connaître les principales propriétés de cette loi, de savoir reconnaître les situations où cette loi s'applique et de savoir comment trouver certains quantiles de cette loi à partir d'une table, d'une calculatrice ou d'un logiciel de statistique comme R.

Principales propriétés de la loi F de Fisher :

(i) La densité de la loi $F_{k,\ell}$ est en forme de cloche asymétrique étirée vers la droite.

(ii) Si $R \sim F_{k,\ell}$ et si $\ell > 2$, alors $\mu_R = \frac{\ell}{\ell-2}$.

(iii) Si $R \sim F_{k,\ell}$ et si $\ell > 4$, alors $\sigma^2_R = \frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}$

Il est bon de noter que si k et ℓ sont grands et si $k \approx \ell$, alors les formules des points (ii) et (iii) ci-dessus nous donnent $\mu_R \approx 1$ et $\sigma_R \approx 2/\sqrt{\ell}$.

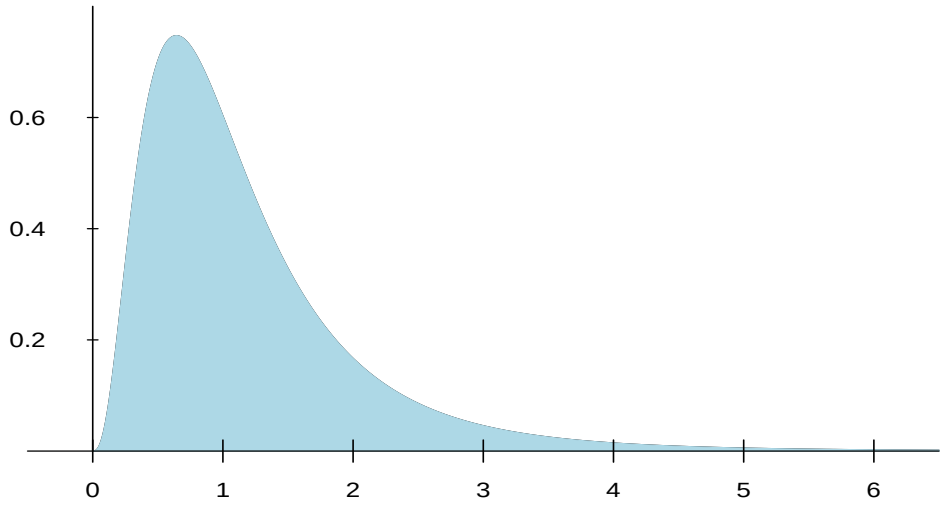
Exemple 1. Considérons la loi de Fisher avec 8 et 12 degrés de liberté. La moyenne et l'écart-type de cette distribution sont, respectivement,

$$\begin{aligned} \mu &= \frac{\ell}{\ell-2} = \frac{12}{10} = 1.20, \\ \sigma &= \sqrt{\frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}} = \sqrt{\frac{2\times 12^2(8+12-2)}{8(12-2)^2(12-4)}} = 0.90. \end{aligned}$$

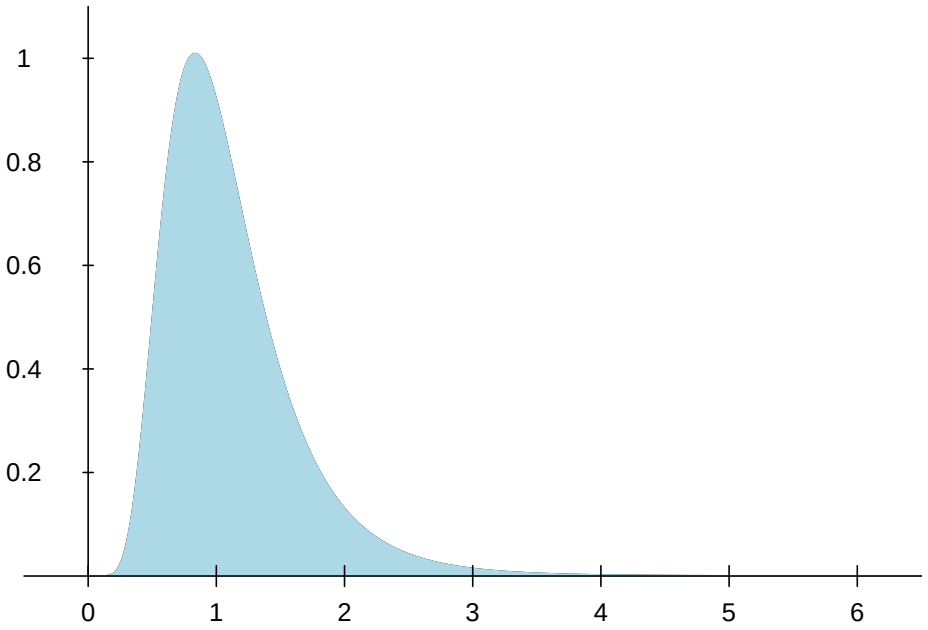
Exemple 2. Considérons maintenant la loi de Fisher avec 24 et 20 degrés de liberté. La moyenne et l'écart-type de cette distribution sont, respectivement,

$$\begin{aligned} \mu &= \frac{\ell}{\ell-2} = \frac{20}{18} \approx 1.11, \\ \sigma &= \sqrt{\frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}} = \sqrt{\frac{2\times 20^2(24+20-2)}{24(20-2)^2(20-4)}} \approx 0.52. \end{aligned}$$

Les densités de ces deux lois de Fisher sont dessinées à la page suivante.



La loi de Fisher avec 8 et 12 degrés de liberté.



La loi de Fisher avec 24 et 20 degrés de liberté.

Principale application de la loi F de Fisher : Nous savons que si les conditions énoncées au début de la présente section sont satisfaites, alors on a

(a) $\frac{(n_1-1)\ S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$,

(b) $\frac{(n_2-1)\ S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$,

(c) Ces deux variables aléatoires sont indépendantes.

Donc, avec $U = \frac{(n_1-1)\ S_1^2}{\sigma_1^2}, V = \frac{(n_2-1)\ S_2^2}{\sigma_2^2}, k = n_1 - 1$ et $\ell = n_2 - 2$, on obtient le résultat suivant :

Théorème : Sous les conditions énoncées au début de la présente section, on a

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Ce résultat peut aussi s'écrire sous la forme suivante :

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}. \tag{4.1}$$

D'une part, ce théorème met en évidence la réalité suivante : lorsqu'il s'agit de comparer les variances théoriques σ_1^2 et σ_2^2, il est plus simple de considérer le rapport σ_1^2/σ_2^2 que la différence $\sigma_1^2 - \sigma_2^2$. D'autre part, ce théorème nous permet d'obtenir un intervalle de confiance pour σ_1^2/σ_2^2 ou d'obtenir une bonne règle de décision pour tester $H_0 : \sigma_1^2 = \sigma_2^2$ (contre l'une ou l'autre des trois hypothèses alternatives usuelles). Mais d'abord, il faut savoir obtenir les quantiles de la loi de Fisher.

4.2.3 Les quantiles de la loi de Fisher

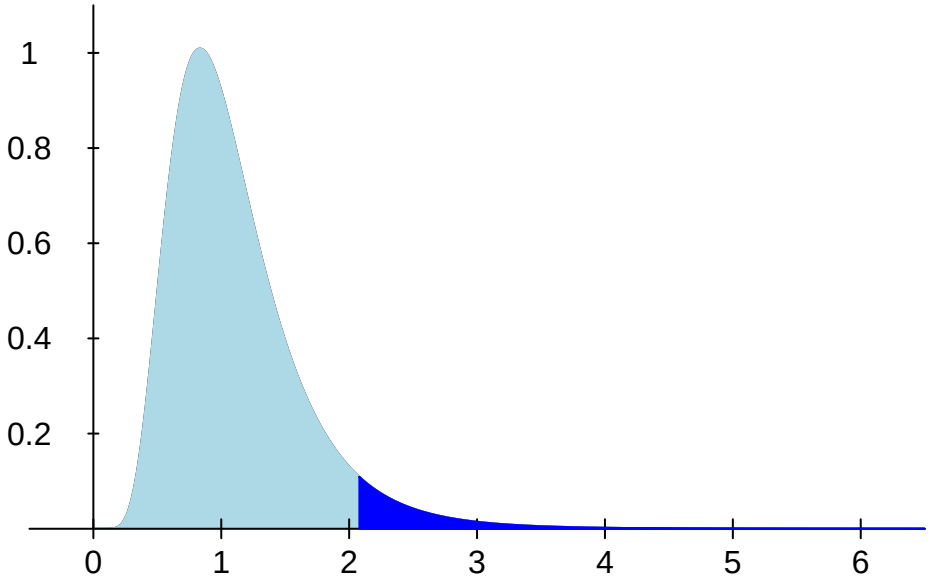
On écrit $F_{k,\ell,\gamma}$ pour dénoter le quantile d'ordre $1-\gamma$ de la loi de Fisher avec k et ℓ degrés de liberté. Donc si $R \sim F_{k,\ell}$, alors on a

$$\mathbb{P}[R \leq F_{k,\ell,\gamma}] = 1 - \gamma$$

ou, de façon équivalente,

$$\mathbb{P}[R > F_{k,\ell,\gamma}] = \gamma.$$

Le graphe suivant est celui de la loi de Fisher avec 24 degrés de liberté au numérateur et 20 degrés de liberté au dénominateur. Sur le graphe, on a indiqué le quantile d'ordre 95%, c'est-à-dire $F_{24,20,0.05} = 2.082$. Ce quantile a été obtenu avec la table de la loi de Fisher qui apparaît un peu plus loin. On peut aussi l'obtenir avec la commande `qf(0.95,24,20)` dans le logiciel R. Sur le graphe, la surface à gauche de 2.082, ombragée pâle, est égale à 0.95. La surface à droite de 2.082, ombragée foncée, est égale à 0.05.



À l'annexe A.4, on présente une table qui donne le quantile d'ordre 95% de la loi de Fisher pour différentes combinaisons des deux nombres de degrés de liberté. Notez que ces tables permettent aussi d'obtenir les quantiles d'ordre 5% grâce à la propriété suivante de la loi de Fisher :

$$F_{k,\ell,\gamma} = \frac{1}{F_{\ell,k,1-\gamma}}. \tag{4.2}$$

Par exemple, pour trouver le quantile d'ordre 5% de la loi $F_{24,20}$, on fait

$$F_{24,20,0.95} = \frac{1}{F_{20,24,0.05}} = \frac{1}{2.027} = 0.4933.$$

À l'aide de la table de la loi de Fisher présentée à l'annexe A.4 et en utilisant la propriété (4.2) ci-dessus, le lecteur devrait pouvoir vérifier les affirmations suivantes :

1. Si $R \sim F_{10,16}$, alors $\mathbb{P}[R > 2.494] = 0.05$.
2. Si $R \sim F_{10,16}$, alors $\mathbb{P}[R < 0.3536] = 0.05$.
3. Si $R \sim F_{10,16}$, alors $\mathbb{P}[0.3536 < R < 2.494] = 0.90$.

4.2.4 Intervalle de confiance de niveau $1-\alpha$ pour le rapport σ_1^2/σ_2^2

Le résultat (4.1) nous permet d'écrire

$$\mathbb{P}\left[F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{n_1-1,n_2-1,\frac{\alpha}{2}}\right] = 1-\alpha.$$

Cette équation peut aussi s'écrire de la façon suivante :

$$\mathbb{P}\left[\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2}\right] = 1-\alpha.$$

On obtient donc l'intervalle de confiance de niveau $1-\alpha$ pour le rapport σ_1^2/σ_2^2 :

$$\left(\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}\ \frac{s_1^2}{s_2^2}\ ,\ \frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}\ \frac{s_1^2}{s_2^2}\right). \tag{4.3}$$

Exemple 3. On a obtenu des échantillons aléatoires indépendants à partir de deux populations. Voici un résumé de nos observations :

Population :	1	2
Taille de l'échantillon :	12	15
Moyenne échantillonnale :	23.37	20.55
Écart-type échantillonnal :	4.83	2.21

Obtenez un intervalle de confiance de niveau 90% pour le rapport des écarts-types théoriques σ_1/σ_2. Sous quelles conditions votre intervalle est-il approprié ?

Solution. La formule (4.3) nous donne l'intervalle pour le rapport σ_1^2/σ_2^2. L'intervalle pour le rapport σ_1/σ_2 est donc

$$\left(\frac{1}{\sqrt{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}}\ \frac{s_1}{s_2}\ ,\ \frac{1}{\sqrt{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}}\ \frac{s_1}{s_2}\right).$$

Avec l'aide du logiciel R on obtient

$$\begin{aligned} F_{n_1-1,n_2-1,\frac{\alpha}{2}} &= F_{11,14,0.05} = \texttt{qf(0.95,11,14)} = 2.5655, \\ F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} &= F_{11,14,0.95} = \texttt{qf(0.05,11,14)} = 0.3651. \end{aligned}$$

L'intervalle désiré est donc

$$\left(\frac{1}{\sqrt{2.5655}}\ \frac{4.83}{2.21}\ ,\ \frac{1}{\sqrt{0.3651}}\ \frac{4.83}{2.21}\right) = (1.36, 3.62).$$

Cet intervalle est approprié à condition qu'on ait des échantillons aléatoires indépendants provenant de distributions normales.

Ci-dessus, on a utilisé R pour obtenir les valeurs $F_{11,14,0.05}$ et $F_{11,14,0.95}$. On peut aussi obtenir ces valeurs à partir de la table de la loi de Fisher qui apparaît à l'annexe A.4. Dans la table, On peut lire directement la valeur $F_{11,14,0.05} = 2.565$. Pour obtenir la valeur $F_{11,14,0.95}$, on utilise la propriété (4.2) et on obtient

$$F_{11,14,0.95} = \frac{1}{F_{14,11,0.05}} = \frac{1}{2.739} \approx 0.365.$$

4.2.5 Tests d'hypothèses sur deux variances

Supposons qu'on veuille tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'hypothèse alternative $H_1 : \sigma_1^2 > \sigma_2^2$. Ces hypothèses peuvent être réécrites sous la forme $H_0 : \sigma_1^2/\sigma_2^2 = 1$ et $H_1 : \sigma_1^2/\sigma_2^2 > 1$. La règle suivante est donc une bonne règle de décision :

$$\text{On rejette } H_0 \text{ si } \frac{S_1^2}{S_2^2} \text{ est trop grand.}$$

Or si l'hypothèse H_0 est vraie, le résultat (4.1) nous assure que la statistique S_1^2/S_2^2 suit la loi F_{n_1-1,n_2-1}. Donc, au seuil α, la règle de décision précédente prend la forme suivante :

$$\text{On rejette } H_0 \text{ si } \frac{S_1^2}{S_2^2} \geq F_{n_1-1,n_2-1,\alpha}.$$

Le cas $H_1 : \sigma_1^2 < \sigma_2^2$ et le cas $H_1 : \sigma_1^2 \neq \sigma_2^2$ peuvent être traités de façon similaire. Le tableau suivant résume les trois cas.

Alternative	On rejette H_0 si
$H_1 : \sigma_1^2 > \sigma_2^2$	$S_1^2/S_2^2 \geq F_{n_1-1,n_2-1,\alpha}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$S_1^2/S_2^2 \leq F_{n_1-1,n_2-1,1-\alpha}$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$S_1^2/S_2^2 \geq F_{n_1-1,n_2-1,\frac{\alpha}{2}}$ ou $S_1^2/S_2^2 \leq F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$

Exemple 4. Si, à l'exemple 3, on nous avait demandé de tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'hypothèse alternative $H_1 : \sigma_1^2 > \sigma_2^2$ au seuil 1%, quelle aurait été notre décision ? Calculez le *p-value*.

Solution. La règle de décision nous dit de rejeter H_0 si $S_1^2/S_2^2 \geq F_{n_1-1,n_2-1,\alpha}$. Ici on obtient $s_1^2/s_2^2 = (4.83)^2/(2.21)^2 = 4.78$ et le logiciel R nous donne

$$F_{n_1-1,n_2-1,\alpha} = F_{11,14,0.01} \;=\; 3.864.$$

Donc au seuil 1%, rejette H_0. Le *p-value* est

$$\begin{aligned} \textit{p-value} &= \text{la surface à droite de 4.78 sous la densité } F_{11,14} \\ &= 0.0038. \end{aligned}$$

4.3 Comparaison de deux moyennes : le cas où $\sigma_1^2 = \sigma_2^2$

On suppose ici que

- $X_{1,1}, X_{1,2}, X_{1,3}, ..., X_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma^2)$.
- $X_{2,1}, X_{2,2}, X_{2,3}, ..., X_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma^2)$.
- Ces deux échantillons sont indépendants l'un de l'autre.
- Les variances théoriques sont égales ; leur valeur commune est dénotée σ^2.
- Les paramètres μ_1, μ_2 et σ^2 sont inconnus.

Il s'agit du même scénario qu'à la section précédente, avec une différence : nous supposons maintenant que les variances théoriques σ_1^2 et σ_2^2 sont égales. Nous dénotons par σ^2 cette variance théorique commune aux deux populations. Notre objectif est de comparer les moyennes théoriques μ_1 et μ_2. Plus précisément, nous aimerions tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'une ou l'autre des hypothèses alternatives usuelles (unilatérale à droite $H_1 : \mu_1 > \mu_2$, unilatérale à gauche $H_1 : \mu_1 < \mu_2$ ou bilatérale $H_1 : \mu_1 \neq \mu_2$). Dans certains cas, nous aimerions plutôt estimer la différence $\mu_1 - \mu_2$, calculer l'erreur type associée à notre estimation, calculer un intervalle de confiance pour $\mu_1 - \mu_2$. Mais avant de s'attaquer à ces problèmes, il faut d'abord considérer le problème de l'estimation de la variance théorique commune à nos deux populations.

Inférence pour la variance théorique σ^2 :

À partir de nos deux échantillons, on calcule nos moyennes échantillonnales et nos variances échantillonnales de la façon usuelle :

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}, \qquad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1,i} - \overline{X}_1)^2,$$

$$\overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}, \qquad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2,i} - \overline{X}_2)^2.$$

La moyenne échantillonnale $\overline{X}_1$ sert à estimer la moyenne théorique μ_1 et la moyenne échantillonnale $\overline{X}_2$ sert à estimer la moyenne théorique μ_2. L'estimation de la variance théorique σ^2 est un peu plus délicate. Nous disposons de deux estimateurs : la variance échantillonnale S_1^2 et la variance échantillonnale S_2^2. Il est naturel de combiner ces deux estimateurs. Une façon simple de combiner ces deux estimateurs serait de prendre leur moyenne arithmétique $(S_1^2 + S_2^2)/2$. Cette approche est valide dans le cas où les tailles de nos échantillons sont identiques, c'est-à-dire dans le cas où on a $n_1 = n_2$. Si les tailles de nos échantillons ne sont pas égales, par exemple si $n_1 < n_2$, alors on s'attend à ce que la variance échantillonnale S_2^2 donne une meilleure estimation de σ^2 que la variance échantillonnale S_1^2. Plutôt que de prendre $(S_1^2 + S_2^2)/2$, on devrait alors prendre une moyenne pondérée de S_1^2 et S_2^2, avec des poids qui tiennent compte des tailles de nos échantillons. En statistique mathématique, on montre que la solution optimale est de

prendre l'estimateur

$$S_c^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{n_1-1}{n_1+n_2-2}\ S_1^2 + \frac{n_2-1}{n_1+n_2-2}\ S_2^2. \tag{4.4}$$

La statistique S_c^2 est appelée la *variance échantillonnale combinée*. Si $n_1 = n_2$, l'équation (4.4) se réduit à

$$S_c^2 = \frac{S_1^2 + S_2^2}{2} = \frac{1}{2}\ S_1^2 + \frac{1}{2}\ S_2^2.$$

Si $n_1 = 6$ et $n_2 = 11$, l'équation (4.4) se réduit à

$$S_c^2 = \frac{5\ S_1^2 + 10\ S_2^2}{15} = \frac{1}{3}\ S_1^2 + \frac{2}{3}\ S_2^2.$$

Pour le problème à un échantillon, nous avions le résultat suivant :

$$\frac{(n-1)\ S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{4.5}$$

C'est à partir de ce résultat qu'on avait obtenu (a) une formule pour l'erreur type associée à la variance, (b) l'intervalle de confiance pour la variance théorique et (c) les règles de décision pour les tests d'hypothèses sur la variance théorique. Pour la variance échantillonnale combinée, le résultat analogue au résultat (4.5) est le suivant :

$$\frac{(n_1+n_2-2)\ S_c^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}. \tag{4.6}$$

Petit truc pour se souvenir du nombre de degrés de liberté associé à S_c^2 : le nombre de degrés de liberté associé à S_1^2 est $n_1 - 1$ et celui associé à S_2^2 est $n_2 - 1$; le nombre de degrés de liberté associé à S_c^2 est donc $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. En procédant comme dans le cas du problème à un échantillon, on arrive aux résultats suivants :

1. La variance échantillonnale combinée S_c^2 est un estimateur sans biais pour la variance théorique σ^2. Autrement dit, $\mathbb{E}[S_c^2] = \sigma^2$.
2. L'erreur type associée à la variance échantillonnale combinée est donnée par

$$\sqrt{2}\ s_c^2/\sqrt{n_1+n_2-2}.$$

3. L'intervalle de confiance de niveau $1-\alpha$ pour σ^2 est donné par

$$\left(\frac{(n_1+n_2-2)\ S_c^2}{\chi^2_{n_1+n_2-2,\frac{\alpha}{2}}},\ \frac{(n_1+n_2-2)\ S_c^2}{\chi^2_{n_1+n_2-2,1-\frac{\alpha}{2}}} \right).$$

4. Pour tester $H_0 : \sigma^2 = \sigma_o^2$, on utilise la règle de décision suivante :
   - Avec $H_1 : \sigma^2 > \sigma_o^2$, la règle nous dit de rejeter H_0 si $U \geq \chi^2_{n_1+n_2-2,\alpha}$.
   - Avec $H_1 : \sigma^2 < \sigma_o^2$, la règle nous dit de rejeter H_0 si $U \leq \chi^2_{n_1+n_2-2,1-\alpha}$.
   - Avec $H_1 : \sigma^2 \neq \sigma_o^2$, on rejette H_0 si $U \geq \chi^2_{n_1+n_2-2,\frac{\alpha}{2}}$ ou si $U \leq \chi^2_{n_1+n_2-2,1-\frac{\alpha}{2}}$.

   La statistique de test est $U = (n_1+n_2-2)S_c^2/\sigma_o^2$.

Inférence pour la différence $\mu_1 - \mu_2$:

L'estimateur naturel de la différence $\mu_1 - \mu_2$ est la différence des moyennes échantillonnales $\overline{X}_1 - \overline{X}_2$. Jusqu'ici, pas de surprise. Mais qu'en est-il de l'erreur type ? Et comment calcule-t-on un intervalle de confiance ? Tout repose sur le résultat suivant :

Théorème : Sous les conditions énoncées au début de la présente section, on a

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1) \tag{4.7}$$

et

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}. \tag{4.8}$$

L'équation (4.7) est l'analogue de l'équation (2.33) du chapitre 2. L'équation (4.8) est l'analogue de l'équation (3.10) du chapitre 3. Voici les principales conséquences de ce théorème :

1. La différence des moyennes échantillonnales $\overline{X}_1 - \overline{X}_2$ est un estimateur sans biais pour la différence des moyennes théoriques $\mu_1 - \mu_2$. Autrement dit, on a

$$\mathbb{E}[\overline{X}_1 - \overline{X}_2] = \mu_1 - \mu_2.$$

2. L'erreur type associée à la différence des moyennes échantillonnales est donnée par

$$\text{erreur type } = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

3. L'intervalle de confiance de niveau $1-\alpha$ pour $\mu_1 - \mu_2$ est donné par

$$\left((\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2,\alpha/2}\ S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2,\alpha/2}\ S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right).$$

4. Pour tester $H_0 : \mu_1 = \mu_2$, on utilise la règle de décision suivante :
   - Avec $H_1 : \mu_1 > \mu_2$, la règle nous dit de rejeter H_0 si $T \geq t_{n_1+n_2-2,\alpha}$.
   - Avec $H_1 : \mu_1 < \mu_2$, la règle nous dit de rejeter H_0 si $T \leq -t_{n_1+n_2-2,\alpha}$.
   - Avec $H_1 : \mu_1 \neq \mu_2$, on rejette H_0 si $|T| \geq t_{n_1+n_2-2,\frac{\alpha}{2}}$.

   La statistique de test est $T = \frac{\overline{X}_1 - \overline{X}_2}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$.

Exemple 5. On veut comparer l'effet de l'engrais A et l'effet de l'engrais B sur la croissance des carottes. On divise notre jardin de 500 carottes en deux sections contenant chacune 250 carottes. Dans une section on utilise l'engrais A, dans l'autre section on utilise l'engrais B. La variable d'intérêt est le poids de la carotte (en grammes) au moment de la récolte. On fait les hypothèses suivantes :

- La loi normale avec moyenne μ_A est un bon modèle pour décrire la distribution des poids des carottes soumises à l'engrais A.
- La loi normale avec moyenne μ_B est un bon modèle pour décrire la distribution des poids des carottes soumises à l'engrais B.
- Ces deux lois normales ont la même variance, disons σ^2, mais cette variance est inconnue.
- Les poids des 250 carottes soumises à l'engrais A peuvent être vus comme étant un échantillon aléatoire de taille $n_A = 250$ issu de la loi $N(\mu_A, \sigma^2)$.
- Les poids des 250 carottes soumises à l'engrais B peuvent être vus comme étant un échantillon aléatoire de taille $n_B = 250$ issu de la loi $N(\mu_B, \sigma^2)$.
- Ces deux échantillons sont indépendants l'un de l'autre.

Voici un résumé des données recueillies à la fin de l'expérience, au moment de la récolte. Une partie du jardin a été détruite lorsque le responsable de l'application des engrais a fait une fausse manoeuvre avec son tracteur, détruisant ainsi 23 carottes.

Engrais :	A	B
Taille de l'échantillon :	250	227
Moyenne échantillonnale :	92.7	80.2
Écart-type échantillonnal :	4.93	4.55

(a) Calculez une estimation pour $\mu_A - \mu_B$.

(b) Calculez l'erreur type associée à l'estimation obtenue en (a).

(c) Calculez un intervalle de confiance de niveau 95% pour $\mu_A - \mu_B$.

(d) Discutez la validité des méthodes utilisées en (a), (b) et (c).

Solution.

(a) Notre estimation pour $\mu_A - \mu_B$ est $\overline{x}_A - \overline{x}_B = 92.7 - 80.2 = 12.5$.

(b) D'abord on calcule s_c et on obtient

$$s_c = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} = \sqrt{\frac{(249 \times (4.93)^2) + (226 \times (4.55)^2)}{475}} = 4.753.$$

L'erreur type associée à l'estimation obtenue en (a) est donc

$$\text{erreur type } = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 4.753 \times \sqrt{\frac{1}{250} + \frac{1}{227}} = 0.436$$

(c) On utilise l'intervalle donné à la page précédente. Il nous faut la quantité

$$t_{n_A+n_B-2,\alpha/2} = t_{475,0.025}.$$

Avec 475 degrés de liberté, la loi de Student est essentiellement identique à la loi $N(0,1)$. Donc on peut prendre $t_{475,0.025} \approx z_{0.025} = 1.96$. L'intervalle désiré est donc l'intervalle

$$\text{estimation } \pm \text{ (1.96 fois l'erreur type).}$$

On utilise l'estimation obtenue en (a) et l'erreur type obtenue en (b) et on obtient l'intervalle $(11.65, 13.35)$.

(d) Si on avait accès aux données (les poids des 250 carottes soumises à l'engrais A et ceux des 227 carottes soumises à l'engrais B), on pourrait examiner nos deux histogrammes pour voir si l'hypothèse de normalité est raisonnable. Mais les tailles d'échantillon étant grandes, l'hypothèse de normalité n'est pas essentielle (en vertu du théorème limite central). Les histogrammes nous permettraient aussi de voir si l'hypothèse d'égalité des variances est raisonnable. Mais on peut utiliser un critère plus objectif (en supposant que les deux histogrammes présentent des formes de lois normales) : le test d'égalité des variances de la section 4.2. On veut tester $H_0 : \sigma_A^2 = \sigma_B^2$ contre $H_1 : \sigma_A^2 \neq \sigma_B^2$. La valeur observée de notre statistique de test est $s_A^2/s_B^2 = 1.174$. Le *p-value* est donc

$$\begin{aligned} \textit{p-value} &= 2 \times \mathbb{P}_{H_0}[S_A^2/S_B^2 \geq 1.174] \\ &= 2 \times \text{ surface à droite de 1.174 sous la densité } F_{249,226} \\ &= 2 \times 0.1095 \\ &\approx 0.22. \end{aligned}$$

Ce *p-value* est très grand. Il n'y a pas lieu de douter de l'hypothèse d'égalité des variances théoriques.

4.4 Comparaison de deux moyennes : le cas où $\sigma_1^2 \neq \sigma_2^2$

Nous reprenons ici le problème traité à la section précédente mais nous ne supposons plus que les variances théoriques sont égales. Le scénario est donc le suivant :

On suppose que

- $X_{1,1}, X_{1,2}, X_{1,3}, ..., X_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma_1^2)$
- $X_{2,1}, X_{2,2}, X_{2,3}, ..., X_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma_2^2)$
- Ces deux échantillons sont indépendants l'un de l'autre.
- Les paramètres μ_1, μ_2, σ_1^2 et σ_2^2 sont inconnus.

À la section précédente, nos procédures statistiques (intervalle de confiance et règles de décision pour tests d'hypothèses) étaient basées sur le résultat (4.8). Dans le présent contexte, on peut utiliser le résultat suivant :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_k \tag{4.9}$$

avec

$$k = \text{ l'entier le plus proche de } \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}. \tag{4.10}$$

Détail technique. Le résultat (4.9) peut être interprété de la façon suivante : sachant $(S_1, S_2) = (s_1, s_2)$, la distribution conditionnelle de la variable aléatoire

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

est approximativement égale à la loi de Student avec k degrés de liberté, où k est l'entier le plus proche de

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

Le test de Welch. Sous les conditions énoncées au début de la présente section, on peut utiliser le résultat (4.9) pour arriver aux règles de décision suivantes pour tester $H_0 : \mu_1 = \mu_2$.

- Avec $H_1 : \mu_1 > \mu_2$, la règle nous dit de rejeter H_0 si $T' \geq t_{k,\alpha}$.
- Avec $H_1 : \mu_1 < \mu_2$, la règle nous dit de rejeter H_0 si $T' \leq -t_{k,\alpha}$.
- Avec $H_1 : \mu_1 \neq \mu_2$, on rejette H_0 si $|T'| \geq t_{k,\frac{\alpha}{2}}$.

La statistique de test est

$$T' = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

et k est donné par l'équation (4.10).

Exemple 6 : On obtient un échantillon aléatoire de $n_1 = 6$ louveteaux du Yukon et un échantillon aléatoire de $n_2 = 9$ louveteaux de Sibérie. On s'intéresse au poids des louveteaux à la naissance. On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. On peut supposer que les deux distributions sont normales mais rien ne nous permet de croire *a priori* que les variances théoriques sont égales. Voici les données :

Yukon : 4.76, 3.58, 5.04, 4.84, 4.29, 4.37.
Sibérie : 3.32, 7.53, 5.83, 8.22, 4.70, 5.20, 6.36, 9.73, 3.38.

Si on fait le test de la section 4.2 pour $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, on obtient un *p-value* égal à 0.0063. Il ne serait donc pas du tout raisonnable de supposer que $\sigma_1^2 = \sigma_2^2$ et d'appliquer la procédure présentée à la section 4.3. L'hypothèse de normalité, quant à elle, est raisonnable. C'est une hypothèse difficile à vérifier à partir d'échantillons de tailles 6 et 9. Mais les histogrammes obtenus avec le logiciel R n'ont rien d'alarmant et on sait que la loi normale est presque toujours un bon modèle pour décrire les distributions de poids

dans les populations animales. On peut donc utiliser le test de Welch. Avec les données ci-dessus, on obtient

$$\overline{x}_1 = 4.48 \qquad s_1 = 0.5253 \qquad \overline{x}_2 = 6.03 \qquad s_2 = 2.1711.$$

Pour déterminer le nombre de degrés de liberté approprié, on calcule

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \;=\; 9.35.$$

On prend donc $k = 9$. La valeur observée de notre statistique de test est

$$T'_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \;=\; -2.0535.$$

Notre *p-value* est donc

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[|T'| \geq 2.0535] \\ &= \text{ deux fois la surface à droite de 2.0535 sous la densité } t_9. \end{aligned}$$

Avec la table, on conclut que le *p-value* est entre 0.05 et 0.10. Avec le logiciel R, on conclut que le *p-value* est 0.07. Conclusion : au seuil 5%, il n'y a pas lieu de rejeter H_0.

4.5 Comparaison de deux moyennes : le cas où les données sont appariées

Voici un exemple illustratif. Supposons qu'une certaine espèce animale soit atteinte d'une maladie qui affecte les griffes des pattes avant. Deux traitements ont été proposés. On dispose de seulement $n = 15$ animaux malades sur lesquels on doit essayer nos 2 traitements afin de les comparer et de déterminer s'il y en a un qui est meilleur que l'autre.

Scénario 1 : Une approche possible serait d'utiliser le traitement A sur 8 animaux et le traitement B sur les 7 autres. On obtiendrait ainsi deux échantillons aléatoires indépendants de tailles respectives $n_1 = 8$ et $n_2 = 7$.

Scénario 2 : Une approche alternative consiste à utiliser les deux traitements sur chacun des $n = 15$ animaux : le traitement A sur une des deux pattes avant et le traitement B sur l'autre patte. Nous aurons ainsi deux échantillons aléatoires de taille $n = 15$. Mais attention ! Nos deux échantillons de tailles $n = 15$ ne seront pas des échantillons indépendants ! (On suppose qu'il est possible d'appliquer un traitement sur une patte et l'autre traitement sur l'autre patte. Ceci est possible par exemple dans le cas où les traitements sont des crèmes qu'on applique directement sur les pattes. Si les traitements étaient des médicaments que les animaux doivent avaler, le scénario 2 ne serait pas applicable).

Le scénario 1 correspond au schéma suivant. On dispose de deux paniers, le panier A et le panier B. Le panier A correspond à la population des animaux malades et la variable

d'intérêt est la réponse au traitement A alors que le panier B correspond à la population des animaux malades et la variable d'intérêt est la réponse au traitement B. Nous faisons 8 tirages à partir du panier A et 7 tirages à partir du panier B. Avec ce scénario, et en supposant que les conditions de normalité et d'égalité des variances théoriques sont vérifiées, on utilise le test présenté à la section 4.3.

Le scénario 2 correspond au schéma suivant. Il y a un seul panier. Ce panier représente la population de tous les animaux atteints de la maladie des griffes des pattes avant. Chaque boule du panier représente un animal malade. Sur chaque boule il y a deux nombres : le premier représente la réponse au traitement A et le deuxième représente la réponse au traitement B. Nous faisons $n = 15$ tirages à partir de ce panier. Au lieu d'avoir deux échantillons aléatoires indépendants, nous avons maintenant un seul échantillon aléatoire. Il s'agit d'un échantillon aléatoire *bivarié* de taille $n = 15$:

$$(X_{1,1}, X_{2,1}), (X_{1,2}, X_{2,2}), (X_{1,3}, X_{2,3}), ..., (X_{1,15}, X_{2,15}).$$

On dit alors que les données sont *appariées*. Elles sont en paires. Dans la paire $(X_{1,7}, X_{2,7})$, le $X_{1,7}$ représente la réponse de l'animal numéro 7 au traitement A alors que le $X_{2,7}$ représente la réponse de l'animal numéro 7 au traitement B. Pour analyser ces données appariées, nous allons considérer les différences

$$D_1 = X_{1,1} - X_{2,1}, \quad D_2 = X_{1,2} - X_{2,2}, \quad D_3 = X_{1,3} - X_{2,3}, ... \quad D_{15} = X_{1,15} - X_{2,15}$$

Nous allons supposer que les observations $D_1, D_2, D_3, ..., D_{15}$ constituent un échantillon aléatoire de taille $n = 15$ issu d'une population avec distribution $N(\mu_D, \sigma_D^2)$. L'hypothèse $H_0 : \mu_1 = \mu_2$ peut alors s'écrire sous la forme $H_0 : \mu_D = 0$. L'hypothèse $H_1 : \mu_1 \neq \mu_2$ prend la forme $H_1 : \mu_D \neq 0$. De même, l'hypothèse $H_1 : \mu_1 > \mu_2$ prend la forme $H_1 : \mu_D > 0$ et l'hypothèse $H_1 : \mu_1 < \mu_2$ prend la forme $H_1 : \mu_D < 0$. Nous sommes maintenant en présence d'un problème de test d'hypothèses sur la moyenne d'une seule distribution normale. Nous avons déjà vu comment traiter ce problème. Avec un échantillon bivarié de taille n, la règle de décision au seuil α prend la forme suivante :

- Avec $H_1 : \mu_D > 0$, on rejette H_0 si $T \geq t_{n-1,\alpha}$.
- Avec $H_1 : \mu_D < 0$, on rejette H_0 si $T \leq -t_{n-1,\alpha}$.
- Avec $H_1 : \mu_D \neq 0$, on rejette H_0 si $|T| \geq t_{n-1,\frac{\alpha}{2}}$.

La statistique de test est $T = \frac{\overline{D}}{S_D/\sqrt{n}}$.

Exemple 7 : Voici un exemple numérique illustratif. On veut comparer deux traitements. Les deux traitements sont utilisés sur 15 animaux malades. Chaque animal reçoit le traitement A sur une patte et le traitement B sur l'autre patte. Avec chaque animal, on lance une pièce de monnaie pour déterminer quelle patte reçoit le traitement A et quelle patte reçoit le traitement B. Nous n'avons aucune raison de soupçonner que l'un ou l'autre des deux traitements est meilleur que l'autre. Nous allons donc faire un test bilatéral :

H_0 : Les deux traitements sont équivalents.

H_1 : Un des deux traitements est meilleur que l'autre.

Notre règle de décision sera basée sur les différences

$D_j =$ (réponse de l'animal j au traitement A) $-$ (réponse de l'animal j au traitement B).

Nous allons supposer que la distribution théorique des différences est la loi $N(\mu_D, \sigma_D^2)$. Nos hypothèses peuvent donc s'écrire sous la forme

$$H_0 : \mu_D = 0 \qquad \text{et} \qquad H_1 : \mu_D \neq 0.$$

Voici les données :

Animal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Trait. A	4.5	2.8	3.7	6.0	2.9	5.6	1.7	3.7	3.6	3.8	4.1	5.2	4.7	5.4	5.8
Trait. B	3.8	2.9	4.0	5.6	2.6	5.4	1.6	3.8	3.5	3.3	3.8	5.0	4.7	4.9	5.4
Dif.	0.7	-0.1	-0.3	0.4	0.3	0.2	0.1	-0.1	0.1	0.5	0.3	0.2	0.0	0.5	0.1

La moyenne échantillonnale des différences est $\overline{d} = 0.213$. L'écart-type échantillonnal des différences est $s_D = 0.270$. La valeur observée de notre statistique de test est donc

$$T_{obs} = \frac{\overline{d}}{s_D/\sqrt{n}} = \frac{0.213}{0.270/\sqrt{15}} = 3.065.$$

Le *p-value* est donc

$$\begin{aligned} \textit{p-value} &= \text{deux fois la surface à droite de 3.065 sous la densité } t_{14} \\ &= 0.0084. \end{aligned}$$

Conclusion : au seuil 1% on rejette H_0.

Remarque : Que se serait-il passé si on n'avait pas tenu compte du fait que les données sont appariées et si on avait naïvement utilisé le test de Student présenté à la section 4.3 ? Faisons les calculs. On obtient $\overline{x}_A = 4.233$ et $s_A = 1.238$. De même on obtient $\overline{x}_B = 4.020$ et $s_B = 1.152$. La variance échantillonnale combinée est

$$s_c = \sqrt{\frac{(n_1-1)s_A^2 + (n_2-1)s_B^2}{n_1+n_2-2}} = \sqrt{\frac{(15-1)(1.238)^2 + (15-1)(1.152)^2}{15+15-2}} = 1.196.$$

La valeur observée de notre statistique de test serait donc

$$T_{obs} = \frac{\overline{x}_A - \overline{x}_B}{s_c\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} = \frac{4.233 - 4.020}{1.196\sqrt{\frac{1}{15}+\frac{1}{15}}} = 0.489.$$

Le *p-value* serait donné par

$$\begin{aligned} \textit{p-value} &= \text{deux fois la surface à droite de 0.489 sous la densité } t_{28} \\ &= 0.629. \end{aligned}$$

La conclusion serait de ne pas rejeter H_0.

4.6 Comparaison de deux proportions

On considère deux populations et on s'intéresse à une certaine variable dichotomique. Pour alléger le texte, supposons que les deux valeurs possibles de cette variable dichotomique sont *malade* et *en santé*. On s'intéresse aux proportions d'individus malades dans ces deux populations. On écrit p_1 pour la proportion d'individus malades dans la population 1 et p_2 pour la proportion d'individus malades dans la population 2.

Les données :

- On obtient un échantillon aléatoire de taille n_1 à partir de la population 1 et on écrit $\hat{p}_1$ pour la proportion échantillonnale calculée à partir de cet échantillon.
- On obtient un échantillon aléatoire de taille n_2 à partir de la population 2 et on écrit $\hat{p}_2$ pour la proportion échantillonnale calculée à partir de cet échantillon.
- On suppose que ces deux échantillons sont indépendants l'un de l'autre.

La distribution de $\hat{p}_1 - \hat{p}_2$:

On sait que si n_1 et n_2 sont suffisamment grands, alors

$$\hat{p}_1 \approx N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \hat{p}_2 \approx N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

Il s'ensuit que si n_1 et n_2 sont suffisamment grands, alors

$$\hat{p}_1 - \hat{p}_2 \;\approx\; N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

et donc

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \;\approx\; N(0,1). \tag{4.11}$$

Tout ce qui suit est basé sur le résultat (4.11).

Estimation de $p_1 - p_2$:

Pour estimer la différence $p_1 - p_2$, on utilise l'estimateur $\hat{p}_1 - \hat{p}_2$. Cet estimateur est sans biais :

$$\mathbb{E}[\hat{p}_1 - \hat{p}_2] = \mathbb{E}[\hat{p}_1] - \mathbb{E}[\hat{p}_2] = p_1 - p_2.$$

La variance de l'estimateur $\hat{p}_1 - \hat{p}_2$ est donnée par

$$\mathbb{V}\text{ar}[\hat{p}_1 - \hat{p}_2] = \mathbb{V}\text{ar}[\hat{p}_1] + \mathbb{V}\text{ar}[\hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

On peut donc calculer l'erreur type associée à l'estimation $\hat{p}_1 - \hat{p}_2$ de la façon suivante :

$$\text{erreur type} \;= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

L'intervalle de confiance de niveau $1-\alpha$ pour $p_1 - p_2$ est donné par

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\, ,\, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right).$$

Notons que cet intervalle a la forme usuelle :

$$\text{estimation } \pm \text{ quelques fois l'erreur type.}$$

Tests d'hypothèses pour comparer p_1 et p_2 :

Nous voulons tester $H_0 : p_1 = p_2$ contre l'une ou l'autre des trois hypothèses alternatives usuelles. Lorsque H_0 est vraie, le résultat (4.11) prend la forme suivante :

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \;\approx\; N(0,1) \tag{4.12}$$

où p_0 dénote la valeur commune des deux proportions théoriques. En supposant que H_0 est vraie, comment estime-t-on p_0 ? Il suffit de prendre

$$\hat{p}_0 = \frac{\text{nombre total de malades dans les 2 échantillons}}{\text{nombre total d'observations}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

À partir du résultat (4.12), on obtient

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \;\approx\; N(0,1). \tag{4.13}$$

C'est à partir de ce résultat (4.13) que nous obtenons les règles de décision suivantes pour tester $H_0 : p_1 = p_2$:

- Avec $H_1 : p_1 > p_2$, la règle nous dit de rejeter H_0 si $Z \geq z_\alpha$.
- Avec $H_1 : p_1 < p_2$, la règle nous dit de rejeter H_0 si $Z \leq -z_\alpha$.
- Avec $H_1 : p_1 \neq p_2$, on rejette H_0 si $|Z| \geq z_{\frac{\alpha}{2}}$.

Ici Z est la statistique donnée par l'équation (4.13).

Exemple 8. On s'intéresse à une certaine maladie chez les perchaudes. On se demande si la proportion de perchaudes malades est la même dans le lac St-Augustin que dans le lac St-Pierre. Nous allons tester $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$. Ici p_1 dénote la proportion de perchaudes malades dans le lac St-Augustin et p_2 dénote cette même proportion dans le lac St-Pierre. Parmi 50 perchaudes tirées du lac St-Augustin, 23 sont malades. Parmi 60 perchaudes tirées du lac St-Pierre, seulement 12 sont malades. Que doit-on conclure ?

Solution. On a $n_1 = 50, \hat{p}_1 = 0.46, n_2 = 60$ et $\hat{p}_2 = 0.20$. On obtient $\hat{p}_0 = (23+12)/(50+60) = 35/110 = 0.3182$. On obtient donc

$$Z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.46 - 0.20}{\sqrt{0.3182(1-0.3182)\left(\frac{1}{50} + \frac{1}{60}\right)}} = 2.915.$$

Le *p-value* est donc

$$p\text{-}value \ = \mathbb{P}_{H_0}[|Z| \geq 2.915] = 2 \times \mathbb{P}_{H_0}[Z \geq 2.915] = 0.0036.$$

Conclusion : au seuil 1% on rejette H_0.

EXEMPLE 9. Suite au résultat obtenu ci-dessus, on décide d'estimer $p_1 - p_2$ avec précision. Parmi 300 perchaudes tirées du lac St-Augustin (incluant les 50 perchaudes ci-dessus), 134 sont malades. Parmi 300 perchaudes tirées du lac St-Pierre (incluant les 60 perchaudes ci-dessus), 77 sont malades. Calculez l'estimation de $p_1 - p_2$. Quelle est l'erreur type associée à cette estimation ? Calculez un intervalle de confiance de niveau 95% pour $p_1 - p_2$.

SOLUTION. L'estimation de $p_1 - p_2$ est $\hat{p}_1 - \hat{p}_2 = \frac{134}{300} - \frac{77}{300} = 0.190$. L'erreur type associée à cette estimation est

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.038.$$

L'intervalle de confiance de niveau 95% est

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

c'est-à-dire $0.190 \pm 1.96 \times 0.038$, c'est-à-dire 0.190 ± 0.075, c'est-à-dire $(0.115, 0.265)$.

4.7 Le test de Wilcoxon-Mann-Whitney [1]

4.7.1 Introduction

Dans tous les problèmes que nous avons considérés jusqu'à maintenant, la distribution de la variable d'intérêt était connue à un ou deux paramètres près. En fait, dans la plupart des cas, cette distribution était simplement la loi normale. On dit alors qu'on est en présence d'un modèle statistique *paramétrique* et les méthodes d'inférence statistique qu'on utilise sont dites *méthodes paramétriques*.

Dans certains problèmes, on préfère ne faire aucune hypothèse sur la forme de la distribution de la variable d'intérêt. On parle alors de modèle statistique *non paramétrique* et les méthodes statistiques propres à ces modèles sont dites *non paramétriques*. Dans la présente section, nous considérons une méthode non paramétrique appelée le *test de la somme des rangs* de Wilcoxon-Mann-Whitney.

4.7.2 Le scénario

On considère deux populations, disons la population 1 et la population 2, et on s'intéresse à une certaine variable numérique de type continu. On veut comparer les moyennes théoriques μ_1 et μ_2. On dispose d'échantillons aléatoires indépendants et on suppose que les deux distributions théoriques ont la même forme mais possiblement des moyennes différentes. On a donc

[1] On peut omettre cette section si on manque de temps

(i) $X_{1,1}, X_{1,2}, ..., X_{1,n_1}$, un échantillon aléatoire de taille n_1 issu de la population 1.

(ii) $X_{2,1}, X_{2,2}, ..., X_{2,n_2}$, un échantillon aléatoire de taille n_2 issu de la population 2.

(iii) Ces deux échantillons sont indépendants l'un de l'autre.

(iv) Les distributions théoriques ont la même forme mais avec, possiblement, des moyennes différentes. Autrement dit, la densité de probabilité de la population 1 et la densité de probabilité de la population 2 sont identiques à une translation près.

On veut tester $H_0 : \mu_1 = \mu_2$ contre l'une ou l'autre des trois hypothèses alternatives usuelles. Il est important de noter que les quatre conditions énoncées ci-dessus sont presque les mêmes que les quatre conditions du test de Student pour comparer deux moyennes. Pour le test de Student on supposait que la loi normale était un bon modèle pour chacune de nos deux populations ; dans ce cas la condition (iv) revient à dire que ces deux lois normales ont la même variance.

4.7.3 La règle de décision

Pour fixer les idées, imaginez qu'on veuille tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. On procède de la façon suivante :

1. On place nos $n_1 + n_2$ observations en ordre croissant, de la plus petite à la plus grande.
2. On attribue à nos $n_1 + n_2$ observations les *rangs* 1 à $n_1 + n_2$ de la façon suivante : la plus petite observation reçoit le rang 1, la deuxième plus petite observation reçoit le rang 2, la troisième plus petite observation reçoit le rang 3, etc.
3. On pose $W =$ *la somme des rangs des observations issues de la population 1.*

Si H_0 est vraie, alors nos deux populations ont exactement la même distribution. Dans ce cas, il n'y a pas de raison pour que les observations issues de la population 1 aient tendance à recevoir des rangs plus élevés que celles issues de la population 2. Mais si c'est H_1 qui est vraie, alors les observations issues de la population 1 auront tendance à être plus grandes que celles issues de la populations 2. Elles auront donc tendance à recevoir des rangs plus élevés. Il est donc raisonnable d'utiliser le règle de décision suivante, proposée en 1945 par Frank Wilcoxon :

$$\text{on rejette } H_0 \text{ si } W \text{ est trop grand.} \tag{4.14}$$

Ici, « trop grand » veut dire « beaucoup plus grand que ce à quoi on devrait s'attendre si l'hypothèse nulle H_0 était vraie ». Mais alors, à quoi devrions-nous nous attendre si H_0 était vraie ? Nous répondons à cette question à la section suivante.

Remarque. La règle de décision (1) fut proposée en 1945 par Wilcoxon. La statistique W est appelée la statistique de la somme des rangs de Wilcoxon. En 1947, Mann et Whitney ont étudié le même problème et ont proposé une règle de décision qui à première vue semblait différente mais qui finalement s'est avérée être équivalente à celle de Wilcoxon. Pour cette raison, le test d'hypothèses décrit dans la présenta section est parfois appelé le test de la somme des rangs de Wilcoxon-Mann-Whitney.

4.7.4 La distribution de la statistique de Wilcoxon

Que H_0 soit vraie ou non, il est facile de voir que la plus petite valeur possible de la statistique W de Wilcoxon est la valeur

$$w_{min} = 1 + 2 + 3 + \cdots + n_1 = \frac{n_1(n_1+1)}{2}.$$

Cette valeur est obtenue si ce sont les n_1 observations issues de la population 1 qui recoivent les rangs 1, 2, 3,..., n_1. Ceci survient si et seulement si les n_1 observations issues de la population 1 sont toutes plus petites que chacune des n_2 observations issues de la population 2. De même, la plus grande valeur possible de W est la valeur

$$w_{max} = (n_2+1) + (n_2+2) + (n_2+3) + \cdots + (n_2+n_1) = \frac{n_1(n_1+1)}{2} + n_1 n_2.$$

Cette valeur est obtenue si ce sont les n_1 observations issues de la population 1 qui recoivent les rangs n_2+1, n_2+2, n_2+3,..., n_2+n_1. Ceci survient si et seulement si les n_1 observations issues de la population 1 sont toutes plus grandes que chacune des n_2 observations issues de la population 2. Finalement, l'ensemble des valeurs possibles de W est simplement l'ensemble de tous les entiers de w_{min} jusqu'à w_{max}.

Que peut-on dire de plus dans le cas où H_0 est vraie ? Bien qu'elle soit difficile à calculer, la distribution, sous H_0, de la statistique W de Wilcoxon est facile à décrire grâce au résultat suivant.

THÉORÈME. Si H_0 est vraie, alors la distribution de la statistique W de Wilcoxon est la même que la distribution de la somme des résultats de n_1 tirages sans remise fait à partir d'un panier contenant $n_1 + n_2$ boules numérotées 1, 2, 3,..., $n_1 + n_2$.

Voici quelques conséquences de ce théorème. Les démonstrations sont omises.

1. $\mathbb{E}_{H_0}[W] = \frac{n_1(n_1+n_2+1)}{2}$
2. $\mathbb{V}\text{ar}_{H_0}[W] = \frac{n_1 n_2(n_1+n_2+1)}{12}$
3. Sous H_0, la distribution de W est symétrique.
4. Si n_1 et n_2 sont grands, disons $n_1 \geq 10$ et $n_2 \geq 10$, alors, toujours sous H_0,

$$\frac{W - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \;\approx\; N(0,1)$$

c'est-à-dire

$$W \;\approx\; N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right). \tag{4.15}$$

REMARQUE 1 : Lorsqu'on calcule un *p-value* à l'aide du résultat (4.15), on utilise toujours la correction pour la continuité.

REMARQUE 2 : Dans le cas où n_1 et n_2 sont petits, il existe des tables pour la distribution de W sous H_0 et ces tables nous permettent d'obtenir nos *p-values*.

4.7.5 Un premier exemple illustratif

Voici un exemple élémentaire pour illustrer les résultats de la section précédente. Supposons que $n_1 = 3$ et $n_2 = 4$. La plus petite valeur possible et la plus grande valeur possible de W sont

$$w_{min} = 1 + 2 + 3 = 6 \qquad \text{et} \qquad w_{max} = 5 + 6 + 7 = 18.$$

L'ensemble des valeurs possibles de W est l'ensemble

$$\{6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\}.$$

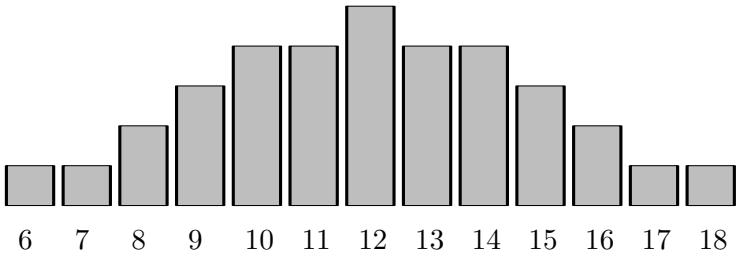
Sous H_0, l'espérance et la variance de W sont données par

$$\begin{aligned} \mathbb{E}_{H_0}[W] &= \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{3(3 + 4 + 1)}{2} = 12, \\ \mathbb{V}\text{ar}_{H_0}[W] &= \frac{n_1 n_2(n_1 + n_2 + 1)}{12} = \frac{(3 \times 4)(3 + 4 + 1)}{12} = 8. \end{aligned}$$

Notez que $\mathbb{E}_{H_0}[W]$ est exactement à mi-chemin entre w_{min} et w_{max}, conformément avec le point 3 de la page précédente. On peut trouver la distribution exacte de W sous H_0. Il suffit de considérer les $\binom{7}{3} = 35$ façons différentes de choisir les 3 rangs qui seront attribués aux 3 observations issues de la population 1. C'est ce que nous faisons dans le tableau de la page suivante. Sous H_0, ces 35 cas sont équiprobables. Pour chacun des 35 cas on calcule la valeur de W. On en déduit ensuite la distribution de W. Par exemple, dans le tableau présenté à la page suivante, on note qu'il y a quatre cas pour lesquels on obtient $W = 10$. On obtient donc $\mathbb{P}_{H_0}[W = 10] = 4/35$. De la même façon on obtient $\mathbb{P}_{H_0}[W = k]$ pour chaque valeur possible k. Voici donc, sous forme de tableau, la distribution de W (sous H_0) :

k	6	7	8	9	10	11	12	13	14	15	16	17	18
$\mathbb{P}_{H_0}[W = k]$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{5}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{3}{35}$	$\frac{2}{35}$	$\frac{1}{35}$	$\frac{1}{35}$

On note que cette distribution est bel et bien symétrique, conformément au point 3 de la page précédente. Voici un graphique de cette distribution :



Même si n_1 et n_2 ne sont pas très grands, on note que la distribution a quand même une forme de cloche symétrique, conformément au point 4 ci-dessus.

Voici le tableau qui nous a permis d'obtenir la distribution de W sous H_0 :

Rangs échantillon 1	Probabilité	Valeur de W
$1-2-3$	$1/35$	6
$1-2-4$	$1/35$	7
$1-2-5$	$1/35$	8
$1-2-6$	$1/35$	9
$1-2-7$	$1/35$	10
$1-3-4$	$1/35$	8
$1-3-5$	$1/35$	9
$1-3-6$	$1/35$	10
$1-3-7$	$1/35$	11
$1-4-5$	$1/35$	10
$1-4-6$	$1/35$	11
$1-4-7$	$1/35$	12
$1-5-6$	$1/35$	12
$1-5-7$	$1/35$	13
$1-6-7$	$1/35$	14
$2-3-4$	$1/35$	9
$2-3-5$	$1/35$	10
$2-3-6$	$1/35$	11
$2-3-7$	$1/35$	12
$2-4-5$	$1/35$	11
$2-4-6$	$1/35$	12
$2-4-7$	$1/35$	13
$2-5-6$	$1/35$	13
$2-5-7$	$1/35$	14
$2-6-7$	$1/35$	15
$3-4-5$	$1/35$	12
$3-4-6$	$1/35$	13
$3-4-7$	$1/35$	14
$3-5-6$	$1/35$	14
$3-5-7$	$1/35$	15
$3-6-7$	$1/35$	16
$4-5-6$	$1/35$	15
$4-5-7$	$1/35$	16
$4-6-7$	$1/35$	17
$5-6-7$	$1/35$	18

EXEMPLE NUMÉRIQUE : On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. La règle de décision nos dit de rejeter H_0 si W est trop grand. On obtient les données suivantes :

échantillon 1	20.6	43.7	18.5	
échantillon 2	12.7	41.7	15.5	17.4

Voici les rangs :

échantillon 1	20.6	43.7	18.5	
rang	5	7	4	
échantillon 2	12.7	41.7	15.5	17.4
rang	1	6	2	3

La valeur observée de notre statistique W est donc

$$W_{obs} = 5 + 7 + 4 = 16.$$

Notre *p-value* est donc

$$p\text{-}value \ = \mathbb{P}_{H_0}[W \geq 16] = \frac{2}{35} + \frac{1}{35} + \frac{1}{35} = \frac{4}{35} \approx 0.1143.$$

Il n'y a pas lieu de rejeter H_0.

4.7.6 Un deuxième exemple illustratif

On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. Attention : il s'agit d'un test bilatéral ! La règle de décision sera donc de la forme *on rejette H_0 si W est ou bien trop petit, ou bien trop grand*. On utilise des échantillons indépendants de tailles $n_1 = 10$ et $n_2 = 10$. Voici les observations et leurs rangs :

échantillon 1	15	55	45	11	26	66	12	27	69	39
rang	3	11	10	1	5	13	2	6	14	9
échantillon 2	81	84	61	17	97	72	28	33	85	73
rang	17	18	12	4	20	15	7	8	19	16

La valeur observée de notre statistique de Wilcoxon W est

$$W_{obs} = 3 + 11 + 10 + 1 + 5 + 13 + 2 + 6 + 14 + 9 = 74.$$

Sous H_0, on a

$$\begin{aligned} \mu_W &= \mathbb{E}_{H_0}[W] = \frac{n_1(n_1 + n_2 + 1)}{2} = 105 \\ \sigma_W &= \sqrt{\mathbb{V}\mathrm{ar}_{H_0}[W]} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{175} = 13.23. \end{aligned}$$

Comme on fait un test bilatéral, et comme la distribution de W sous H_0 est symétrique, notre *p-value* est donné par

$$p\text{-}value \ = 2 \times \mathbb{P}_{H_0}[W \leq 74].$$

À l'aide d'une table de la distribution du W de Wilcoxon, on obtient un *p-value* de 0.01854. Pour construire une telle table à la main, il faudrait procéder comme à la section 4.7.5. Pas facile ! Le tableau de la section 4.7.5 comprenait $\binom{7}{3} = 35$ lignes. Avec $n_1 = n_2 = 10$, le tableau compterait maintenant $\binom{20}{10} = 184\,756$ lignes ! Heureusement de telles tables existent.

On peut aussi obtenir une bonne approximation du *p-value* à l'aide de l'approximation gaussienne grâce au point 4 de la section 4.7.4 :

$$W \ \approx \ N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$

c'est-à-dire, toujours sous H_0, $W \approx N(105, 175)$. On obtient donc

$$\begin{aligned} p\text{-}value &= 2 \times \mathbb{P}_{H_0}[W \leq 74] \\ &\approx 2 \times \mathbb{P}\left[Z \leq \frac{74.5 - 105}{\sqrt{175}}\right] \\ &= 2 \times \mathbb{P}[Z \leq -2.3056] \\ &= 2 \times 0.0106 \ = \ 0.0212. \end{aligned}$$

Cette approximation du *p-value* est très proche de la valeur exacte de 0.01854 obtenue à partir d'une table.

Enfin, on peut faire le test de Wilcoxon avec le logiciel R. On tape la commande

```
wilcox.test(c(15,55,45,11,26,66,12,27,69,39),c(81,84,61,17,97,72,28,33,85,73))
```

et le logiciel R nous donne le *p-value* 0.01854.

Conclusion. Est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Pas facile ! Nous sommes ici dans la zone grise : au seuil 5% on rejette H_0 alors qu'au seuil 1% on accepte H_0.

4.8 Exercices

Numéro 1. Pour la loi de Fisher avec 23 et 29 degrés de liberté (c'est-à-dire 23 degrés de liberté au numérateur et 29 degrés de liberté au dénominateur), trouvez la moyenne, l'écart-type, le 5^e centile et le 95^e centile.

Numéro 2. Deux machines sont utilisées pour remplir des sacs de carottes de 2 kg. La distribution des poids des sacs remplis par la machine A est la loi $N(2.080, (0.050)^2)$. Celle des poids des sacs remplis par la machine B est la loi $N(2.050, (0.050)^2)$.

(a) Quel pourcentage des sacs remplis par la machine A pèsent moins de 2 kg ?

(b) Quel pourcentage des sacs remplis par la machine B pèsent moins de 2 kg ?

Je choisis au hasard 24 sacs remplis par la machine A et 30 sacs remplis par la machine B. Je calcule $\overline{x}_{\mathrm{A}}$, s_{A}, $\overline{x}_{\mathrm{B}}$, s_{B}.

(c) Je m'attends à ce que $\overline{x}_{\mathrm{A}} - \overline{x}_{\mathrm{B}}$ soit environ _____, plus ou moins environ _____.

(d) Je m'attends à ce que $s_{\mathrm{A}}^2/s_{\mathrm{B}}^2$ soit environ _____, plus ou moins environ _____.

Je pose $N =$ *le nombre de sacs pesant moins de 2 kg parmi les 24 sacs remplis par la machine A.*

(e) Quelle est la distribution de la variable aléatoire N ?

(f) Quelle est l'espérance de la variable aléatoire N ?

(g) Quel est l'écart-type de la variable aléatoire N ?

(h) Que vaut $\mathbb{P}[N \geq 4]$?

Numéro 3. À partir de la population A, on obtient un échantillon aléatoire de taille 16. La moyenne de ces 16 observations est 36.7 et l'écart-type est 20.60. À partir de la population B, on obtient un échantillon aléatoire de taille 21. La moyenne de ces 21 observations est 45.9 et l'écart-type est 8.20. On suppose que la loi $N(\mu_A, \sigma_A^2)$ est un bon modèle pour la population A et que la loi $N(\mu_B, \sigma_B^2)$ est un bon modèle pour la population B. Obtenez un intervalle de confiance de niveau 90% pour le rapport des écarts-types théoriques σ_A/σ_B.

Numéro 4. Je veux tester $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 > \sigma_2^2$. J'ai des échantillons indépendants de tailles n_1 et n_2. J'obtiens $s_1^2/s_2^2 = 1.887$. Est-ce que le *p-value* est plus petit dans le cas $n_1 = 25$ et $n_2 = 31$ ou dans le cas $n_1 = 38$ et $n_2 = 41$? Suggestion : de quoi a l'air la loi de Fisher avec $n_1 - 1$ et $n_2 - 1$ degrés de liberté ?

Numéro 5. On a mesuré les poids de 16 kiwis provenant d'une ferme de Bay of Plenty en Nouvelle-Zélande. Voici ces 16 poids, en grammes :

65.06	71.44	67.93	69.02	67.28	62.34	66.23	64.16
68.56	70.45	64.91	69.90	65.52	66.75	68.54	67.90

On suppose que la loi normale avec moyenne μ_1 et variance σ^2 est un bon modèle pour décrire la distribution des poids des kiwis de Bay of Plenty.

On a mesuré les poids de 18 kiwis provenant d'une ferme de la péninsule de Banks en Nouvelle-Zélande. Voici ces 18 poids, en grammes :

66.00	71.79	65.19	67.25	65.12	61.17
69.72	64.04	67.93	63.95	63.85	68.82
67.54	63.22	61.82	66.81	65.40	69.02

On suppose que la loi normale avec moyenne μ_2 et variance σ^2 est un bon modèle pour décrire la distribution des poids des kiwis de la péninsule de Banks.

On veut tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'alternative $H_1 : \mu_1 > \mu_2$.

(a) Énoncez la règle de décision au seuil 1%.

(b) Calculez votre statistique de test et comparez-la à la valeur critique appropriée au seuil 1%. Au seuil 1%, est-ce que vous acceptez ou est-ce que vous rejetez l'hypothèse nulle ?

(c) Quel est votre *p-value* ?

(d) Les hypothèses de normalité et d'égalité des variances théoriques semblent-elles raisonnables ? Justifiez votre réponse.

Numéro 6.
Voici les poids de 15 fraises provenant du champ A :

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

Voici les poids de 15 fraises provenant du champ B :

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

(a) Calculez une estimation pour $\mu_{\text{A}} - \mu_{\text{B}}$.

(b) Calculez l'erreur type associée à l'estimation obtenue en (a).

(c) Calculez un intervalle de confiance de niveau 95% pour $\mu_{\text{A}} - \mu_{\text{B}}$.

(d) La méthode utilisée à la partie (c) est valide sous certaines conditions. Énoncez ces conditions.

(e) Vérifiez si les conditions énoncées en (d) sont satisfaites.

(f) Si on avait utilisé 200 fraises du champ A et 200 fraises du champ B (au lieu de 15 et 15), quelle aurait été la longueur de l'intervalle obtenu en (c) ?

Numéro 7. Au numéro précédent, on suppose qu'on a des distributions normales de même variance, disons σ^2. Calculez un intervalle de confiance de niveau 95% pour σ^2.

Numéro 8. J'utilise l'intervalle de confiance

$$\left((\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2,\alpha/2}\ S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2,\alpha/2}\ S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \right).$$

pour une différence de moyennes théoriques, $\mu_1 - \mu_2$. Cette méthode est appropriée lorsque certaines conditions sont satisfaites. Dans chacun des cas suivants, au moins une condition n'est sans doute pas satisfaite. Laquelle ?

(a) μ_1 est le poids moyen des garçons québécois de 6 ans et μ_2 est le poids moyen des garçons québécois à la naissance. J'obtiens un échantillon aléatoire de 20 garçons de 6 ans et un échantillon aléatoire de 20 garçons nouveau-nés. $X_{1,i}$ est le poids du i^e garçon de 6 ans. $X_{2,i}$ est le poids du i^e garçon nouveau-né.

(b) μ_1 est le salaire moyen des employés de la compagnie Microsoft à Seattle et μ_2 est le salaire moyen des employés de McDonald. J'obtiens un échantillon aléatoire de 20 employés de Microsoft et un échantillon aléatoire de 20 employés de McDonald. $X_{1,i}$ est le salaire du i^e employé de Microsoft. $X_{2,i}$ est le salaire du i^e employé de McDonald.

(c) On veut comparer deux crèmes pour la peau sèche. μ_1 est la réponse moyenne à la crème A. μ_2 est la réponse moyenne à la crème B. On travaille avec 20 patients qui ont la peau sèche. Pour chaque patient, on applique la crème A sur une main et la crème B sur l'autre main. $X_{1,i}$ est la réponse à la crème A pour le i^e patient. $X_{2,i}$ est la réponse à la crème B pour le i^e patient.

NUMÉRO 9. Pour chacune des lois suivantes, trouver la moyenne, l'écart-type, le 5^e centile et le 95^e centile.

(a) La loi $N(0, 1)$.

(b) La loi $N(40, 16)$.

(c) La loi du khi-deux avec 24 degrés de liberté.

(d) La loi de Student avec 24 degrés de liberté.

(e) La loi de Fisher avec 8 et 11 degrés de liberté.

NUMÉRO 10. On veut comparer l'efficacité de 2 types de cire (ou *fart*) pour le ski de fond sur de la neige granuleuse, sous une température de -3^o C à -2^o C. Vingt-huit skieurs ont participé à notre expérience. Nos skieurs étaient tous à peu près du même niveau, tous à peu près du même poids, et ils utilisaient tous le même type de skis. Chaque skieur a skié la même boucle de 20 km de niveau intermédiaire. Pour les 12 skieurs qui ont utilisé la cire A, le temps moyen pour parcourir la boucle a été de 85.50 minutes et l'écart-type a été de 4.10 minutes. On suppose que ces 12 observations constituent un échantillon aléatoire issu d'une loi normale de moyenne μ_A et d'écart-type σ. Pour les 16 skieurs qui ont utilisé la cire B, le temps moyen pour parcourir la boucle a été de 82.25 minutes et l'écart-type a été de 4.80 minutes. On suppose que ces 16 observations constituent un échantillon aléatoire issu d'une loi normale de moyenne μ_B et d'écart-type σ, le même σ pour les 2 types de cire.

(a) Calculez un intervalle de confiance de niveau 90% pour l'écart-type σ.

(b) Calculez un intervalle de confiance de niveau 80% pour la différence $\mu_A - \mu_B$.

NUMÉRO 11. Le VO-2 MAX d'un athlète est une mesure de sa capacité aérobique. Pour les épreuves de longue distance dans les sports d'endurance comme la course à pied, le ski de fond, le cyclisme et la natation, le VO-2 MAX permet de prédire la performance de l'athlète. Par exemple, en course à pied, les coureurs ayant un VO-2 MAX de 55.0 ml/kg par minute peuvent s'attendre à courir le 10 000 mètres en 38 minutes et 6 secondes alors que ceux qui ont un VO-2 MAX de 60.0 ml/kg par minute peuvent s'attendre à courir cette même distance en 35 minutes et 22 secondes. Bien que le VO-2 MAX d'un individu soit en grande partie une affaire d'hérédité, il est possible de l'augmenter par l'entraînement.

Seize nageuses du Club de Natation Rouge et Or de l'Université Laval ont participé, en début de saison, à une étude pour comparer la nouvelle méthode d'entraînement *Immersion Totale* (voir www.totalimmersion.net) et la méthode d'entraînement Kinsella, une méthode développée dans les années 70 par John Kinsella, cet Américain qui gagna la traversée du Lac St-Jean à 6 reprises consécutives, de 1974 à 1979 (voir www.traversee.qc.ca). Les 16 nageuses ont d'abord été regroupées en 8 paires de nageuses de niveaux comparables. Pour

chacune des 8 paires, une nageuse a suivi la méthode d'entraînement Kinsella et l'autre a suivi la méthode *Immersion Totale*. Après 6 mois d'entraînement, on a mesuré, pour chaque nageuse, le gain en VO-2 MAX. Voici les résultats (en ml/kg par min) :

Numéro de la paire de nageuses :	1	2	3	4	5	6	7	8
Gain VO-2 MAX pour la nageuse ayant suivi la méthode *Immersion Totale*	2.17	1.06	1.84	2.44	3.61	2.73	1.94	2.29
Gain VO-2 MAX pour la nageuse ayant suivi la méthode Kinsella	1.35	1.16	0.32	1.81	2.28	1.01	0.80	1.71

En supposant que ces nageuses sont représentatives de l'ensemble des nageuses de niveau universitaire canadien, et en supposant que l'hypothèse de normalité est valide, calculez un intervalle de confiance de niveau 90% pour $\mu_{IT} - \mu_{JK}$. Ici, μ_{IT} représente l'espérance du gain VO-2 MAX pour les nageuses qui suivent le programme d'entraînement *Immersion Totale* et μ_{JK} représente l'espérance du gain de VO-2 MAX pour les nageuses qui suivent le programme d'entraînement de John Kinsella.

NUMÉRO 12. À l'Université de Montréal, deux types d'étudiants prennent le cours IFT-12550 *Introduction au langage C++* : les étudiants inscrits au baccalauréat en informatique et les étudiants inscrits au programme de génie informatique. On veut comparer ces 2 groupes. On suppose que la loi normale avec moyenne μ_{BI} et variance σ^2_{BI} est un bon modèle pour les notes à l'examen final de IFT-12550 pour les étudiants inscrits au baccalauréat en informatique et que la loi normale avec moyenne μ_{GI} et variance σ^2_{GI} est un bon modèle pour les notes à l'examen final de IFT-12550 pour les étudiants inscrits au programme de génie informatique. De plus, on suppose que $\sigma^2_{BI} = \sigma^2_{GI}$ et on écrit tout simplement σ^2 pour dénoter cette variance théorique commune.

(a) On veut tester $H_0 : \mu_{GI} = \mu_{BI}$ contre $H_1 : \mu_{GI} \neq \mu_{BI}$. On a obtenu les notes pour un échantillon aléatoire de 12 étudiants inscrits au baccalauréat en informatique. La moyenne de ces 12 notes est 58.4 et l'écart-type est 6.30. On a également obtenu les notes pour un échantillon aléatoire de 18 étudiants inscrits en génie informatique. La moyenne de ces 18 notes est 66.2 et l'écart-type est 5.80. Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Quel est le *p-value* ?

(b) Parmi un échantillon de 80 étudiants inscrits au baccalauréat en informatique, il y avait 36 femmes et 44 hommes. Parmi un échantillon de 100 étudiants inscrits en génie informatique, il y avait 28 femmes et 72 hommes. Testez $H_0 : p_{BI} = p_{GI}$ contre $H_1 : p_{BI} \neq p_{GI}$. Ici p_{GI} représente la proportion de femmes en génie informatique à l'Université de Montréal et où p_{BI} représente la proportion de femmes au baccalauréat en informatique à l'Université de Montréal. Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Quel est le *p-value* ?

Numéro 13. Un échantillon aléatoire de 150 étudiantes de l'Université Laval révèle que 24 d'entre elles fréquentent le PEPS régulièrement. Pour les étudiants, un échantillon aléatoire de taille 196 révèle que 49 d'entre eux fréquentent le PEPS régulièrement. Par *fréquentation régulière* on veut dire au moins 3 visites au PEPS par semaine. Calculez un intervalle de confiance de niveau 90% pour la différence $p_H - p_F$, où p_H et p_F représentent les proportions d'étudiants (H = homme) et d'étudiantes (F = femme) de l'Université Laval qui fréquentent le PEPS régulièrement.

Numéro 14. Considérons l'intervalle de confiance pour une différence de proportions avec des échantillons de même taille n :

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}} \ , \ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}} \right)$$

Quelle valeur de n nous assure que la longueur de cet intervalle sera au plus 0.04 ?

Numéro 15. On obtient d'abord un échantillon de taille 10 à partir de la population 1. La moyenne échantillonnale est 37.65 et l'écart-type échantillonnal est 12.33. On obtient ensuite un échantillon de taille 18 à partir de la population 2. La moyenne échantillonnale est 26.51 et l'écart-type échantillonnal est 6.28. On suppose que les histogrammes sont en forme de cloche symétrique. Au seuil 1%, testez $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$. Quel est votre *p-value* ? Justifiez le choix de votre règle de décision.

Numéro 16. Avec les données du numéro précédent, obtenez un intervalle de confiance de niveau 90% pour la différence des moyennes, $\mu_1 - \mu_2$. Justifiez votre démarche.

Numéro 17. Sous quelles conditions le test de la somme des rangs de Wilcoxon-Mann-Whitney est-il approprié ?

Numéro 18. Considérons le test de la somme des rangs de Wilcoxon-Mann-Whitney dans le cas où $n_1 = 2$ et $n_2 = 4$. On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. Oui, oui, je sais, le cas $n_1 = 2$ et $n_2 = 4$ n'est pas très utile en pratique car avec de telles tailles d'échantillon on ne peut pas conclure grand chose. Mais la vraie vie, c'est pour demain ! Aujourd'hui on essaie de comprendre ce qui se passe !

(a) Déterminez l'ensemble des valeurs possibles de la statistique de Wilcoxon.

(b) Calculez $\mathbb{E}_{H_0}[W]$ et $\mathbb{V}\mathrm{ar}_{H_0}[W]$.

(c) [Optionnel, mais vous devriez lire la solution] En procédant comme à la section 4.7.5 des notes de cours, obtenez la distribution exacte de la statistique W de Wilcoxon et dessinez son graphe.

Numéro 19. On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. Nos tailles d'échantillons sont égales : $n_1 = n_2$. Dénotons par n cette taille commune aux deux échantillons. Notre règle de décision est de la forme

$$\text{on rejette } H_0 \text{ si } W \text{ est trop grand.}$$

Supposons que les n observations issues de la population 1 soient toutes plus grandes que chacune des n observations issues de la population 2. Est-ce qu'on rejette H_0 ? Calculez le

p-value dans le cas $n = 1, 2, 3, ...$ À partir de quelle valeur de n obtient-on un *p-value* plus petit que 0.01 ?

Numéro 20. Le test de la somme des rangs de Wilcoxon-Mann-Whitney peut être très utile dans les scénarios où la variable d'intérêt est difficile à quantifier. Supposons qu'on veuille comparer les fleurs du champ A avec les fleurs du champ B. Ici la variable d'intérêt n'est ni le poids ni la hauteur mais bien *la beauté*. Et qui donc a dit qu'il n'y avait rien de poétique en statistique ! ? ! Nous obtenons 20 fleurs au hasard à partir du champ A et 20 fleurs au hasard à partir du champ B. Pour s'assurer qu'il n'y aura pas de biais, on a demandé à un aveugle de cueillir les 40 fleurs. En revenant au laboratoire, l'aveugle perd une fleur du champ A et deux fleurs du champ B. Bref, nos tailles d'échantillon sont $n_1 = 19$ et $n_2 = 18$. Prochaine étape : évaluer la beauté de chaque fleur ! Pas facile. Et pas nécessaire ! Il est difficile de quantifier la beauté, mais il est relativement facile de comparer deux fleurs et de déterminer laquelle des deux est la plus belle. Nous demandons à un comité d'experts de mettre nos 37 fleurs en ordre, de la moins belle à la plus belle. Ensuite, nous attribuons le rang 1 à la moins belle fleur, le rang 2 à la deuxième moins belle fleur, etc. Voici les résultats :

rang	1	2	3	4	5	6	7	8	9	10
champ de provenance	*A*	*A*	*A*	*B*	*A*	*B*	*A*	*A*	*B*	*A*
rang	11	12	13	14	15	16	17	18	19	20
champ de provenance	*A*	*A*	*B*	*A*	*A*	*B*	*A*	*A*	*B*	*A*
rang	21	22	23	24	25	26	27	28	29	30
champ de provenance	*B*	*A*	*B*	*A*	*B*	*B*	*B*	*A*	*B*	*B*
rang	31	32	33	34	35	36	37			
champ de provenance	*A*	*B*	*B*	*A*	*B*	*B*	*B*			

(a) Exprimez H_0 et H_1 en quelques mots.

(b) Quelle est la valeur observée de la statistique de Wilcoxon ?

(c) Si H_0 était vraie, à quoi devrait-on s'attendre ? Autrement dit, complétez la phrase suivante : *Si H_0 était vraie, je m'attendrais à ce que le W de Wilcoxon soit environ* _____, *plus ou moins environ* _____. Autrement dit, calculez

$$\mathbb{E}_{H_0}[W] \qquad \text{et} \qquad \sqrt{\mathbb{V}\text{ar}_{H_0}[W]}.$$

(d) Déterminez l'ensemble des valeurs possibles de la statistique W.

(e) D'après les résultats obtenus au points (c) et (d), est-ce que le W_{obs} obtenue en (b) vous semble cohérent ou incohérent avec H_0 ?

(f) À l'aide de l'approximation gaussienne de la distribution de W sous H_0, et en utilisant la correction pour la continuité, obtenez le *p-value* approprié.

Chapitre 5

Introduction à l'analyse de la variance

5.1 Introduction

Jusqu'à maintenant, nous avons considéré les problèmes à un échantillon et les problèmes à deux échantillons. Dans le présent chapitre, nous allons considérer certains problèmes à I échantillons. En particulier, nous allons étudier le test de Fisher pour l'égalité des moyennes de I populations. Ce test est une généralisation du test bilatéral de Student pour l'égalité des moyennes de deux populations lorsqu'on dispose d'échantillons aléatoires indépendants.

Le scénario :

On considère I populations, disons la population 1, la population 2,..., la population I. On s'intéresse à une certaine variable quantitative de type continue. On veut comparer les moyennes théoriques $\mu_1, \mu_2, ..., \mu_I$. On dispose d'échantillons aléatoires indépendants et on suppose que les I distributions théoriques sont des lois normales avec la même variance mais possiblement avec des moyennes différentes. On suppose donc que les conditions suivantes sont satisfaites :

- $Y_{1,1}, Y_{1,2}, Y_{1,3}, ..., Y_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma^2)$.
- $Y_{2,1}, Y_{2,2}, Y_{2,3}, ..., Y_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma^2)$.

  $\vdots$
- $Y_{I,1}, Y_{I,2}, Y_{I,3}, ..., Y_{I,n_I}$ est un échantillon aléatoire de taille n_I issu d'une population avec distribution $N(\mu_I, \sigma^2)$.
- Ces I échantillons aléatoires sont indépendants les uns des autres.
- Les variances théoriques sont égales ; la valeur commune est notée σ^2.
- Les paramètres $\mu_1, \mu_2, ..., \mu_I$ et σ^2 sont inconnus.

5.2 L'estimation des paramètres du modèle

Pour estimer la moyenne de la population i on utilise la moyenne des observations provenant de la population i. Autrement dit, pour estimer la moyenne théorique μ_i, on utilise la moyenne échantillonnale

$$\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

Le « point » qui apparaît juste à côté du i dans la notation $\overline{Y}_{i\cdot}$ signifie que cette moyenne échantillonnale a été calculée en fixant le premier indice et en faisant la moyenne sur toutes les valeurs de l'indice j. Nous verrons plus loin que ce type de notation est très pratique. Puisque nos observations $Y_{i,1}, Y_{i,2}, Y_{i,3}, ..., Y_{i,n_i}$ constituent un échantillon aléatoire de taille n_i issu d'une population avec distribution $N(\mu_i, \sigma^2)$, on a le résultat suivant :

$$\overline{Y}_{i\cdot} \sim N(\mu_i, \sigma^2/n_i). \tag{5.1}$$

Pour estimer la variance théorique σ^2, on procède comme dans le problème à deux échantillons. Rappelons que dans le problème à deux échantillons on combinait nos variances échantillonnales S_1^2 et S_2^2 de la façon suivante :

$$S_c^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Pour le problème à I échantillons, on procède de la même façon. On combine nos I variances échantillonnales de la façon suivante :

$$S_c^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots + (n_I-1)S_I^2}{(n_1+n_2+\cdots+n_I) - I}.$$

Dans cette expression, S_i^2 dénote la variance échantillonnale calculée à partir des observations $Y_{i,1}, Y_{i,2}, Y_{i,3}, ..., Y_{i,n_i}$, c'est-à-dire

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 \tag{5.2}$$

Lorsque nous avons étudié le problème à deux échantillons, nous avons vu que la distribution de la statistique $(n_1+n_2-2)S_c^2/\sigma^2$ était la loi du khi-deux avec n_1+n_2-2 degrés de liberté. Dans le présent contexte, le résultat analogue est le suivant :

$$\frac{(n_1+n_2+\cdots+n_I-I)S_c^2}{\sigma^2} \sim \chi^2_{n_1+n_2+\cdots+n_I-I}.$$

Notation et terminologie :

- Pour simplifier l'écriture, la somme des tailles d'échantillons sera dénotée N. On écrit donc $N = n_1+n_2+\cdots+n_I$. La dernière équation peut donc s'écrire sous la forme
$$\frac{(N-I)S_c^2}{\sigma^2} \sim \chi^2_{N-I}. \tag{5.3}$$

- Dans le présent contexte, la variance échantillonnale combinée S_c^2 est habituellement dénotée MSE. Cette notation vient de l'anglais *Mean Squared Errors*. Nous avons donc

$$MSE = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_I - 1)S_I^2}{N - I}. \tag{5.4}$$

À l'aide de l'équation (5.2), on peut réécrire l'équation (5.4) sous la forme suivante :

$$MSE = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2}{N - I}. \tag{5.5}$$

Le numérateur du terme de droite dans l'équation (5.5) représente la variation à l'intérieur des échantillons c'est-à-dire la somme des carrés des distances entre *observation* et *moyenne des observations* à l'intérieur de l'échantillon. Ce numérateur est souvent dénoté SSE, de l'anglais *Sum of Squared Errors*. On a donc

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2 \quad \text{et} \quad MSE = \frac{SSE}{N - I}.$$

L'équation (5.3) peut donc aussi s'écrire sous la forme

$$\frac{(N - I)MSE}{\sigma^2} \sim \chi^2_{N-I}. \tag{5.6}$$

On peut utiliser ce résultat pour calculer un intervalle de confiance pour le paramètre σ^2. On obtient l'intervalle

$$\left(\frac{(N - I)MSE}{\chi^2_{N-I,\frac{\alpha}{2}}}, \frac{(N - I)MSE}{\chi^2_{N-I,1-\frac{\alpha}{2}}} \right).$$

L'étudiant devrait comparer cet intervalle avec l'intervalle pour σ^2 dans le problème à un échantillon provenant d'une loi normale avec variance σ^2 et avec l'intervalle pour σ^2 dans le problème à deux échantillons indépendants provenant de lois normales de même variance σ^2. On peut aussi se servir de l'équation (5.6) pour faire des tests d'hypothèses sur σ^2. Par exemple, pour tester $H_0 : \sigma^2 = \sigma_o^2$ contre $H_1 : \sigma^2 > \sigma_o^2$ au seuil α, on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{(N - I)MSE}{\sigma_o^2} \geq \chi^2_{N-I,\alpha}.$$

Enfin, à partir des résultats (5.1) et (5.6) on peut obtenir des intervalles de confiance pour les moyennes individuelles μ_i et les différences de moyennes $\mu_i - \mu_k$. Pour une moyenne μ_i, on obtient l'intervalle

$$\overline{y}_{i\cdot} \pm t_{N-I,\alpha/2} \sqrt{\frac{MSE}{n_i}}. \tag{5.7}$$

L'étudiant devrait comparer cet intervalle avec l'intervalle pour la moyenne μ dans le problème à un échantillon issu d'une loi normale. Pour une différence de moyennes $\mu_i - \mu_k$, on obtient l'intervalle

$$(\overline{y}_{i\cdot} - \overline{y}_{k\cdot}) \pm t_{N-I,\alpha/2} \sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}. \tag{5.8}$$

L'étudiant devrait comparer cet intervalle avec l'intervalle pour $\mu_1 - \mu_2$ dans le problème à deux échantillons indépendants issus de lois normales de même variance.

5.3 Le test d'égalité des moyennes théoriques

On suppose toujours le scénario décrit à la section 5.1. On veut tester

$$H_0 : \text{ les moyennes } \mu_i \text{ sont toutes égales}$$
$$H_1 : \text{ les moyennes } \mu_i \text{ ne sont pas toutes égales}$$

c'est-à-dire

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$
$$H_1 : \mu_i \neq \mu_j \text{ pour au moins un choix de } i \text{ et } j.$$

L'idée derrière la règle de décision est très simple :

- S'il n'y a pas beaucoup de variation entre nos I moyennes échantillonnales, alors il n'y a pas lieu de rejeter H_0.
- S'il y a beaucoup de variation entre nos I moyennes échantillonnales, alors on rejette l'hypothèse nulle H_0.

Pour mettre en pratique cette idée, il faut *mesurer* la variation entre nos I moyennes échantillonnales $\overline{Y}_{1\cdot}, \overline{Y}_{2\cdot}, ..., \overline{Y}_{I\cdot}$. On mesure cette variation avec la statistique

$$SST_R = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

où $\overline{Y}_{\cdot\cdot}$ dénote la moyenne échantillonnale globale, c'est-à-dire

$$\overline{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N} \sum_{i=1}^{I} n_i \overline{Y}_{i\cdot}$$

Dans les applications, il arrive souvent que les I échantillons correspondent aux réponses de divers sujets, ou unités expérimentales, à I traitements. La notation SST_R vient de l'anglais *Sum of Squares between Treatments*. Le nombre de degrés de liberté associés à SST_R est $I - 1$ et le rapport

$$MST_R = \frac{SST_R}{I - 1}$$

est appelé le *Mean Square Treatment* ou *Mean Square between Treatments*. Voici donc une façon simple de préciser la règle de décision :

- Si MST_R n'est pas très grand par rapport à MSE, alors il n'y a pas lieu de rejeter l'hypothèse nulle H_0.
- Si MST_R est très grand par rapport à MSE, alors on rejette H_0.

Autrement dit, notre règle de décision sera de la forme

$$\text{on rejette } H_0 \text{ si } \frac{MST_R}{MSE} \geq c$$

où c est une constante convenablement choisie. Le résultat suivant nous permet de déterminer la constante c.

THÉORÈME. Si les conditions énoncées à la section 5.1 sont satisfaites et si l'hypothèse d'égalité des moyennes théoriques est satisfaite, alors

$$\frac{MST_R}{MSE} \sim F_{I-1,N-I}. \tag{5.9}$$

Ce théorème nous permet d'écrire notre règle de décision au seuil α sous la forme

$$\text{on rejette } H_0 \text{ si } \frac{MST_R}{MSE} \geq F_{I-1,N-I,\alpha} \tag{5.10}$$

et il explique pourquoi on écrit parfois F pour dénoter le rapport MST_R/MSE. Le théorème nous dit aussi comment calculer notre *p-value*. Si F_{obs} dénote la valeur observée de la statistique $F = MST_R/MSE$, alors notre *p-value* est donné par

$$p\text{-}value = \mathbb{P}_{H_0}[F \geq F_{obs}].$$

REMARQUES : Le test décrit dans la présente section est appelé le test de Fisher. On peut montrer que dans le cas $I = 2$, le test de Fisher est équivalent au test de Student du problème à deux échantillons indépendants.

5.4 Une interprétation du rapport MST_R/MSE

Voici une autre façon d'interpréter la règle de décision énoncée à la section précédente. Examinons séparément le numérateur et le dénominateur de notre statistique de test MST_R/MSE.

LE DÉNOMINATEUR MSE :

Rappelons que l'espérance de la loi du khi-deux est simplement son nombre de degrés de liberté. Ainsi, à partir du résultat $(N-I)MSE/\sigma^2 \sim \chi^2_{N-I}$ on obtient

$$\mathbb{E}\left[\frac{(N-I)MSE}{\sigma^2}\right] = N - I$$

et donc

$$\mathbb{E}[MSE] = \sigma^2. \tag{5.11}$$

Ce résultat est valide, peu importe que les moyennes théoriques $\mu_1, \mu_2, ..., \mu_I$ soient toutes égales ou non. La statistique MSE est donc un estimateur sans biais pour σ^2, peu importe que H_0 soit vraie ou non.

LE NUMÉRATEUR MST_R :

La statistique MST_R est un peu plus difficile à analyser. Si on pose

$$\mu = \frac{1}{N}\sum_{i=1}^{I} n_i \mu_i \tag{5.12}$$

alors on peut montrer que

$$\mathbb{E}[MST_R] = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i(\mu_i - \mu)^2. \tag{5.13}$$

Ce résultat est valide, peu importe que les moyennes théoriques $\mu_1, \mu_2, ..., \mu_I$ soient toutes égales ou non. Notez que la quantité μ définie par l'équation (5.12) est une moyenne pondérée des moyennes théoriques $\mu_1, \mu_2, ..., \mu_I$. Par conséquent, si ces moyennes théoriques sont toutes égales, alors on obtient $\mu = \mu_1 = \mu_2 = \cdots = \mu_I$ et donc

$$\frac{1}{I-1}\sum_{i=1}^{I} n_i(\mu_i - \mu)^2 = 0.$$

Par contre, si les moyennes théoriques $\mu_1, \mu_2, ..., \mu_I$ ne sont pas toutes égales, alors on obtient

$$\frac{1}{I-1}\sum_{i=1}^{I} n_i(\mu_i - \mu)^2 > 0.$$

L'équation (5.13) nous donne donc

$$\begin{array}{rcll} \mathbb{E}[MST_R] & = & \sigma^2 & \text{si } H_0 \text{ est vraie} \\ \mathbb{E}[MST_R] & > & \sigma^2 & \text{si } H_0 \text{ n'est pas vraie.} \end{array}$$

Bref, si H_0 est vraie, MST_R est un estimateur sans biais pour σ^2 alors que si H_0 n'est pas vraie MST_R a tendance à surestimer σ^2.

Le rapport MST_R/MSE :

À la lumière des paragraphes précédents, on obtient les conclusions suivantes.

(a) Si H_0 est vraie, alors MSE et MST_R sont tous les deux des estimateurs sans biais pour σ^2. Le rapport MST_R/MSE a donc tendance à être aux alentours de 1.

(b) Si H_0 n'est pas vraie, alors MSE est un estimateur sans biais pour σ^2 alors que MST_R a tendance à surestimer σ^2. Le rapport MST_R/MSE a donc tendance à être plus grand que 1.

Ces deux observations justifient la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{MST_R}{MSE} \text{ est trop grand.}$$

Analyse des moyennes ou analyse de la variance ?

Nous avons développé une procédure pour tester les hypothèses

$$\begin{array}{ll} H_0 : & \text{les moyennes } \mu_i \text{ sont toutes égales} \\ H_1 : & \text{les moyennes } \mu_i \text{ ne sont pas toutes égales.} \end{array}$$

Ces hypothèses portent sur des moyennes mais la technique utilisée s'appelle l'*analyse de la variance*. Pourquoi ? Parce que cette technique est basée sur une analyse de deux

estimateurs de la variance théorique σ^2, à savoir l'estimateur MSE et l'estimateur MST_R. On utilise l'acronyme anova, parfois écrit ANOVA, pour *analyse de la variance*. Le « O » de l'acronyme nous vient de l'anglais *analysis of variance*. Le problème étudié dans les sections 1 à 12 du présent chapitre est appelé le modèle d'anova à un facteur. À la section 14 nous allons considérer brièvement le modèle d'anova à deux facteurs.

5.5 Un exemple illustratif

Un chercheur a mené une expérience dans le but de comparer trois types d'engrais pour plants de tomates. Le tableau suivant résume les données :

Type d'engrais	Taille de l'échantillon	Moyenne échantillonnale	Écart-type échantillonnal
A	21	6.50 lbs	1.25 lb
B	21	4.75 lbs	1.05 lb
C	21	5.10 lbs	1.40 lb

Ce résumé a été obtenu de la façon suivante. Le chercheur a fait pousser 21 plants de tomates avec l'engrais A. Pour chacun de ces 21 plants, il a mesuré le poids, en livres, de la production de tomates pour la saison entière. La moyenne de ces 21 poids est de 6.50 livres et l'écart-type est de 1.25 livre. Même chose pour les engrais B et C. On désire tester l'hypothèse d'égalité des moyennes théoriques. On nous demande de faire l'analyse appropriée.

Solution. Calculons d'abord la moyenne échantillonnale globale :

$$\begin{aligned} \overline{y}_{..} &= \frac{1}{N}\sum_{i=1}^{I} n_i \overline{y}_{i.} \\ &= \frac{(21 \times 6.50) + (21 \times 4.75) + (21 \times 5.10)}{63} = 5.45. \end{aligned}$$

Calculons ensuite SST_R et SSE :

$$\begin{aligned} SST_R &= \sum_{i=1}^{I} n_i(\overline{y}_{i.} - \overline{y}_{..})^2 \\ &= 21(6.50 - 5.45)^2 + 21(4.75 - 5.45)^2 + 21(5.10 - 5.45)^2 = 36.015, \\ SSE &= \sum_{i=1}^{I}(n_i - 1)s_i^2 \\ &= 20(1.25)^2 + 20(1.05)^2 + 20(1.40)^2 = 92.500. \end{aligned}$$

Enfin, calculons MST_R et MSE :

$$\begin{aligned} MST_R &= \frac{SST_R}{I-1} = \frac{36.015}{2} = 18.0075 \\ MSE &= \frac{SSE}{N-I} = \frac{92.500}{60} = 1.5417. \end{aligned}$$

La valeur observée de notre statistique de test est donc

$$F_{obs} = \frac{MST_R}{MSE} = \frac{18.0075}{1.5417} = 11.68.$$

Cette valeur est beaucoup plus grande que ce à quoi on s'attendrait si les trois moyennes théoriques étaient égales. Notre *p-value* est

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[F \geq F_{obs}] \\ &= \mathbb{P}_{H_0}[F \geq 11.68] \\ &= \text{ surface à droite de 11.68 sous la densité } F_{2,60} \\ &= 0.00005 \qquad \text{(obtenu avec le logiciel R).} \end{aligned}$$

Le *p-value* étant extrêmement petit, on rejette l'hypothèse nulle et on conclut qu'il y a une différence entre les effets des différents engrais sur le rendement des plants de tomates.

REMARQUE 1. Dans cet exemple, on a laissé tomber l'indice « *obs* » pour les valeurs observées de SST_R, MST_R, SSE et MSE. C'est simplement pour alléger la notation.

REMARQUE 2. Dans cet exemple, nos trois populations avec distributions $N(\mu_A, \sigma^2)$, $N(\mu_B, \sigma^2)$ et $N(\mu_C, \sigma^2)$ sont des populations hypothétiques ! On fait comme si nos mesures sur les 21 plants de tomates soumis à l'engrais A constituent un échantillon aléatoire de taille 21 issu d'une population imaginaire de plants de tomates soumis à l'engrais A et on suppose que la loi $N(\mu_A, \sigma^2)$ est un bon modèle pour cette population. Même chose pour les engrais B et C.

REMARQUE 3. Dans le langage des statisticiens, les 63 jeunes plants de tomate choisis pour réaliser cette expérience sont appelés les *unités expérimentales*. Idéalement, on *randomise* nos unités expérimentales avant d'appliquer le traitement : on choisit au hasard les 21 plants qui recevront l'engrais A, les 21 plants qui recevront l'engrais B et les 21 plants qui recevront l'engrais C.

5.6 Décomposition de la somme des carrés et table d'anova

Aux sections 5.2 et 5.3, nous avons introduit les sommes de carrés suivantes :

$$\begin{aligned} SST_R &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \\ SSE &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2 = \sum_{i=1}^{I} (n_1 - 1) S_i^2. \end{aligned}$$

Une troisième somme importante est la somme totale des carrés, dénotée SST de l'anglais *Sum of Squares Total* :

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\cdot\cdot})^2.$$

Le résultat suivant précise le lien entre ces trois sommes de carrés.

Théorème.

$$SST = SST_R + SSE.$$

Démonstration. En utilisant le fait que pour chaque i on a

$$\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot}) = \sum_{i=1}^{n_i} Y_{ij} - n_i\overline{Y}_{i\cdot} = n_i\overline{Y}_{i\cdot} - n_i\overline{Y}_{i\cdot} = 0$$

et donc

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) = \sum_{i=1}^{I}\left\{(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})\right\} = 0.$$

on obtient

$$\begin{aligned}
SST &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot} + \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}((Y_{ij} - \overline{Y}_{i\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}))^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left\{(Y_{ij} - \overline{Y}_{i\cdot})^2 + 2(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2\right\} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \\
&= SSE + SST_R.
\end{aligned}$$

La table d'anova :

À l'époque de Fisher, les ordinateurs n'existaient pas. L'anova nécessitait donc de longs calculs faits à la main. Fisher prit l'habitude de présenter de façon claire et concise le résumé des calculs d'une analyse de la variance sous forme d'un tableau qu'on appelle maintenant *la table d'anova*. Aujourd'hui, les logiciels de statistique résument les calculs d'analyse de la variance sous forme de tables d'anova à la Fisher. Dans le cas de l'anova étudiée ici, la table d'anova prend la forme suivante :

Source de Variation	Somme de carrés	Degrés de liberté	Moyenne des carrés	Statistique F	*p-value*
Traitement	SST_R	$I-1$	$MST_R = \frac{SST_R}{I-1}$	$F = \frac{MST_R}{MSE}$	$\mathbb{P}_{H_0}[F \geq F_o]$
Erreur	SSE	$N-I$	$MSE = \frac{SSE}{N-I}$		
Total	SST	$N-1$			

On note que pour remplir ce tableau, les seuls longs calculs sont le calcul de SST_R et le calcul de SSE. Une fois qu'on a calculé ces deux quantités, la table d'anova est facile à compléter.

Voici la table d'anova pour l'exemple de la section précédente :

Source de Variation	Somme de carrés	Degrés de liberté	Moyenne des carrés	Statistique F	*p-value*
Engrais	36.015	2	18.0075	11.6805	0.00005
Erreur	92.500	60	1.5417		
Total	128.515	62			

5.7 Un autre exemple illustratif

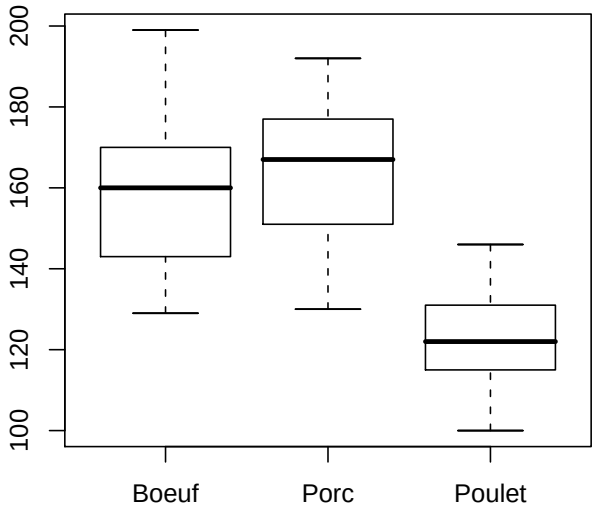
Dans une étude menée par *Consumer Reports* (Juin 1986, pages 366-367), des chercheurs ont mesuré le contenu en calories pour 51 marques de saucisses à *hot dog*. Parmi ces 51 marques, il y en avait 17 au boeuf, 17 au porc et 17 au poulet. Les données (légèrement modifiées) sont présentées ci-dessous.

Boeuf	176	129	169	144	170	133	162	163	131	155
	160	143	176	184	199	147	143			
Porc	143	177	172	192	183	155	173	172	167	161
	144	151	145	164	177	181	130			
Poulet	125	131	130	140	116	100	127	116	136	118
	113	122	115	138	107	106	146			

Le tableau suivant résume les données :

Type de viande	Taille de l'échantillon	Moyenne échantillonnale	Écart-type échantillonnal
Boeuf	17	157.88	19.82
Porc	17	163.94	16.97
Poulet	17	122.71	13.00

Voici les diagrammes en boîte juxtaposés :



Le résumé et les diagrammes en boîte suggèrent que d'une part il n'y a pas beaucoup de différence entre les saucisses au boeuf et les saucisses au porc et que d'autre part les saucisses au poulet contiennent vraiment moins de calories que celles au boeuf et celles au porc. Pour voir si cette différence est vraiment significative, obtenons notre table d'anova.

Voici quelques calculs préliminaires :

$$\begin{aligned}
\overline{y}_{..} &= \frac{1}{N}\sum_{i=1}^{I} n_i \overline{y}_{i.} = \frac{157.88 + 163.94 + 122.71}{3} = 148.18 \\
SST_R &= \sum_{i=1}^{I} n_i (\overline{y}_{i.} - \overline{y}_{..})^2 \\
&= 17\left((157.88 - 148.18)^2 + (163.94 - 148.18)^2 + (122.71 - 148.18)^2\right) = 16855.18 \\
SSE &= \sum_{i=1}^{I} (n_i - 1) s_i^2 = 16\ \left((19.82)^2 + (16.97)^2 + (13.00)^2\right) = 13598.24.
\end{aligned}$$

Voici la table d'anova :

Source de Variation	Somme de carrés	Degrés de liberté	Moyenne des carrés	Statistique F	*p-value*
Viande	16855.18	2	8427.59	29.75	< 0.0001
Erreur	13598.24	48	283.30		
Total	30453.42	50			

Le *p-value* est extrêmement petit. On rejette donc l'hypothèse d'égalité des moyennes théoriques.

5.8 Inférence pour une combinaison linéaire des moyennes

Lorsqu'on rejette l'hypothèse d'égalité des moyennes théoriques, il peut arriver qu'on veuille estimer une certaine combinaison linéaire de ces moyennes. Dans l'exemple des saucisses à hot dog, on aimerait peut-être estimer la quantité

$$\frac{\mu_1 + \mu_2}{2} - \mu_3.$$

Notre estimation est simplement

$$\frac{\overline{y}_{1\cdot} + \overline{y}_{2\cdot}}{2} - \overline{y}_{3\cdot} = \frac{157.88 + 163.94}{2} - 122.71 = 38.20 \text{ calories.}$$

Quelle est l'erreur type associée à cette estimation ?

Pour répondre à cette question, nous utilisons le résultat suivant, énoncé ici dans le contexte général des sections 5.1 et 5.2.

THÉORÈME. Pour tout choix des constantes $c_1, c_2, ..., c_I$, on a

$$\sum_{i=1}^{I} c_i \overline{Y}_{i\cdot} \sim N\left(\sum_{i=1}^{I} c_i \mu_i, \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{n_i}\right).$$

Voici les principales conséquences de ce théorème :

(a) La statistique $\sum_{i=1}^{I} c_i \overline{Y}_{i\cdot}$ est un estimateur sans biais pour la quantité $\sum_{i=1}^{I} c_i \mu_i$.

(b) Si on utilise l'estimateur $\sum_{i=1}^{I} c_i \overline{Y}_{i\cdot}$ pour estimer la combinaison linéaire $\sum_{i=1}^{I} c_i \mu_i$, alors l'erreur type associée à l'estimation $\sum_{i=1}^{I} c_i \overline{y}_{i\cdot}$ est donnée par

$$\text{erreur type} \;=\; \sqrt{\sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{n_i}} \;\approx\; \sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}.$$

(c) L'intervalle de confiance de niveau $1 - \alpha$ pour $\sum_{i=1}^{I} c_i \mu_i$ est donné par

$$\sum_{i=1}^{I} c_i \overline{y}_{i\cdot} \;\pm\; t_{N-I,\alpha/2} \sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}. \tag{5.14}$$

(d) Pour tester $H_0 : \sum_{i=1}^{I} c_i\mu_i = m_o$ vs $H_1 : \sum_{i=1}^{I} c_i\mu_i > m_o$ on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } T \geq t_{N-I,\alpha}$$

avec

$$T = \frac{\sum_{i=1}^{I} c_i\overline{Y}_{i\cdot} - m_o}{\sqrt{MSE\sum_{i=1}^{I}\frac{c_i^2}{n_i}}}.$$

REMARQUE. Si tous les c_i sont nuls, sauf un qui est égal à 1, alors l'intervalle donné par l'équation (5.14) se réduit à l'intervalle pour une moyenne donné à l'équation (5.7). Si tous les c_i sont nuls, sauf un qui est égal à 1 et un autre qui est égal à -1, alors l'intervalle donné par l'équation (5.14) se réduit à l'intervalle pour une différence de moyennes donné à l'équation (5.8).

Revenons à l'exemple des saucisses à hot dog. On a $c_1 = 1/2, c_2 = 1/2$ et $c_3 = -1$. L'erreur type associée à notre estimation est

$$\sqrt{MSE\sum_{i=1}^{I}\frac{c_i^2}{n_i}} = \sqrt{283.3\left(\frac{(1/2)^2}{17} + \frac{(1/2)^2}{17} + \frac{(-1)^2}{17}\right)} = 5.00 \text{ calories}$$

et l'intervalle de confiance de niveau 90% pour $\frac{\mu_1+\mu_2}{2} - \mu_3$ est l'intervalle $(29.8, 46.6)$.

REMARQUE : La combinaison linéaire $\frac{\mu_1+\mu_2}{2} - \mu_3$ est un exemple de combinaison linéaire appelée un *contraste*. Dans le cas général, un contraste des moyennes $\mu_1, \mu_2, ..., \mu_I$ est une combinaison linéaire $\sum_{i=1}^{I} c_i\mu_i$ pour laquelle les constantes c_1, c_2,..., c_I vérifient la condition $c_1 + c_2 + \cdots + c_I = 0$. Dans notre exemple on avait $I = 3$, $c_1 = 1/2, c_2 = 1/2$ et $c_3 = -1$. La condition $c_1 + c_2 + c_3 = 0$ était donc satisfaite. Dans certains problèmes, on s'intéresse à plusieurs combinaisons linéaires en même temps. Dans le cas où ces combinaisons linéaires sont toutes des contrastes, il existe des méthodes spéciales permettant d'obtenir des intervalles de confiance *simultanés*. Ce sujet ne sera pas abordé ici.

5.9 Description alternative du modèle d'anova [1]

Rappelons les conditions du modèle d'anova à un facteur présentées à la section 1 :

- $Y_{1,1}, Y_{1,2}, Y_{1,3}, ..., Y_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma^2)$.
- $Y_{2,1}, Y_{2,2}, Y_{2,3}, ..., Y_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma^2)$.

  $\vdots$
- $Y_{I,1}, Y_{I,2}, Y_{I,3}, ..., Y_{I,n_I}$ est un échantillon aléatoire de taille n_I issu d'une population avec distribution $N(\mu_I, \sigma^2)$.
- Ces I échantillons aléatoires sont indépendants les uns des autres.
- Les variances théoriques sont égales ; leur valeur commune est dénotée σ^2.

[1] On peut omettre cette section si on manque de temps.

- Les paramètres $\mu_1, \mu_2, ..., \mu_I$ et σ^2 sont inconnus.

Si on pose

$$\epsilon_{ij} = Y_{ij} - \mu_i$$

alors les variables aléatoires ϵ_{ij} sont des $N(0, \sigma^2)$ indépendantes les unes des autres et on peut écrire

$$Y_{ij} = \mu_i + \epsilon_{ij}.$$

Maintenant posons

$$\alpha_i = \mu_i - \mu \tag{5.15}$$

avec

$$\mu = \frac{1}{N} \sum_{i=1}^{I} n_i \mu_i$$

comme à l'équation (5.12). On a alors

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{5.16}$$

Notons en passant que

$$\sum_{i=1}^{I} n_i \alpha_i = \sum_{i=1}^{I} n_i (\mu_i - \mu) = \sum_{i=1}^{I} n_i \mu_i - \mu \sum_{i=1}^{I} n_i = N\mu - N\mu = 0.$$

L'équation (5.16) possède une belle interprétation. Prenons par exemple le scénario où un chercheur veut comparer I traitements. Il utilise $N = n_1 + n_2 + \cdots + n_I$ unités expérimentales : n_1 unités reçoivent le traitement 1, n_2 unités reçoivent le traitement 2, etc. L'équation (5.16) exprime la réponse au traitement i comme une somme de trois composantes : le paramètre μ représente l'effet moyen des I traitements, le paramètre α_i représente l'effet particulier du traitement i et la variable aléatoire ϵ_{ij} représente la variation résiduelle non expliquée par le traitement.

Avec cette façon de décrire le modèle d'anova à un facteur, les hypothèses

$$\begin{aligned} H_0 &: \text{ les moyennes } \mu_i \text{ sont toutes égales} \\ H_1 &: \text{ les moyennes } \mu_i \text{ ne sont pas toutes égales} \end{aligned}$$

peuvent s'écrire sous la forme suivante :

$$\begin{aligned} H_0 &: \text{ les paramètres } \alpha_i \text{ sont tous nuls} \\ H_1 &: \text{ les paramètres } \alpha_i \text{ ne sont pas tous nuls} \end{aligned}$$

On utilisait la moyenne échantillonnale $\overline{Y}_{i\cdot}$ pour estimer la moyenne théorique μ_i. De même, on utilise la moyenne échantillonnale globale $\overline{Y}_{\cdot\cdot}$ pour estimer la moyenne globale μ. Puisque $\alpha_i = \mu_i - \mu$, on utilise $\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ pour estimer α_i. L'estimateur $\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ est parfois dénoté $\hat{\alpha}_i$. On a donc

$$\hat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$$

Remarque. Cette façon alternative du présenter l'anova devient très utile lorsqu'on considère des modèles plus complexes, comme celui qui sera présenté à la section 14.

5.10 L'hypothèse d'homogénéité des variances

Dans notre modèle d'anova, on a supposé que les I populations à partir desquelles on échantillonne ont toutes la même variance. Dans la pratique, comment peut-on vérifier si cette hypothèse est raisonnable ? Autrement dit, comment peut-on tester les hypothèses

$$\begin{array}{rcl} H_0 & : & \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 \\ H_1 & : & \text{les variances } \sigma_1^2, \sigma_2^2, ..., \sigma_I^2 \text{ ne sont pas toutes égales.} \end{array}$$

Il existe plusieurs procédures :

- Le test de Bartlett (1937).
- Le test de Cochran (1941).
- Le test de Hartley (1950).
- Le test de Levene (1960).

Dans le présent chapitre, nous allons considérer uniquement le test de Bartlett. Au seuil α, la règle de décision du test de Bartlett s'énonce ainsi :

$$\text{on rejette } H_0 \text{ si } B \geq \chi^2_{I-1,\alpha},$$

avec

$$B = \frac{(N-I)\ln(MSE) - \sum_{i=1}^{I}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(I-1)}\left\{\left(\sum_{i=1}^{I}\frac{1}{n_i - 1}\right) - \frac{1}{N-I}\right\}}.$$

Cette règle de décision est basée sur les résultats suivants :

- Si les variances théoriques sont toutes égales, on s'attend à ce que la statistique B de Bartlett soit aux alentours de $I - 1$.
- Si les variances théoriques ne sont pas toutes égales, on s'attend à ce que la statistique B de Bartlett soit bien plus grande que $I - 1$.
- Si les variances théoriques sont toutes égales et si les n_i sont suffisamment grands, disons $n_i \geq 5$, alors la statistique B de Bartlett suit à peu près la loi du khi-deux avec $I - 1$ degrés de liberté.

Exemple. Reprenons l'exemple de la section 5.5. On obtient $B_{obs} = 1.60$. Sous $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, on a $B \approx \chi_2^2$. Le *p-value* est donc la surface à droite de 1.60 sous la densité de la loi du khi-deux à 2 degrés de liberté. Pas besoin de table ou de logiciel R ! La moyenne de la loi du khi-deux avec 2 degrés de liberté est 2 et l'écart-type est aussi 2. On a obtenu la valeur 1.6. C'est cohérent avec la khi-deux avec 2 degrés de liberté. Autrement dit, c'est cohérent avec l'hypothèse nulle d'égalité des variances. Il n'y a donc pas lieu de rejeter de l'hypothèse $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$. (Avec R on obtient *p-value* = 0.45).

Le test de Bartlett avec le logiciel R :

Reprenons l'exemple des saucisses à hot dog de la section 5.7. On peut calculer la statistique B de Bartlett à la main :

$$\begin{aligned} B_{obs} &= \frac{(N-I)\ln(MSE) - \sum_{i=1}^{I}(n_i-1)\ln(S_i^2)}{1+\frac{1}{3(I-1)}\left\{\left(\sum_{i=1}^{I}\frac{1}{n_i-1}\right) - \frac{1}{N-I}\right\}} \\ &= \frac{(51-3)\ln(283.30) - \{(17-1)\ln(392.36) + (17-1)\ln(287.93) + (17-1)\ln(169.10)\}}{1+\frac{1}{3(3-1)}\left\{\left(\frac{1}{17-1}+\frac{1}{17-1}+\frac{1}{17-1}\right) - \frac{1}{51-3}\right\}} \\ &= \frac{48\ln(283.30) - 16\{\ln(392.36)+\ln(287.93)+\ln(169.10)\}}{1+\frac{1}{6}\left\{\frac{3}{16}-\frac{1}{48}\right\}} = 2.71. \end{aligned}$$

Le *p-value* est donc la surface à droite de 2.71 sous la densité χ^2_2, c'est-à-dire 0.258. Ce *p-value* est grand. Il n'y a donc pas lieu de rejeter l'hypothèse d'égalité des variances théoriques σ^2_{boeuf}, σ^2_{porc} et σ^2_{poulet}.

On peut aussi calculer cette statistique à l'aide du logiciel R. Voici une façon de faire ça.

```
boeuf <- c(176, 129, 169, 144, 170, 133, 162, 163, 131, 155, 160,
143, 176, 184, 199, 147, 143)
porc <- c(143, 177, 172, 192, 183, 155, 173, 172, 167, 161, 144,
151, 145, 164, 177, 181, 130)
poulet <- c(125, 131, 130, 140, 116, 100, 127, 116, 136, 118, 113,
122, 115, 138, 107, 106, 146)
calorie <- c(boeuf, porc, poulet)
viande <- c(rep("Boeuf",17), rep("Porc",17), rep("Poulet",17))
```

Le vecteur `calorie` contient alors nos 51 observations correspondant, dans l'ordre, aux calories des 17 saucisses au boeuf, des 17 saucisses au porc et des 17 saucisses au poulet. Le vecteur `viande` contient, dans l'ordre, 17 fois le mot `Boeuf`, 17 fois le mot `Porc` et 17 fois le mot `Poulet`. On tape ensuite la commande

```
bartlett.test(calorie, viande)
```

et le logiciel R nous donne les résultats suivants :

```
Bartlett test of homogeneity of variances
data : calorie and viande
Bartlett's K-squared = 2.6906, df = 2, p-value = 0.2605
```

Le `K-squared` du logiciel R est notre statistique B de Bartlett. La différence entre le $B_{obs} = 2.71$ obtenu à la main et le $B_{obs} = K^2_{obs} = 2.6906$ obtenu par le logiciel R est simplement due aux erreurs d'arrondissement dans le calcul à la main.

Une justification pour la statistique de Bartlett :

D'où vient la statistique de Bartlett ? Qu'est-ce qu'elle représente ? Pourquoi une grande valeur de B nous incite-t-elle à rejeter l'hypothèse d'égalité des variances théoriques ? Voici une réponse à ces questions dans le cas où les tailles d'échantillon sont toutes égales. Dans ce cas, si on écrit n pour dénoter la taille commune de nos I échantillons, alors on note que la statistique B de Bartlett peut s'écrire sous la forme suivante :

$$\begin{aligned}
B &= \frac{(N-I)\ln(MSE) - \sum_{i=1}^{I}(n_i-1)\ln(S_i^2)}{1+\frac{1}{3(I-1)}\left\{\left(\sum_{i=1}^{I}\frac{1}{n_i-1}\right) - \frac{1}{N-I}\right\}} \\
&= \frac{(nI-I)\ln(MSE) - \sum_{i=1}^{I}(n-1)\ln(S_i^2)}{1+\frac{1}{3(I-1)}\left\{\left(\sum_{i=1}^{I}\frac{1}{n-1}\right) - \frac{1}{nI-I}\right\}} \\
&= \frac{(n-1)I\ln(MSE) - (n-1)\sum_{i=1}^{I}\ln(S_i^2)}{1+\frac{1}{3(I-1)}\left\{\frac{I}{n-1} - \frac{1}{(n-1)I}\right\}} \\
&= \frac{(n-1)I}{1+\frac{1}{3(I-1)}\left\{\frac{I}{n-1} - \frac{1}{(n-1)I}\right\}}\left(\ln(MSE) - \frac{1}{I}\sum_{i=1}^{I}\ln(S_i^2)\right) \\
&= c\left(\ln(MSE) - \frac{1}{I}\sum_{i=1}^{I}\ln(S_i^2)\right)
\end{aligned}$$

où c est la constante positive définie par

$$c = \frac{(n-1)I}{1+\frac{1}{3(I-1)}\left\{\frac{I}{n-1} - \frac{1}{(n-1)I}\right\}}.$$

Rappelons les principales propriétés de la fonction logarithme : $\ln(x)+\ln(y) = \ln(xy)$, $\ln(x)-\ln(y) = \ln(x/y)$ et $a\ln(x) = \ln(x^a)$. À l'aide de ces propriétés on obtient

$$\begin{aligned}
B &= c\left(\ln(MSE) - \frac{1}{I}\sum_{i=1}^{I}\ln(S_i^2)\right) = c\left(\ln(MSE) - \frac{1}{I}\ln\left(\prod_{i=1}^{I}S_i^2\right)\right) \\
&= c\left(\ln(MSE) - \ln\left(\left(\prod_{i=1}^{I}S_i^2\right)^{1/I}\right)\right) = c\ln\left(\frac{MSE}{\left(\prod_{i=1}^{I}S_i^2\right)^{1/I}}\right) \\
&= c\ln\left(\frac{\frac{S_1^2+S_2^2+\cdots+S_I^2}{I}}{\left(S_1^2\,S_2^2\,\cdots\,S_I^2\right)^{1/I}}\right).
\end{aligned}$$

Pour la dernière égalité, on a utilisé le fait que lorsque les échantillons sont tous de même taille, la statistique MSE est simplement la moyenne arithmétique des I variances échantillonnales. Ce calcul montre que la règle de décision

on rejette H_0 si B est trop grand

peut aussi s'écrire sous la forme

$$\text{on rejette } H_0 \text{ si } B^* \text{ est trop grand}$$

où B^* est la statistique définie par

$$B^* = \frac{\frac{S_1^2+S_2^2+\cdots+S_I^2}{I}}{\left(S_1^2\, S_2^2\, \cdots\, S_I^2\right)^{1/I}}.$$

Ce B^* peut être vu comme une estimation du rapport

$$\frac{\frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_I^2}{I}}{\left(\sigma_1^2\, \sigma_2^2\, \cdots\, \sigma_I^2\right)^{1/I}}.$$

Le numérateur $(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_I^2)/I$ est ce qu'on appelle la *moyenne arithmétique* des nombres $\sigma_1^2, \sigma_2^2, ..., \sigma_I^2$ et le dénominateur $(\sigma_1^2\, \sigma_2^2\, \cdots\, \sigma_I^2)^{1/I}$ est ce qu'on appelle la *moyenne géométrique* des nombres $\sigma_1^2, \sigma_2^2, ..., \sigma_I^2$. Le petit résultat suivant va nous permettre d'interpréter la statistique B^*.

THÉORÈME. *Si I est un entier positif et si $v_1, v_2, ..., v_I$ sont des nombres réels positifs alors on a*

$$\frac{v_1 + v_2 + \cdots + v_I}{I} \geq (v_1 v_2 \cdots v_I)^{1/I}. \tag{5.17}$$

On a égalité si et seulement si les nombres v_1, v_2,..., v_I sont tous égaux.

Si on applique ce théorème à nos I variances théoriques, on obtient

$$\frac{\frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_I^2}{I}}{\left(\sigma_1^2\, \sigma_2^2\, \cdots\, \sigma_I^2\right)^{1/I}} \geq 1$$

avec égalité si et seulement si on a $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$. Autrement dit, si l'hypothèse d'égalité des variances théoriques est vraie, alors on a

$$\frac{\frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_I^2}{I}}{\left(\sigma_1^2\, \sigma_2^2\, \cdots\, \sigma_I^2\right)^{1/I}} = 1$$

et si cette hypothèse n'est pas vraie, alors on a

$$\frac{\frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_I^2}{I}}{\left(\sigma_1^2\, \sigma_2^2\, \cdots\, \sigma_I^2\right)^{1/I}} > 1.$$

Donc, si les variances théoriques sont toutes égales, on s'attend à ce que B^* soit aux alentours de 1 et si ces variances théoriques ne sont pas toutes égales alors on s'attend à ce que B^* soit plus grand que 1. Il est donc raisonnable d'utiliser la règle de décision

$$\text{on rejette } H_0 \text{ si } B^* \text{ est trop grand.}$$

L'argument qu'on vient de présenter est valide seulement dans le cas où les tailles d'échantillons sont toutes égales. Le cas où les tailles d'échantillon ne sont pas toutes égales peut être traité essentiellement de la même façon mais il faut alors invoquer une généralisation du résultat (5.17) pour les moyennes arithmétiques pondérées et les moyennes géométriques pondérées. Ce sera pour une autre fois...

REMARQUE : On peut montrer que dans le cas $I = 2$ le test de Bartlett est équivalent au test de Fisher que nous avons étudié dans le contexte des problèmes à deux échantillons.

5.11 L'hypothèse de normalité

Dans un problème d'anova à un facteur, nous supposons que nos I échantillons proviennent de distributions normales (ayant toutes la même variance). Comment vérifie-t-on l'hypothèse de normalité ? Si les tailles d'échantillons n_1, n_2,..., n_I sont suffisamment grandes, on peut appliquer les méthodes usuelles (histogramme, graphe quantile-quantile gaussien, test de Shapiro et Wilk, etc.) à chacun de nos I échantillons. Si nos n_i ne sont pas très grands, alors on peut procéder de la façon suivante. D'abord on note que sous les conditions énoncées à la section 5.1, on a

$$\epsilon_{ij} = Y_{ij} - \mu_i \sim N(0, \sigma^2).$$

Les variables

$$\epsilon_{ij} = Y_{ij} - \mu_i \quad j = 1, 2, 3, ..., n_i \qquad i = 1, 2, ..., I$$

sont donc indépendantes et identiquement distribuées $N(0, \sigma^2)$. On s'attend donc à ce que les *résidus*

$$e_{ij} = y_{ij} - \overline{y}_{i\cdot} \quad j = 1, 2, 3, ..., n_i \qquad i = 1, 2, ..., I$$

aient l'air d'un échantillon provenant d'une loi normale centrée à 0. On peut donc appliquer les méthodes usuelles (histogramme, graphe quantile-quantile gaussien, test de Shapiro et Wilk, etc.) à l'ensemble des résidus.

EXERCICE : Reprenez l'exemple de la section 5.7 (les saucisses à hot dog) et vérifiez si l'hypothèse de normalité est raisonnable.

5.12 Transformation des données

On considère ici un scénario d'anova à un facteur avec l'hypothèse d'homogénéité des variances et l'hypothèse de normalité non satisfaites. On suppose simplement qu'on a

- $Y_{11}, ..., Y_{1n_1}$, un échantillon issu d'une loi de moyenne μ_1 et de variance σ_1^2.
- $Y_{21}, ..., Y_{2n_2}$, un échantillon issu d'une loi de moyenne μ_2 et de variance σ_2^2.

  $\vdots$
- $Y_{I1}, ..., Y_{In_I}$, un échantillon issu d'une loi de moyenne μ_I et de variance σ_I^2.
- Ces I échantillons sont indépendants les uns des autres.

Dans un tel problème d'anova à un facteur, le scénario suivant est relativement fréquent :

- La variable d'intérêt est une variable positive.
- Les diagrammes en boîtes suggèrent des distributions asymétriques étirées vers la droite.
- Le graphe des I points $(\overline{y}_{i\cdot}, s_i)$ suggère l'existence d'une relation du type $\sigma_i = c\mu_i^p$ pour un certain $p > 0$.

On reconnaît facilement ce scénario en examinant les diagrammes en boîte juxtaposés. Dans un tel scénario, il est souvent possible de transformer les données de façon à ce que l'hypothèse d'homogénéité des variances et l'hypothèse de normalité soient satisfaites. Autrement dit, avec un choix approprié de la transformation $g(y)$, les données $y'_{ij} = g(y_{ij})$ vont satisfaire l'hypothèse d'homogénéité des variances et l'hypothèse de normalité. Le choix de la transformation $g(y)$ dépend de la puissance p qui apparaît dans la relation $\sigma_i = c\mu_i^p$. Voici les quatre cas les plus rencontrés en pratique :

Relation $\sigma_i = c\mu_i^p$	Transformation
$\sigma_i = c\sqrt{\mu_i}$	$y' = \sqrt{y}$
$\sigma_i = c\mu_i$	$y' = \ln(y)$
$\sigma_i = c\mu_i^{3/2}$	$y' = 1/\sqrt{y}$
$\sigma_i = c\mu_i^2$	$y' = 1/y$

On peut donc procéder de la façon suivante.

1. Est-ce que la variable d'intérêt est bel et bien une variable à valeurs positives ?
2. Est-ce que les diagrammes en boîte juxtaposés suggèrent des distributions asymétriques étirées vers la droite ?
3. Est-ce que les diagrammes en boîte juxtaposés suggèrent que plus la moyenne est grande, plus l'écart-type est grand ?
4. Si on a répondu « oui » aux trois questions ci-dessus, alors on trace le graphe des I points $(\overline{y}_{i\cdot}, s_i)$ et on choisit une transformation selon le tableau ci-dessus.
5. Une fois qu'on a choisi notre transformation $y' = g(y)$, on calcule nos *observations transformées* $y'_{ij} = g(y_{ij})$ et on essaie de voir si les conditions de l'anova à un facteur sont satisfaites par ces y'_{ij}.

Alternativement, on essaie chacune des quatre transformations suggérées ci-dessus et avec un peu de chance il y en aura une qui fera l'affaire !

5.13 Le test de Kruskal et Wallis [2]

Dans certains cas où l'hypothèse de normalité n'est pas satisfaite, on peut utiliser une procédure basée sur les rangs des observations. Cette procédure, appelée le test de Kruskal et Wallis, est une généralisation du test de la somme des rangs de Wilcoxon-Mann-Whitney étudié au chapitre 4.

[2] On peut omettre cette section si on manque de temps.

Le scénario :

On considère I populations, disons la population 1, la population 2,... et la population I. On s'intéresse à une certaine variable quantitative de type continu. On veut comparer les moyennes théoriques μ_1, μ_2,..., μ_I. On dispose d'échantillons aléatoires indépendants et on suppose que les I distributions théoriques ont la même forme mais possiblement des moyennes différentes. On a donc

- $Y_{11}, ..., Y_{1n_1}$, un échantillon aléatoire de taille n_1 issu de la population 1,
- $Y_{21}, ..., Y_{2n_2}$, un échantillon aléatoire de taille n_2 issu de la population 2,

  $\vdots$
- $Y_{I1}, ..., Y_{In_I}$, un échantillon aléatoire de taille n_I issu de la population I.
- Ces I échantillons sont indépendants les uns des autres.
- Les distributions théoriques ont la même forme mais avec, possiblement, des moyennes différentes. Autrement dit, les I densités de probabilité sont identiques à des translations près.

On veut tester $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ contre l'hypothèse alternative qui dit que ces I moyennes ne sont pas toutes égales.

La règle de décision :

On procède de la façon suivante :

1. On place nos $N = n_1 + n_2 + \cdots + n_I$ observations en ordre croissant, de la plus petite à la plus grande.
2. On attribue à nos N observations les *rangs* 1 à N de la façon suivante : la plus petite observation reçoit le rang 1, la deuxième plus petite observation reçoit le rang 2, la troisième plus petite observation reçoit le rang 3, etc.
3. On pose

$$\begin{aligned} R_1 &= \text{la somme des rangs des observations issues de la population 1,} \\ R_2 &= \text{la somme des rangs des observations issues de la population 2,} \\ &\vdots \\ R_I &= \text{la somme des rangs des observations issues de la population I.} \end{aligned}$$

4. On calcule la statistique de Kruskal-Wallis :

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{I} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2} \right)^2. \tag{5.18}$$

5. On rejette H_0 si K est trop grand.

On peut montrer que si H_0 est vraie et si les n_i sont tous plus grands ou égaux à 5, alors

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{I} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2} \right)^2 \approx \chi^2_{I-1}. \tag{5.19}$$

La règle de décision est donc la suivante :

$$\text{on rejette } H_0 \text{ si } K \geq \chi^2_{I-1,\alpha}.$$

Une justification pour la statistique de Kruskal et Wallis :

Supposons que H_0 soit vraie. Dans ce cas là, nos I populations ont toutes la même distribution et nos N observations peuvent être vues comme étant un échantillon aléatoire issu de cette distribution. Fixons $i \in \{1, 2, 3, ..., I\}$ et examinons les rangs des n_i observations de l'échantillon numéro i. Sous H_0, ces i rangs peuvent être vus comme étant le résultat de n_i tirages sans remise à partir d'un panier contenant N boules numérotées 1 à N. L'espérance de R_i est donc n_i fois la moyenne de ce panier, c'est-à-dire

$$\mathbb{E}_{H_0}[R_i] = n_i \times \frac{N+1}{2}.$$

On a donc

$$\mathbb{E}_{H_0}\left[\frac{R_i}{n_i}\right] = \frac{N+1}{2}.$$

Ceci est vrai pour chaque $i \in \{1, 2, 3, ..., I\}$. Bref, si H_0 est vraie, on s'attend à ce que les moyennes de rangs

$$\frac{R_1}{n_1}, \quad \frac{R_2}{n_2}, \quad \frac{R_3}{n_3}, \quad \cdots, \quad \frac{R_I}{n_I}$$

soient toutes assez proches de $(N+1)/2$. Autrement dit, si H_0 est vraie, on s'attend à ce que les écarts carrés

$$\left(\frac{R_1}{n_1} - \frac{N+1}{2}\right)^2, \quad \left(\frac{R_2}{n_2} - \frac{N+1}{2}\right)^2, \quad \left(\frac{R_2}{n_2} - \frac{N+1}{2}\right)^2, \quad \cdots, \quad \left(\frac{R_I}{n_I} - \frac{N+1}{2}\right)^2$$

soient tous petits. Donc, si H_0 est vraie, on s'attend à ce que la somme

$$\sum_{i=1}^{I} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2}\right)^2$$

soit petite tandis que si H_0 n'est pas vraie alors on s'attend à ce que cette somme soit grande. Notons que dans cette somme on a tenu compte des tailles d'échantillons. On conclut qu'une bonne règle de décision est la suivante :

$$\text{on rejette } H_0 \text{ si } \sum_{i=1}^{I} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2}\right)^2 \text{ est trop grand.}$$

Le facteur $12/(N(N+1))$ qui apparaît devant la somme dans l'équation (5.18) est là simplement pour que le résultat (5.19) soit valide.

Une forme alternative por la statistique de Kruskal-Wallis

La statistique de Kruskal-Wallis est souvent donnée sous la forme suivante :

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} \; - \; 3(N+1). \tag{5.20}$$

Il est facile de voir que les formes (5.18) et (5.20) sont équivalentes :

$$\begin{aligned}
&\frac{12}{N(N+1)}\sum_{i=1}^{I} n_i\left(\frac{R_i}{n_i}-\frac{N+1}{2}\right)^2 \\
&= \frac{12}{N(N+1)}\sum_{i=1}^{I} n_i\left(\frac{R_i^2}{n_i^2}-2\frac{R_i}{n_i}\frac{N+1}{2}+\frac{(N+1)^2}{4}\right) \\
&= \frac{12}{N(N+1)}\sum_{i=1}^{I}\left(\frac{R_i^2}{n_i}-R_i(N+1)+n_i\frac{(N+1)^2}{4}\right) \\
&= \frac{12}{N(N+1)}\left(\sum_{i=1}^{I}\frac{R_i^2}{n_i}-\sum_{i=1}^{I}R_i(N+1)+\sum_{i=1}^{I}n_i\frac{(N+1)^2}{4}\right) \\
&= \frac{12}{N(N+1)}\left(\sum_{i=1}^{I}\frac{R_i^2}{n_i}-(N+1)\sum_{i=1}^{I}R_i+\frac{(N+1)^2}{4}\sum_{i=1}^{I}n_i\right) \\
&= \frac{12}{N(N+1)}\left(\sum_{i=1}^{I}\frac{R_i^2}{n_i}-(N+1)\frac{N(N+1)}{2}+\frac{(N+1)^2}{4}N\right) \\
&= \frac{12}{N(N+1)}\left(\sum_{i=1}^{I}\frac{R_i^2}{n_i}-\frac{N(N+1)^2}{4}\right) \\
&= \frac{12}{N(N+1)}\sum_{i=1}^{I}\frac{R_i^2}{n_i} \;-\; 3(N+1).
\end{aligned}$$

REMARQUE : On peut montrer que dans le cas $I = 2$ le test de Kruskal-Wallis est équivalent au test de la somme des rangs de Wilcoxon-Mann-Whitney.

5.14 L'anova à deux facteurs [3]

5.14.1 Description du modèle

On s'intéresse ici à l'effet de deux facteurs, disons le facteur A et le facteur B, sur une certaine variable, disons la variable Y. Le nombre de niveaux du facteur A est dénoté a et le nombre de niveaux du facteur B est dénoté b. Normalement on écrit n_{ij} pour dénoter la taille de l'échantillon numéro (i, j), c'est-à-dire le nombre d'unités expérimentales soumises au niveau i du traitement A et au niveau j du traitement B. Pour simplifier la présentation, nous allons supposer que les n_{ij} sont tous égaux et leur valeur commune sera dénotée n. Le nombre total d'observation N sera donc $N = nab$. On écrit Y_{ijk} pour dénoter la k^e observation pour la combinaison (i, j) des niveaux des facteurs A et B. On suppose que les conditions suivantes sont satisfaites.

- Pour chaque combinaison (i, j), les variables $Y_{ij1}, Y_{ij2}, ..., Y_{ijn}$ sont des $N(\mu_{ij}, \sigma^2)$ indépendantes les unes des autres.

[3] On peut omettre cette section si on manque de temps.

- Ces $a \times b$ échantillons aléatoires sont indépendants les uns des autres.
- Les variances théoriques sont inconnues mais on suppose qu'elles sont égales.

Ces conditions impliquent qu'on a

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \tag{5.21}$$

avec $\epsilon_{ijk} \sim N(0, \sigma^2)$, indépendantes les unes des autres. Si on pose

$$\begin{aligned}
\mu &= \overline{\mu}_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij} \\
\alpha_i &= \overline{\mu}_{i.} - \overline{\mu}_{..} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij} - \mu \\
\beta_j &= \overline{\mu}_{.j} - \overline{\mu}_{..} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij} - \mu \\
\gamma_{ij} &= \mu_{ij} - \overline{\mu}_{i.} - \overline{\mu}_{.j} + \overline{\mu}_{..} = \mu_{ij} - \alpha_i - \beta_j - \mu,
\end{aligned}$$

alors l'équation (5.21) peut s'écrire sous la forme

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}. \tag{5.22}$$

Notez que nos définitions des paramètres μ, α_i, β_j et γ_{ij} impliquent que

$$\begin{aligned}
\sum_{i=1}^{a}\alpha_i &= 0, \\
\sum_{j=1}^{b}\beta_j &= 0, \\
\sum_{i=1}^{a}\gamma_{ij} &= 0 \quad \text{pour chaque } j, \\
\sum_{j=1}^{b}\gamma_{ij} &= 0 \quad \text{pour chaque } i.
\end{aligned}$$

Le paramètre μ représente la moyenne globale, le paramètre α_i représente l'effet du niveau i du facteur A, le paramètre β_j représente l'effet du niveau j du facteur B et le paramètre γ_{ij} représente l'effet d'interaction entre le niveau i du facteur A et le niveau j du facteur B.

Le modèle

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{5.23}$$

est parfois appelé le modèle saturé. Lorsque le terme d'interaction est absent, c'est-à-dire lorsque les paramètres γ_{ij} sont tous nuls, on obtient

$$\mu_{ij} = \mu + \alpha_i + \beta_j.$$

Ce modèle est parfois appelé le modèle additif.

5.14.2 Les moyennes échantillonnales

On pose

$$\begin{aligned}
\overline{Y}_{...} &= \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} \\
\overline{Y}_{i..} &= \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} \\
\overline{Y}_{.j.} &= \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} Y_{ijk} \\
\overline{Y}_{ij.} &= \frac{1}{n}\sum_{k=1}^{n} Y_{ijk} \quad .
\end{aligned}$$

Les estimateurs des paramètres définis à la section précédente sont donnés par

$$\begin{aligned}
\hat{\mu} &= \overline{Y}_{...} \\
\hat{\alpha}_i &= \overline{Y}_{i..} - \overline{Y}_{...} \\
\hat{\beta}_j &= \overline{Y}_{.j.} - \overline{Y}_{...} \\
\hat{\gamma}_{ij} &= \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...} \\
\hat{\mu}_{ij} &= \overline{Y}_{ij.} \quad .
\end{aligned}$$

5.14.3 Le modèle saturé

La somme totale des carrés, en anglais *sum of squares total*, est donnée par

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{...}\right)^2 .$$

Cette somme totale des carrés peut être décomposée de la façon suivante :

$$SST = SS_{model} + SSE \tag{5.24}$$

avec

$$\begin{aligned}
SS_{model} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\overline{Y}_{ij.} - \overline{Y}_{...}\right)^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij.} - \overline{Y}_{...}\right)^2 \\
SSE &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{ij.}\right)^2 .
\end{aligned}$$

Dans l'equation (5.24), le terme SST représente la variation totale, le terme SS_{model} représente la variation expliquée par le modèle saturé et le terme SSE représente la variation résiduelle c'est-à-dire la variation qui ne s'explique pas par le modèle. Le rapport

$$R^2 = \frac{SS_{model}}{SST}$$

représente donc la proportion de la variation expliquée par le modèle. Le terme SS_{model} peut être décomposé en une composante SS_A qui représente la variation expliquée par le facteur A, une composante SS_B qui représente la variation expliquée par le facteur B et une composante SS_{AB} qui représente la variation expliquée par l'interaction ente le facteur A et le facteur B. Plus précisément, on obtient

$$SS_{model} = SS_A + SS_B + SS_{AB}$$

avec

$$\begin{aligned}
SS_A &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2 = bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2 \\
SS_B &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2 = an\sum_{j=1}^{b}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2 \\
SS_{AB} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2
\end{aligned}$$

et donc

$$SST = SS_A + SS_B + SS_{AB} + SSE. \tag{5.25}$$

Le nombre de degrés de liberté associé à SST est, comme dans le modèle d'anova à un facteur, $N-1$, où N est le nombre total d'observation. Comme on suppose ici que les tailles d'échantillon sont toutes égales, on a $N = nab$. Le nombre de degrés de liberté associé à SST est donc $nab-1$. À chaque somme de carrés qui apparaît du côté droit de l'égalité (5.25) correspond un nombre de degrés de liberté et une moyenne de carrés. Le tableau suivant nous donne les degrés de liberté de chaque terme ainsi que l'espérance de la moyenne des carrés correspondante.

Source de Variation	Somme de carrés	Degrés de liberté	Espérance de la moyenne des carrés
Modèle	SS_{model}	$ab-1$	$\mathbb{E}[MS_{model}] = \sigma^2 + \frac{n}{ab-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij}-\mu)^2$
Facteur A	SS_A	$a-1$	$\mathbb{E}[MS_A] = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\alpha_i^2$
Facteur B	SS_B	$b-1$	$\mathbb{E}[MS_B] = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_j^2$
Interaction AB	SS_{AB}	$(a-1)(b-1)$	$\mathbb{E}[MS_{AB}] = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}^2$
Erreur	SSE	$N-ab$	$\mathbb{E}[MSE] = \sigma^2$

Test pour la présence d'une interaction entre les deux traitements :

Pour tester

$$\begin{aligned} H_0 &: \text{ les } \gamma_{ij} \text{ sont tous nuls} \\ H_1 &: \text{ les } \gamma_{ij} \text{ ne sont pas tous nuls} \end{aligned}$$

on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{MS_{AB}}{MSE} \geq F_{(a-1)(b-1),N-ab,\alpha}. \tag{5.26}$$

Cette règle de décision est basée sur le fait que sous ce H_0 on a

$$\frac{MS_{AB}}{MSE} \sim F_{(a-1)(b-1),N-ab}. \tag{5.27}$$

Test pour la présence d'un effet du traitement B :

Pour tester

$$\begin{aligned} H_0 &: \text{ les } \beta_j \text{ sont tous nuls} \\ H_1 &: \text{ les } \beta_j \text{ ne sont pas tous nuls} \end{aligned}$$

on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{MS_B}{MSE} \geq F_{b-1,N-ab,\alpha}.$$

Cette règle de décision est basée sur le fait que sous ce H_0 on a

$$\frac{MS_B}{MSE} \sim F_{b-1,N-ab}.$$

Test pour la présence d'un effet du traitement A :

Pour tester

$$\begin{aligned} H_0 &: \text{ les } \alpha_i \text{ sont tous nuls} \\ H_1 &: \text{ les } \alpha_i \text{ ne sont pas tous nuls} \end{aligned}$$

on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{MS_A}{MSE} \geq F_{a-1,N-ab,\alpha}.$$

Cette règle de décision est basée sur le fait que sous ce H_0 on a

$$\frac{MS_A}{MSE} \sim F_{a-1,N-ab}.$$

Ces règles de décision peuvent toutes être justifiées de la même façon. Considérons par exemple la première, c'est-à-dire celle qui concerne les γ_{ij}. Le tableau ci-dessus nous dit que

$$\mathbb{E}[MS_{AB}] = \sigma^2 + \frac{n}{(a-1)(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij}^2.$$

Donc si les γ_{ij} sont tous nuls on a $\mathbb{E}[MS_{AB}] = \sigma^2$ et si les γ_{ij} ne sont pas tous nuls alors on a $\mathbb{E}[MS_{AB}] > \sigma^2$. Par ailleurs, on a toujours $\mathbb{E}[MSE] = \sigma^2$. Donc si les γ_{ij} sont tous nuls on s'attend à ce que le rapport MS_{AB}/MSE soit proche de 1 alors que si les γ_{ij} ne sont pas tous nuls on s'attend à ce que le rapport MS_{AB}/MSE soit beaucoup plus grand que 1. Cette observation justifie la règle de décision qui dit de rejeter l'hypothèse « H_0 : *les γ_{ij} sont tous nuls* » lorsque MS_{AB}/MSE est trop grand.

5.14.4 Table d'anova pour le modèle saturé

Les logiciels d'analyse statistique produisent habituellement une table d'anova plus ou moins semblable à la table d'anova suivante :

Source de Variation	Somme de carrés	Degrés de liberté	Moyenne de carrés	Statistique F	*p-value*
Modèle	SS_{model}	$ab-1$	$MS_{model} = \frac{SS_{model}}{a-1}$	$F_{model} = \frac{MS_{model}}{MSE}$	*p-value*
Facteur A	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MSE}$	*p-value*
Facteur B	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_B = \frac{MS_B}{MSE}$	*p-value*
Interaction AB	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{MS_{AB}}{MSE}$	*p-value*
Erreur	SSE	$N-ab$	$MSE = \frac{SSE}{N-ab}$	$***$	$***$
Total	SST	$N-1$	$***$	$***$	$***$

5.15 Exercices

Numéro 1. On considère un problème d'anova à un facteur avec 4 niveaux. Voici nos moyennes échantillonnales :

$$\overline{y}_{1\cdot} = 23.5 \qquad \overline{y}_{2\cdot} = 17.2 \qquad \overline{y}_{3\cdot} = 25.7 \qquad \overline{y}_{4\cdot} = 29.7$$

La moyenne échantillonnale globale $\overline{y}_{..}$ peut-elle être égale à 17.3 ? Peut-elle être égale à 16.8 ? Expliquez.

NUMÉRO 2. [Suite du numéro 1] Supposons qu'au numéro 1 les tailles des échantillons sont

$$n_1 = 7 \qquad n_2 = 18 \qquad n_3 = 11 \qquad n_4 = 8.$$

Calculez la moyenne échantillonnale globale $\overline{y}_{..}$. Calculez la somme des carrés inter-groupes SST_R. Calculez la moyenne des carrés inter-groupes MST_R.

NUMÉRO 3. [Suite des numéros 1 et 2] Voici les variances échantillonnales :

$$s_1^2 = 4.30 \qquad s_2^2 = 6.15 \qquad s_3^2 = 8.84 \qquad s_4^2 = 3.61.$$

Calculez l'erreur quadratique moyenne MSE.

NUMÉRO 4. [Suite des numéros 1, 2 et 3] Calculez la statistique de Bartlett. Que peut-on conclure ?

NUMÉRO 5. [Suite des numéros 1, 2, 3 et 4] En supposant l'homogénéité des variances théoriques, obtenez un intervalle de confiance de niveau 90% pour σ^2, la variance théorique commune aux quatre populations.

NUMÉRO 6. [Suite des numéros 1, 2, 3, 4 et 5] Toujours en supposant l'homogénéité des variances théoriques, obtenez un intervalle de confiance de niveau 95% pour $\frac{\mu_1+\mu_3+\mu_4}{3} - \mu_2$.

NUMÉRO 7. On fait une anova pour comparer 5 populations. Nos tailles d'échantillons sont toutes égales à 10. Quels sont l'espérance et l'écart-type de la statistique MST_R/MSE ?

NUMÉRO 8. [Suite du numéro 7] On obtient $SST_R = 109.2$ et $SST = 820.7$. Complétez la table d'anova. Avec le logiciel R ou avec une calculatrice scientifique munie de la loi de Fisher, obtenez le *p-value*. Tracez le graphe de la loi de Fisher et illustrez votre *p-value* sur ce graphe.

NUMÉRO 9. Dans un problème à deux échantillons indépendants issus de populations avec lois normales de même variance, on veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$. Le statisticien A affirme qu'on devrait utiliser le test de Student :

$$\text{On rejette } H_0 \text{ si } \left| \frac{\overline{Y}_1 - \overline{Y}_2}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \geq t_{n_1+n_2-2,\alpha/2}.$$

Le statisticien B affirme qu'on devrait utiliser le test de Fisher :

$$\text{On rejette } H_0 \text{ si } \frac{MST_R}{MSE} \geq F_{I-1,N-I,\alpha}$$

avec $I = 2$. Qui a raison ?

NUMÉRO 10. Dix-sept patients ont participé à une expérience visant à comparer 3 médicaments pour réduire la pression sanguine. On a mesuré la baisse de pression après 2 semaines. Voici les résultats :

Médicament A : 25.6, 32.1, 29.3, 28.8, 22.7

Médicament B : 31.6, 28.1, 23.4, 33.3, 24.8, 28.6

Médicament C : 28.7, 36.4, 31.3, 39.5, 33.7, 36.2

Ces trois médicaments sont-ils tous aussi bons les uns que les autres ? Faites le test approprié au seuil 5%

NUMÉRO 11. [Suite du numéro 10]. L'hypothèse d'égalité des variances est-elle raisonnable ? Faites le test de Bartlett au seuil 5%.

NUMÉRO 12. [Suite des numéros 10 et 11]. L'hypothèse de normalité est-elle raisonnable ? Obtenez les 17 résidus et faites les analyses appropriées.

NUMÉRO 13. Imaginez un problème d'anova à un facteur. Le nombre de niveaux du facteur est 4. Les tailles d'échantillons sont, dans l'ordre, 6, 6, 9 et 8. Nos observations sont dénotées de la façon usuelle :

Niveau du facteur	Observations								
1	$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	$Y_{1,4}$	$Y_{1,5}$	$Y_{1,6}$			
2	$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	$Y_{2,4}$	$Y_{2,5}$	$Y_{2,6}$			
3	$Y_{3,1}$	$Y_{3,2}$	$Y_{3,3}$	$Y_{3,4}$	$Y_{3,5}$	$Y_{3,6}$	$Y_{3,7}$	$Y_{3,8}$	$Y_{3,9}$
4	$Y_{4,1}$	$Y_{4,2}$	$Y_{4,3}$	$Y_{4,4}$	$Y_{4,5}$	$Y_{4,6}$	$Y_{4,7}$	$Y_{4,8}$	

Si les moyennes théoriques sont $\mu_1 = 9$, $\mu_2 = 10$, $\mu_3 = 8$ et $\mu_4 = 13$ et si la variance théorique commune aux quatre populations est $\sigma^2 = 4$, alors...

(a) la statistique $\overline{Y}_{3\cdot}$ sera environ _____, plus ou moins environ _____,

(b) la statistique $\overline{Y}_{\cdot\cdot}$ sera environ _____, plus ou moins environ _____,

(c) la statistique MSE sera environ _____, plus ou moins environ _____,

(d) la statistique MST_R sera environ _____.

NUMÉRO 14. Voici les résultats d'une expérience visant à comparer quatre populations. Ces données sont dans le fichier `STT-1920-chap-5-no-14.xls` disponible sur le site web du cours.

Échantillon	Observations										
1	34.5	36.8	29.5	32.1	33.6	29.7	34.2	34.0	31.9	33.7	38.2
2	31.8	28.2	27.0	30.7	29.3	27.9	35.1	30.9	28.5	29.9	28.7
3	33.0	31.6	35.7	35.8	32.6	33.7	32.3	34.8	33.9	33.1	31.8
4	30.8	34.6	31.3	35.0	34.5	37.8	34.7	32.6	32.4	34.4	31.9

On veut tester

H_0 : les moyennes théoriques sont toutes égales,

H_1 : les moyennes théoriques ne sont pas toutes égales.

(a) Énoncez les conditions sous lesquelles le test de Fisher est approprié.

(b) À l'aide de graphiques appropriés et de tests appropriés, vérifiez si les conditions énoncées à la partie (a) sont satisfaites.

(c) Effectuez le test de Fisher. Quelle est la valeur observée de votre statistique de Fisher ? Quel est le *p-value* associé ? Quelle est votre décision ?

(d) Obtenez un intervalle de confiance de niveau 90% pour l'écart-type théorique commun aux quatre populations.

(e) Obtenez un intervalle de confiance de niveau 95% pour le contraste

$$\frac{\mu_1 + \mu_3 + \mu_4}{3} - \mu_2.$$

NUMÉRO 15. Voici 125 observations classées selon la couleur. Ces données sont dans le fichier `STT-1920-chap-5-no-15.xls` disponible sur le site web du cours. On veut comparer les 5 distributions.

Bleu	Vert	Rouge	Noir	Jaune
1536005	1037156	817551	25466714	122327
4036940	95369	132725	2495468	26223
9121176	17630083	10988993	46771675	414274
561372	2179846	2362123	872203	43217
226819	176567	2358609	2056733	569447
19184439	1565221	7302478	1772772	248264
6380594	2211955	4942741	454121201	1134885
1486093	4745914	2345216	27672896	1414549
918371	620104	2152119	804446	1714239
4180787	110603	21810245	1824559	6132893
1595535	182505	1365467	6826294	882554
84018	2182961	1635582	3115058	55275
594743	156285	1964677	67859808	2952695
212286	983439	179832	5998559	101552
1094786	148723	3209975	21003165	10072737
454107	184586	97936	51585435	671710
733912	35609	4690566	21608034	3169040
700530	1579854	149800106	117701	132736
2976364	1690361	320830	4975482	428612
1655097	676494	11656708	60008674	3702570
330622	1526254	195161	117212057	9368103
425860	420161	27372667	3511415	42111
1698051	414037	78380565	47033958	13978
5583657	100809560	910144	7477766	38904
2649982	2036286	443837	6741484	1928170

(a) Examinez et interprétez les diagrammes en boîte juxtaposés pour les données originales y_{ij}.

(b) Examinez et interprétez les diagrammes en boîte juxtaposés pour les données transformées $\sqrt{y_{ij}}$.

(c) Examinez et interprétez les diagrammes en boîte juxtaposés pour les données transformées $1/y_{ij}$.

(d) Examinez et interprétez les diagrammes en boîte juxtaposés pour les données transformées $\log(y_{ij})$.

(e) Faites le test de Fisher avec les données convenablement transformées.

NUMÉRO 16. On a fait une expérience avec 4 types d'engrais (a, b, c et d) et 3 espèces de blé d'Inde (I, II et III). Pour notre expérience, on disposait de 32 plants de l'espèce I, 32 plants de l'espèce II et 32 plants de l'espèce III. Pour chacune des 3 espèces, on a formé 4 groupes de 8 plants et chaque groupe a reçu un des 4 engrais. À la fin de l'expérience, on a mesuré la hauteur des plants en cm. On a donc en tout 96 observations c'est-à-dire 8 obvervations pour chacune des 12 combinaisons possibles engrais-espèce. Les données sont dans le fichier Excel `STT-1920-chap-5-no-16.xls` disponible sur le site web du cours. Avec R-Commander, faites l'anova à deux facteurs. Y a-t-il un effet engrais ? Y a-t-il un effet espèce ? Y a-t-il un effet d'interaction ?

NUMÉRO 17. Une étude a été réalisée afin de comprendre l'effet de deux facteurs sur le rendement d'un certain processus. Trois niveaux du facteur A et quatre niveaux du facteur B ont été utilisés. Il y avait donc en tout 12 combinaisons différentes. Nous disposions de 72 unités expérimentales, soit 6 unités expérimentales pour chacune des 12 combinaisons. Voici les résultats. Vous les trouverez également dans le fichier `STT-1920-chap-5-no-17.xls` disponible sur le site web du cours.

Facteur B

Facteur A	*I*		*II*		*III*		*IV*	
1	19.2	16.1	20.9	19.6	21.9	20.7	21.0	20.9
	16.8	18.5	20.5	20.0	21.2	22.3	20.8	21.6
	16.4	19.3	19.0	20.3	20.7	20.9	21.0	22.1
2	17.9	18.2	20.7	20.6	20.3	21.8	21.2	20.3
	18.3	17.8	20.4	19.1	21.5	19.9	20.7	21.1
	19.3	18.6	20.0	20.4	21.4	20.3	23.9	19.5
3	17.4	18.0	19.7	20.1	21.5	20.3	22.1	21.8
	17.1	18.7	19.7	20.0	21.6	20.4	21.5	20.9
	17.6	17.4	18.7	19.6	20.9	21.5	22.0	21.2

(a) Obtenez la table d'anova appropriée. Quelles sont vos conclusions ? Y a-t-il un effet d'interaction ? Y a-t-il un effet principal du facteur A ? Y a-t-il un effet principal du facteur B ? Examinez les statistiques appropriées et les *p-values* appropriés.

(b) À l'aide de graphes appropriés et de tests appropriés, déterminez s'il est raisonnable de supposer que les conditions de normalité et d'égalité des variances sont satisfaites.

NUMÉRO 18. Une expérience a été réalisée afin de comparer différents types d'engrais. On a utilisé le même nombre d'unités expérimentales pour chacun des types d'engrais

considérés. Une fois l'expérience terminée, on a obtenu la table d'anova suivante :

Analysis of Variance Table

Source	Df	Sum Sq	Mean Sq	F value	p-value
Traitment	4	113.870	28.4675	2.1940	0.0848
Error	45	583.880	12.9751		
Total	49	697.750			

(a) Combien de types d'engrais ont été comparés ?

(b) Combien d'unités expérimentales a-t-on utilisées pour chaque type d'engrais ?

(c) Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette l'hypothèse d'égalité des moyennes de rendement sous les différents types d'engrais ?

(d) Les types d'engrais qu'on a comparés dans cette expérience s'appellent l'engrais A, l'engrais B, etc. Donnez une estimation de l'écart-type théorique de la population des rendements sous l'engrais B.

(e) Calculez un intervalle de confiance de niveau 95% pour l'écart-type théorique de la population des rendements sous l'engrais B.

Numéro 19.

(a) Dans un problème d'anova à un facteur, on suppose que nos échantillons sont les valeurs observées d'échantillons aléatoires indépendants issus de populations normales de même variance. Complétez la phrase suivante.

> Pour vérifier si l'hypothèse d'égalité des variances théoriques est raisonnable, on peut examiner les _____ juxtaposés et on peut utiliser le test de _____.

(b) Dans un problème d'anova à un facteur, on suppose que nos échantillons sont les valeurs observées d'échantillons aléatoires indépendants issus de populations normales de même variance. Choisissez la bonne réponse.

> Pour vérifier si l'hypothèse de normalité est raisonnable, j'ai calculé mes résidus, j'ai dessiné l'histogramme des résidus, j'ai tracé un *normal Q-Q plot* (graphe quantile-quantile gaussien) et j'ai fait un test de Shapiro et Wilk. Ma statistique de Shapiro et Wilk est 0.942 et mon *p-value* est 0.3131. Si ma statistique de Shapiro et Wilk avait été 0.842, mon *p-value* aurait été
>
> ☐ plus petit que 0.3131
>
> ☐ plus grand que 0.3131
>
> ☐ Nous n'avons pas assez d'informations pour répondre à cette question.

(c) Je veux comparer trois populations. Je soupçonne que les distributions théoriques sont des lois asymétriques de même forme. J'utilise la statistique de Kruskal et Wallis :

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

Si l'hypothèse d'égalité des distributions théoriques est vraie, alors je m'attends à ce que

la statistique K soit environ _____,

plus ou moins environ _____.

NUMÉRO 20. Voici les résultats d'une expérience visant à comparer quatre populations. Ces données sont dans le fichier `STT-1920-chap-5-no-20.xls` disponible sur le site web du cours.

Échantillon	Observations
Nord	8.47 5.42 15.87 0.21 1.79 2.30 8.03 21.23
Sud	6.51 6.44 4.01 4.61 3.50 4.79 3.12 12.42 2.20 24.02
Est	9.48 20.92 5.95 4.57 19.00 9.67 9.83 34.42 7.79 16.34 8.02
Ouest	1.94 1.68 2.50 7.19 16.07 1.16 2.02 5.03

On veut tester

H_0 : les moyennes théoriques sont toutes égales

H_1 : les moyennes théoriques ne sont pas toutes égales

(a) Les conditions du test de Fisher semblent-elles être satisfaites ? Justifiez votre réponse.

(b) Énoncez les conditions sous lesquelles le test de Kruskal-Wallis est approprié.

(c) Les conditions du test de Kruskal-Wallis semblent-elles être satisfaites ? Justifiez votre réponse.

(c) Effectuez le test de Kruskal-Wallis. Quelle est la valeur observée de votre statistique de Kruskal-Wallis ? Quel est le *p-value* associé ? Quelle est votre décision ?

Chapitre 6

Introduction à la régression

6.1 Deux exemples illustratifs

6.1.1 Exemple 1

Une expérience a été réalisée afin de mieux comprendre l'effet du nombre d'heures d'ensoleillement quotidien sur la production de tomates. Dans une serre située dans un désert du sud de la Californie où il fait soleil 365 jours par année, on a ensemencé 60 plants de tomates. À l'aide d'immenses rideaux noirs, on a contrôlé la quantité quotidienne de soleil :

- 10 plants ont reçu 3 heures de soleil par jour,
- 10 plants ont reçu 4 heures de soleil par jour,
- 10 plants ont reçu 5 heures de soleil par jour,
- 10 plants ont reçu 6 heures de soleil par jour,
- 10 plants ont reçu 7 heures de soleil par jour,
- 10 plants ont reçu 8 heures de soleil par jour.

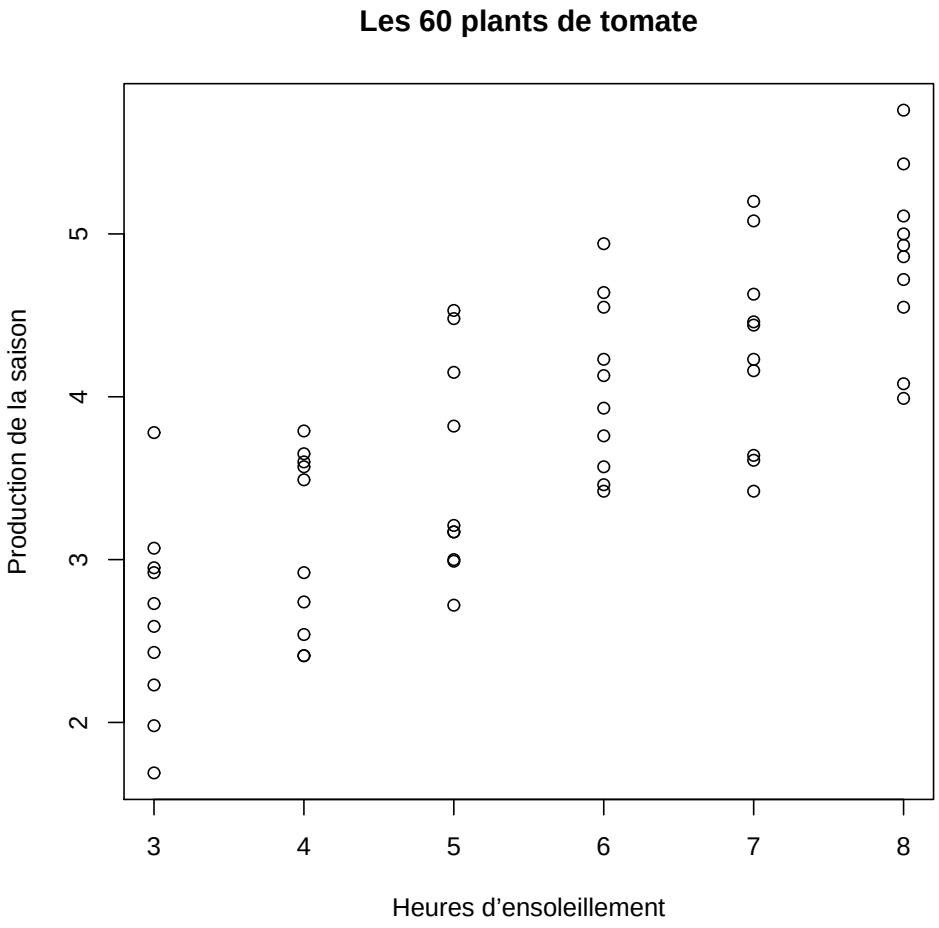
Quatre mois plus tard, à la fin de l'expérience, on a pris note de la production totale pour chaque plant. Voici les résultats en kg.

Soleil	Production (en kg)									
3 h	2.92	2.59	2.73	2.43	3.78	2.23	3.07	1.69	1.98	2.95
4 h	2.54	2.92	3.65	3.57	3.60	2.41	3.79	2.74	2.41	3.49
5 h	3.17	2.72	4.53	3.21	3.00	4.15	2.99	4.48	3.17	3.82
6 h	4.13	4.64	4.55	3.42	4.23	3.76	3.57	4.94	3.93	3.46
7 h	3.42	5.20	4.63	4.16	3.64	3.61	4.44	4.46	5.08	4.23
8 h	4.08	5.43	5.76	3.99	4.55	4.93	5.11	5.00	4.86	4.72

Voici (à la page suivante) le graphe de ces données. On note, entre autres choses, qu'il y a une relation linéaire entre les variables

X = nombre d'heures d'ensoleillement par jour

Y = nombre de kg de tomates produites par le plant durant la saison

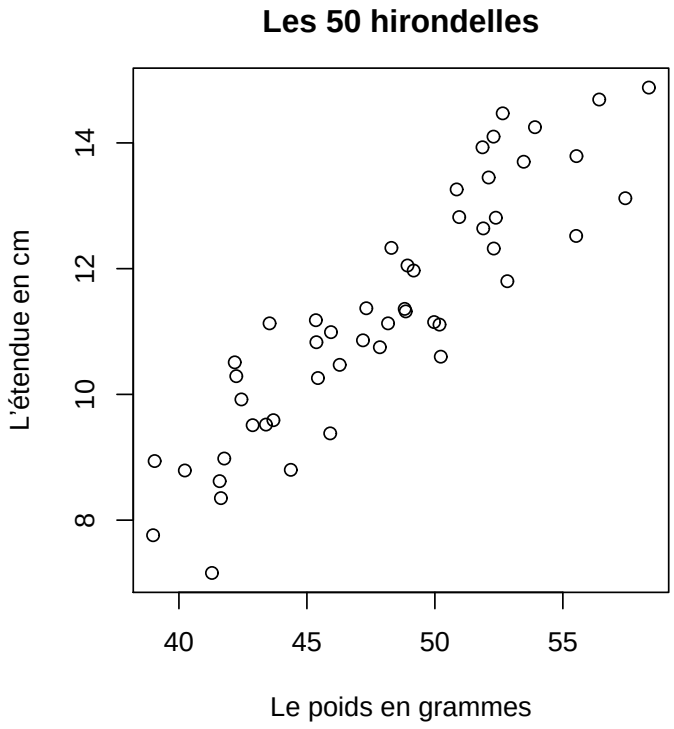


6.1.2 Exemple 2

On a obtenu un échantillon aléatoire de 50 hirondelles et on a mesuré, pour chaque oiseau, la variable `x = poids`, c'est-à-dire le poids de l'hirondelle en grammes, et la variable `y = étendue`, c'est-à-dire la distance en cm du bout de l'aile gauche jusqu'au bout de l'aile droite. Les données sont présentées à la page suivante. Il s'agit de 50 *points* (x_i, y_i), un pour chaque oiseau. Le graphe de ces 50 points, parfois appelé *diagramme de dispersion* (en anglais *scatterplot*), apparaît à la page suivante. À nouveau on note, entre autres choses, qu'il y a une relation linéaire entre la variable `x = poids` et la variable `y = étendue`.

Oiseau	Poids	Étendue
1	52.29	14.10
2	39.05	8.94
3	40.23	8.79
4	43.69	9.59
5	50.23	10.60
6	57.44	13.12
7	42.18	10.51
8	43.54	11.13
9	41.59	8.62
10	52.38	12.81
11	45.35	11.18
12	52.30	12.32
13	45.91	9.38
14	47.19	10.86
15	45.94	10.99
16	58.36	14.88
17	51.89	12.64
18	52.10	13.45
19	55.52	12.52
20	50.18	11.11
21	47.32	11.37
22	47.85	10.75
23	48.82	11.36
24	44.37	8.80
25	46.28	10.47

Oiseau	Poids	Étendue
26	48.93	12.05
27	48.17	11.13
28	55.53	13.79
29	50.85	13.26
30	42.88	9.51
31	48.86	11.32
32	52.65	14.47
33	45.43	10.26
34	43.40	9.52
35	50.95	12.82
36	42.24	10.29
37	48.30	12.33
38	56.42	14.69
39	49.97	11.15
40	41.64	8.35
41	42.44	9.92
42	51.86	13.93
43	38.99	7.76
44	53.91	14.25
45	41.29	7.16
46	49.17	11.97
47	52.83	11.80
48	41.77	8.98
49	53.47	13.70
50	45.37	10.83



6.2 Le modèle classique de régression linéaire simple

Dans les exemples de la section précédente, nous avons des données de la forme suivante :

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n).$$

Dans l'exemple 1, on avait $n = 60$ et

$$\begin{aligned}(x_1, y_1) &= (3, 2.92)\\ (x_2, y_2) &= (3, 2.59)\\ (x_3, y_3) &= (3, 2.73)\\ &\vdots\\ (x_{59}, y_{59}) &= (8, 4.86)\\ (x_{60}, y_{60}) &= (8, 4.72)\end{aligned}$$

Dans l'exemple 2, on avait $n = 50$ et

$$\begin{aligned}(x_1, y_1) &= (52.29, 14.10)\\ (x_2, y_2) &= (39.05, 8.94)\\ (x_3, y_3) &= (40.23, 8.79)\\ &\vdots\\ (x_{49}, y_{49}) &= (53.47, 13.70)\\ (x_{50}, y_{50}) &= (45.37, 10.83)\end{aligned}$$

Dans ces deux exemples, et dans bien des exemples qu'on rencontre en pratique, il est raisonnable de supposer qu'il existe des constantes β_0, β_1 et σ^2 telles que pour chaque point (x_i, y_i), le y_i est la valeur observée de la variable aléatoire

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{6.1}$$

où les *erreurs* $\epsilon_1, \epsilon_2, ..., \epsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées, avec distribution $N(0, \sigma^2)$. La droite $y = \beta_0 + \beta_1 x$ sera appelée la *droite de régression théorique* ou tout simplement la *droite de régression*.

Deux scénarios d'échantillonnage

Les exemples de la section 1 illustrent les deux principaux scénarios d'échantillonnage qu'on rencontre en pratique. Dans l'exemple 1, les x_i sont fixés à l'avance par la personne qui réalise l'expérience. Le scénario d'échantillonnage est semblable au scénario qu'on avait avec le modèle d'anova à un facteur, à l'exception que dans notre modèle de régression on suppose que la moyenne théorique de Y est une fonction linéaire de x. Donc, dans l'exemple 1, on suppose que les conditions suivantes sont satisfaites.

- $Y_1, Y_2, ..., Y_{10}$ est un échantillon aléatoire issu de la loi $N(\beta_0 + 3\beta_1, \sigma^2)$.
- $Y_{11}, Y_{12}, ..., Y_{20}$ est un échantillon aléatoire issu de la loi $N(\beta_0 + 4\beta_1, \sigma^2)$.
- $Y_{21}, Y_{22}, ..., Y_{30}$ est un échantillon aléatoire issu de la loi $N(\beta_0 + 5\beta_1, \sigma^2)$.
- $Y_{31}, Y_{32}, ..., Y_{40}$ est un échantillon aléatoire issu de la loi $N(\beta_0 + 6\beta_1, \sigma^2)$.
- $Y_{41}, Y_{42}, ..., Y_{50}$ est un échantillon aléatoire issu de la loi $N(\beta_0 + 7\beta_1, \sigma^2)$.
- $Y_{51}, Y_{52}, ..., Y_{60}$ est un échantillon aléatoire issu de la loi $N(\beta_0 + 8\beta_1, \sigma^2)$.
- Les variables Y_i sont indépendantes les unes des autres.
- Les paramètres β_0, β_1 et σ^2 sont inconnus et doivent être estimés.

Dans l'exemple 2, la situation est différente. Au lieu d'avoir plusieurs populations (une pour chaque valeur possible de x), on considère une population bivariée à partir de laquelle on obtient un échantillon aléatoire

$$(X_1, Y_1),\ (X_2, Y_2),\ (X_3, Y_3),\ \ldots,\ (X_{50}, Y_{50}).$$

Contrairement à l'exemple 1, cette fois-ci les x_i ne sont pas fixés à l'avance par la personne qui recueille les données. Les x_i sont les valeurs observées des variables aléatoires X_i.

Deux scénarios, une méthode

Malgré cette distinction importante entre les deux scénarios d'échantillonnage, la même méthode d'analyse statistique est utilisée. Cette méthode statistique s'appelle la régression.

À partir de l'équation (6.1), on peut écrire

$$\begin{aligned} \mathbb{E}[Y|X = x] &= \beta_0 + \beta_1\, x, \\ \mathbb{V}\mathrm{ar}[Y|X = x] &= \sigma^2. \end{aligned}$$

Les notations $\mathbb{E}[Y|X = x]$ et $\mathbb{V}\mathrm{ar}[Y|X = x]$ se lisent respectivement *l'espérance de Y sachant que $X = x$* et *la variance de Y sachant que $X = x$*.

6.3 Estimation des paramètres du modèle

Pour estimer les paramètres β_0 et β_1, on utilise la méthode des moindres carrés. Comme son nom l'indique, cette méthode consiste à prendre comme estimation de β_0 et β_1 les valeurs b_0 et b_1 qui minimisent la somme des carrés des distances entre les valeurs observées y_i et les valeurs *prédites* $b_0 + b_1 x_i$. Autrement dit, nos estimations de β_0 et β_1 sont les valeurs b_0 et b_1 qui minimisent la somme suivante :

$$\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2. \tag{6.2}$$

L'interprétation géométrique est simple : parmi toutes les droites possibles, on choisit celle pour laquelle la somme des carrés des distances verticales entre les points et la droite est

la plus petite possible. À l'aide des techniques du calcul différentiel, on montre facilement que les valeurs b_0 et b_1 qui minimisent cette somme de carrés sont

$$b_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{6.3}$$

$$b_0 = \overline{y} - b_1 \overline{x} \tag{6.4}$$

La droite $y = b_0 + b_1 x$ ainsi obtenue sera appelée la *droite de régression empirique* ou tout simplement la *droite de régression*.

En imaginant que les x_i sont fixés, on voit que b_0 et b_1 sont des combinaisons linéaires des y_i. Donc, vus comme estimateurs de β_0 et β_1, avec les x_i fixés, notre b_0 et notre b_1 sont des combinaisons linéaires des variables aléatoires Y_i. Il s'ensuit que b_0 et b_1 sont normalement distribués. En fait, il est facile de montrer que, conditionnellement aux x_i, on a

$$b_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}\right), \tag{6.5}$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right). \tag{6.6}$$

Pour estimer σ^2, on calcule d'abord la somme des erreurs au carré, en anglais SSE pour *sum of squares error* :

$$SSE = \sum_{i=1}^n \left(y_i - (b_0 + b_1 x_i)\right)^2. \tag{6.7}$$

Le nombre de degrés de liberté associé à cette somme est $n - 2$ car dans la somme SSE, les deux paramètres β_0 et β_1 ont été remplacés par les estimations b_0 et b_1. On a donc *perdu* deux degrés de liberté. Pour estimer le paramètre σ^2, on utilise donc

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}. \tag{6.8}$$

Le résultat suivant est semblable aux résultats qu'on avait dans le problème à un échantillon, le problème à deux échantillons et le problème à I échantillons :

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}. \tag{6.9}$$

6.4 Intervalle de confiance et tests d'hypothèses

Les résultats (6.5), (6.6) et (6.9) nous permettent de faire de l'inférence statistique (estimation, intervalle de confiance, tests d'hypothèses) de la façon usuelle.

6.4.1 Inférence pour le paramètre σ^2

Le résultat (6.9) entraîne, comme d'habitude, les conséquences suivantes :

1. MSE est un estimateur sans biais pour σ^2.
2. L'erreur type associée à l'estimation MSE est $\frac{\sqrt{2}\,MSE}{\sqrt{n-2}}$.
3. L'intervalle de confiance de niveau $1-\alpha$ pour σ^2 est

$$\left(\frac{(n-2)MSE}{\chi^2_{n-2,\frac{\alpha}{2}}}\,,\,\frac{(n-2)MSE}{\chi^2_{n-2,1-\frac{\alpha}{2}}}\right).$$

4. Pour tester $H_0 : \sigma^2 = \sigma_o^2$ contre $H_1 : \sigma^2 > \sigma_o^2$ au seuil α, on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{(n-2)MSE}{\sigma_o^2} \geq \chi^2_{n-2,\alpha}.$$

   L'étudiant peut deviner la règle de décision pour le cas $H_1 : \sigma^2 < \sigma_o^2$ et pour le cas $H_1 : \sigma^2 \neq \sigma_o^2$.

Dans la pratique, ce sont surtout les points 1, 2 et 3 qui nous intéressent.

6.4.2 Inférence pour le paramètre β_0

Les résultats (6.5) et (6.9) entraînent, comme d'habitude, les conséquences suivantes :

1. b_0 est un estimateur sans biais pour β_0.
2. L'erreur type associée à l'estimation b_0 est

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

3. L'intervalle de confiance de niveau $1-\alpha$ pour β_0 est

$$b_0 \pm t_{n-2,\frac{\alpha}{2}}\,\sqrt{\frac{MSE\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2}}.$$

4. Pour tester $H_0 : \beta_0 = \beta_0^*$ contre $H_1 : \beta_0 > \beta_0^*$ au seuil α, on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \frac{b_0-\beta_0^*}{\sqrt{\frac{MSE\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2}}} \geq t_{n-2,\alpha}.$$

   L'étudiant peut deviner la règle de décision pour le cas $H_1 : \beta_0 < \beta_0^*$ et pour le cas $H_1 : \beta_0 \neq \beta_0^*$.

Dans la pratique, ce sont surtout les points 1, 2 et 3 qui nous intéressent. Il arrive aussi qu'on veuille tester $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$. L'interprétation géométrique de l'hypothèse $H_0 : \beta_0 = 0$ est claire. La condition $\beta_0 = 0$ signifie que la droite de régression théorique passe par l'origine du plan cartésien.

6.4.3 Inférence pour le paramètre β_1

Les résultats (6.6) et (6.9) entraînent, comme d'habitude, les conséquences suivantes :

1. b_1 est un estimateur sans biais pour β_1.
2. L'erreur type associée à l'estimation b_1 est
$$\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}.$$
3. L'intervalle de confiance de niveau $1-\alpha$ pour β_1 est
$$b_1 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}.$$
4. Pour tester $H_0 : \beta_1 = \beta_1^*$ contre $H_1 : \beta_1 > \beta_1^*$ au seuil α, on utilise la règle de décision
$$\text{on rejette } H_0 \text{ si } \frac{b_1-\beta_1^*}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \geq t_{n-2,\alpha}.$$
L'étudiant peut deviner la règle de décision pour le cas $H_1 : \beta_1 < \beta_1^*$ et pour le cas $H_1 : \beta_1 \neq \beta_1^*$.

Dans la pratique, ce sont surtout les points 1, 2 et 3 qui nous intéressent. Il arrive aussi qu'on veuille tester $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. L'interprétation géométrique de l'hypothèse $H_0 : \beta_1 = 0$ est claire. La condition $\beta_1 = 0$ signifie que la droite de régression théorique est horizontale. D'un point de vue statistique, cela signifie que la variable X n'affecte pas la variable Y.

6.4.4 Inférence pour $\beta_0 + \beta_1 x_*$

Supposons que pour un certain x_* on veuille estimer la quantité $\beta_0+\beta_1 x_*$. On peut montrer que les résultats (6.5), (6.6) et (6.9) entraînent les conséquences suivantes :

1. $b_0 + b_1 x_*$ est un estimateur sans biais pour $\beta_0 + \beta_1 x_*$.
2. L'erreur type associée à l'estimation $b_0 + b_1 x_*$ est
$$\sqrt{MSE\left(\frac{1}{n}+\frac{(x_*-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}.$$
3. L'intervalle de confiance de niveau $1-\alpha$ pour $\beta_0 + \beta_1 x_*$ est
$$(b_0+b_1x_*) \pm t_{n-2,\frac{\alpha}{2}} \sqrt{MSE\left(\frac{1}{n}+\frac{(x_*-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}.$$
4. Pour tester $H_0 : \beta_0 + \beta_1 x_* = c$ contre $H_1 : \beta_0 + \beta_1 x_* > c$ au seuil α, on utilise la règle de décision
$$\text{on rejette } H_0 \text{ si } \frac{(b_0+b_1x_*)-c}{\sqrt{MSE\left(\frac{1}{n}+\frac{(x_*-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}} \geq t_{n-2,\alpha}.$$
L'étudiant peut deviner la règle de décision pour le cas $H_1 : \beta_0 + \beta_1 x_* < c$ et pour le cas $H_1 : \beta_0 + \beta_1 x_* \neq c$.

6.5 Retour à l'exemple 1

Quelques calculs élémentaires pour l'exemple 1 :

$$\overline{x} = 5.500, \qquad \overline{y} = 3.744, \qquad \sum_{i=1}^{60}(x_i - \overline{x})^2 = 175.000, \qquad \sum_{i=1}^{60}(x_i - \overline{x})(y_i - \overline{y}) = 75.470.$$

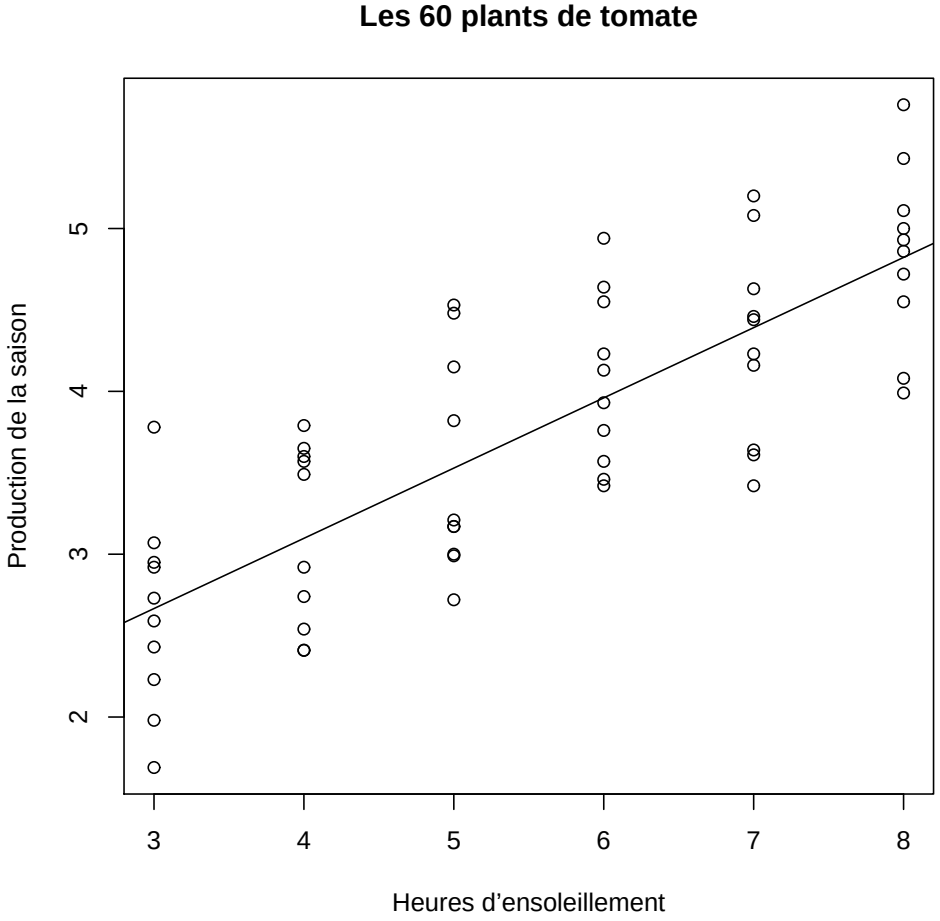
On obtient donc

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{75.470}{175} = 0.4316, \\ b_0 &= \overline{y} - b_1 \overline{x} = 3.744 - (0.43 \times 5.5) = 1.3724, \\ \hat{\sigma}^2 &= MSE = \frac{1}{n-2} \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 = 0.3228, \\ \hat{\sigma} &= \sqrt{MSE} = 0.5681. \end{aligned}$$

La droite de régression (empirique ou estimée) est donc

$$y = 1.3724 + 0.4316\ x$$

Son graphe apparaît ci-dessous.



Voici l'interprétation graphique des estimations b_0, b_1 et $\hat{\sigma}^2$.

Interprétation du b_0

L'estimation $b_0 = 1.3724$ est l'ordonnée à l'origine de la droite de régression empirique. Imaginez que dans le graphe ci-dessus on dessine l'axe des x jusqu'à la valeur $x = 0$ et on prolonge notre droite de régression jusqu'à ce qu'elle intersecte l'axe des y à $x = 0$. La droite de régression coupera alors l'axe des y à la hauteur $y = 1.3724$.

Interprétation du b_1

La pente de notre droite de régression est $b_1 = 0.4316$. Cela signifie que pour chaque heure d'ensoleillement additionnelle, la production augmente en moyenne de 0.4316 kg.

Interprétation du $\hat{\sigma}^2$

Dans le graphe ci-dessus, chaque point du graphe est à une certaine distance verticale de la droite de régression. La distance typique est $\hat{\sigma} = \sqrt{0.3227} = 0.5681$ kg.

Voici les erreurs types associées à nos estimations :

$$\begin{aligned}
\text{erreur type associée à } b_0 &= \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}} = 0.2473, \\
\text{erreur type associée à } b_1 &= \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} = 0.04295, \\
\text{erreur type associée à } \hat{\sigma}^2 &= \frac{\sqrt{2}\,\hat{\sigma}^2}{\sqrt{n-2}} = 0.0599.
\end{aligned}$$

6.6 Retour à l'exemple 2

Quelques calculs élémentaires pour l'exemple 2 :

$$\overline{x} = 47.9864, \qquad \overline{y} = 11.3106, \qquad \sum_{i=1}^{50} (x_i - \overline{x})^2 = 1244.766, \qquad \sum_{i=1}^{50} (x_i - \overline{x})(y_i - \overline{y}) = 429.3165.$$

On obtient donc

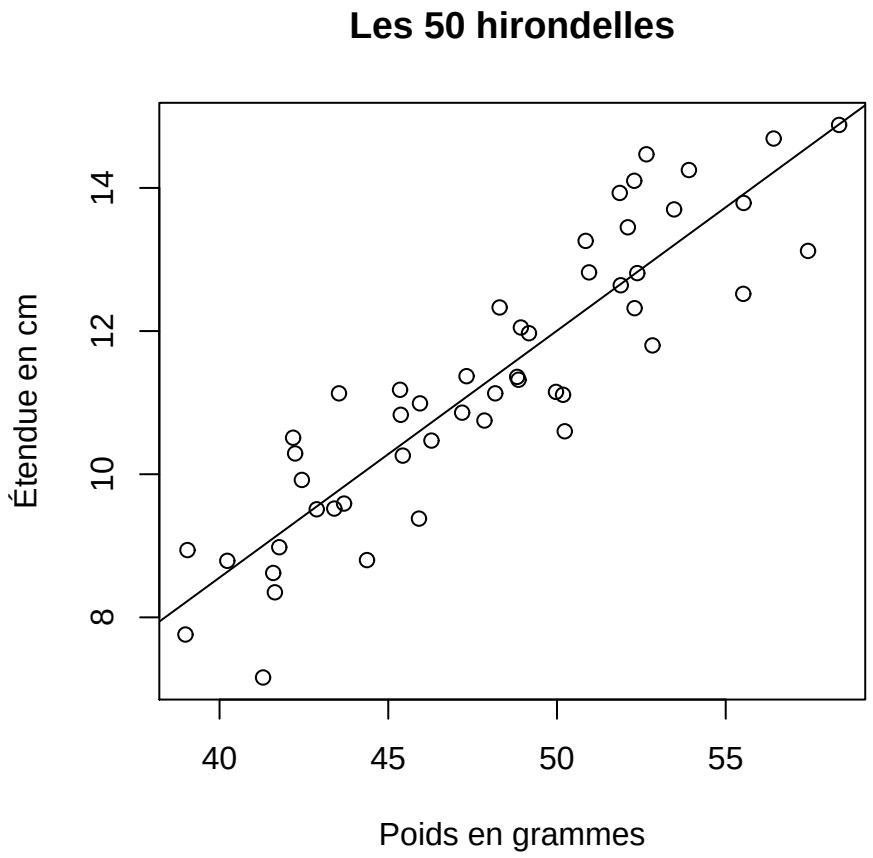
$$\begin{aligned}
b_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{429.3165}{1244.766} = 0.3445, \\
b_0 &= \overline{y} - b_1 \overline{x} = 11.3106 - (0.3445 \times 47.9864) = -5.2398, \\
\hat{\sigma}^2 &= MSE = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - (b_0 + b_1 x_i)\right)^2 = 0.6983, \\
\hat{\sigma} &= \sqrt{MSE} = 0.8356.
\end{aligned}$$

Ces estimations peuvent être interprétées de la même façon qu'à l'exemple 1. Les erreurs types associées à ces estimations peuvent être calculées de la même façon que celles calculées à la section précédente pour l'exemple 1.

Voici la droite de régression (empirique ou estimée) :

$$y = -5.2398 + 0.3445\ x$$

Son graphe apparaît ci-dessous.



6.7 Décomposition de la somme des carrés et table d'anova pour la régression

Considérons un scénario de régression linéaire simple et examinons à nouveau le SSE introduit à la section 6.3 :

$$SSE = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

En utilisant le fait que $b_0 = \overline{y} - b_1\overline{x}$, on obtient

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \\ &= \sum_{i=1}^{n}(y_i - (\overline{y} - b_1\overline{x} + b_1 x_i))^2 \\ &= \sum_{i=1}^{n}((y_i - \overline{y}) - b_1(x_i - \overline{x}))^2 \\ &= \sum_{i=1}^{n}\left\{(y_i - \overline{y})^2 - 2b_1(x_i - \overline{x})(y_i - \overline{y}) + b_1^2(x_i - \overline{x})^2\right\} \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2b_1\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \end{aligned}$$

Mais puisque

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

on voit que

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = b_1\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

On a donc

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2b_1\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2 + b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 - b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2. \end{aligned} \tag{6.10}$$

On a donc

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2 + SSE \tag{6.11}$$

Le terme de gauche dans l'équation (6.11) représente la variation totale en Y. Il est donc naturel de l'appeler SST (comme nous l'avions fait en ANOVA). On pose donc

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2. \tag{6.12}$$

Le premier terme du côté droit de l'égalité (6.11) s'appelle SSR, de l'anglais *Sum of Squares Regression*. On a donc

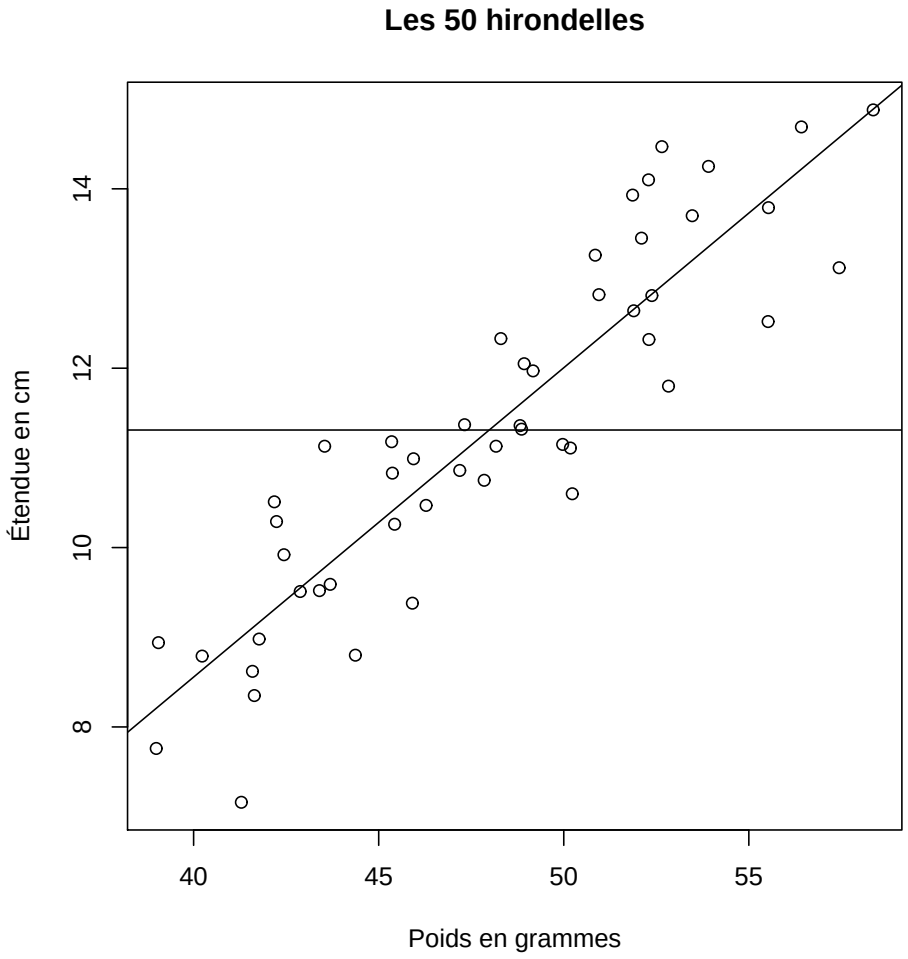
$$SSR = b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{6.13}$$

L'équation (6.11) peut donc s'écrire sous la forme suivante :

$$SST = SSR + SSE. \tag{6.14}$$

Cette équation nous montre que la variation totale est la somme de deux composantes : la composante SSR représente la variation en Y qui est expliquée par la variable X et la composante SSE représente la variation résiduelle c'est-à-dire la variation non expliquée par la variable X.

On peut interpréter les trois termes de l'équation (6.14) à l'aide d'un graphe. Reprenons l'exemple des 50 hirondelles. Sur le graphe ci-dessous, la droite oblique est la droite de régression obtenue à la section 6. La droite horizontale intersecte l'axe des y à la hauteur $\overline{y} = 11.31$ (l'étendue moyenne pour nos 50 hirondelles).



Interprétation du SST : Pour chacun des 50 points du graphe, on calcule la distance verticale entre le point et la droite horizontale. La statistique SST est la somme des carrés de ces 50 distances. L'équation (6.12) justifie cette interprétation.

Interprétation du SSE : Pour chacun des 50 points du graphe, on calcule la distance

verticale entre le point et la droite de régression. La statistique SSE est la somme des carrés de ces 50 distances. L'équation (6.7) justifie cette interprétation.

Interprétation du SSR : Pour chacun des 50 points du graphe, on trace une droite verticale passant par le point. Cette droite verticale intersecte la droite de régression et la droite horizontale. On calcule la distance entre ces deux points d'intersection. La statistique SSR est la somme des carrés de ces 50 distances. Cette interprétation est justifiée par le fait que le SSR de l'équation (6.13) peut aussi s'écrire sous la forme suivante :

$$SSR = \sum_{i=1}^{n} (\overline{y} - (b_0 + b_1 x_i))^2.$$

Le coefficient de détermination

L'équation (6.14) peut aussi s'écrire sous la forme

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}.$$

Les quantités SSR/SST et SSE/SST représentent respectivement la fraction de variation en Y expliquée par le modèle de régression et la fraction de variation en Y inexpliquée par le modèle de régression. Le rapport SSR/SST est appelée le *coefficient de détermination* et est souvent notée R^2. On a donc

$$R^2 = \frac{SSR}{SST}.$$

La table d'anova pour la régression

On a vu que le nombre de degrés de liberté associé à SSE est $n-2$ et que

$$\mathbb{E}[MSE] = \sigma^2.$$

On peut montrer que le nombre de degrés de liberté associé à SSR est 1 et que

$$\mathbb{E}[MSR] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Donc si $\beta_1 = 0$ alors on s'attend à ce que le rapport MSR/MSE soit aux alentours de 1 tandis que si $\beta_1 \neq 0$ alors on s'attend à ce que le rapport MSR/MSE soit beaucoup plus grand que 1. On peut montrer que si $\beta_1 = 0$, alors on a

$$\frac{MSR}{MSE} \sim F_{1,n-2}.$$

Pour tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ on peut donc utiliser la règle de décision suivante :

$$\text{on rejette } H_0 \text{ si } \frac{MSR}{MSE} \geq F_{1,n-2,\alpha}. \tag{6.15}$$

Le calcul du rapport MSR/MSE est habituellement présenté sous forme de la table d'anova suivante, appelée *table d'anova pour la régression* :

Source	Sum Sq	df	Mean Sq	F	p-value
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$\mathbb{P}_{H_0}[F \geq F_{obs}]$
Residuals	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
Total	SST	$n-1$			

REMARQUE. D'après ce qui a été fait à la section 4, pour tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$, on peut utiliser la règle de décision

$$\text{on rejette } H_0 \text{ si } \left| \frac{b_1}{\sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \right| \geq t_{n-2,\alpha/2}. \tag{6.16}$$

La règle de décision (6.16) peut s'écrire sous la forme suivante :

$$\text{on rejette } H_0 \text{ si } \left(\frac{b_1}{\sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \right)^2 \geq t^2_{n-2,\alpha/2},$$

c'est-à-dire

$$\text{on rejette } H_0 \text{ si } \frac{b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2}{MSE} \geq t^2_{n-2,\alpha/2},$$

c'est-à-dire

$$\text{on rejette } H_0 \text{ si } \frac{MSR}{MSE} \geq t^2_{n-2,\alpha/2}.$$

On peut montrer que $t^2_{n-2,\alpha/2} = F_{1,n-2,\alpha}$. Les règles de décision (6.15) et (6.16) sont donc complètement équivalentes !

6.8 Intervalle de prédiction [1]

Revenons à l'exemple des 50 hirondelles. Imaginez qu'on capture une 51-ième hirondelle. Le poids de cette hirondelle est $X = 53.40$ grammes. On aimerait pouvoir prédire l'étendue Y de cette hirondelle. Si on connaissait les paramètres β_0, β_1 et σ^2, alors on pourrait raisonner de la façon suivante. Sachant que $X = 53.40$, la distribution de Y est la loi $N(\beta_0 + \beta_1(54.40), \sigma^2)$. On est donc 95% certain que Y se situe dans l'intervalle suivant :

$$((\beta_0 + 53.40\,\beta_1) - 1.96\,\sigma\,,\ (\beta_0 + 53.40\beta_1) + 1.96\,\sigma)\,.$$

Si on remplace 53.40 par une valeur quelconque x_* et 95% par un niveau quelconque $1-\alpha$, alors cet intervalle devient

$$\big((\beta_0 + \beta_1 x_*) - z_{\alpha/2}\,\sigma\,,\ (\beta_0 + \beta_1 x_*) + z_{\alpha/2}\,\sigma\big)\,. \tag{6.17}$$

[1] On peut omettre cette section si on manque de temps.

Cet intervalle s'appelle un intervalle de prédiction de niveau $1-\alpha$ pour une future observation Y prise à $X = x$. Or on ne connaît pas les vraies valeurs des paramètres β_0, β_1 et σ^2. On connaît seulement les estimations b_0, b_1 et MSE. Sous ces conditions, on peut montrer que l'intervalle de prédiction pour l'étendue Y de notre 51-ième hirondelle est

$$(b_0 + 53.40\, b_1) \pm t_{48,0.025} \sqrt{MSE \left(1 + \frac{1}{50} + \frac{(53.40 - \overline{x})^2}{\sum_{i=1}^{50}(x_i - \overline{x})^2}\right)}.$$

Voici la forme générale de l'intervalle de prédiction de niveau $1-\alpha$ pour une future observation Y pour laquelle la variable X prend la valeur x_* :

$$(b_0 + b_1 x_*) \pm t_{n-2,\frac{\alpha}{2}} \sqrt{MSE \left(1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}. \tag{6.18}$$

Il ne faut pas confondre cet intervalle de prédiction avec l'intervalle de confiance pour $\beta_0 + \beta_1 x_*$ obtenu à la section 4 et reproduit ci-dessous :

$$(b_0 + b_1 x_*) \pm t_{n-2,\frac{\alpha}{2}} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}. \tag{6.19}$$

Pour comprendre la différence entre l'intervalle (6.18) et l'intervalle (6.19), imaginez que n est très très grand. Qu'arrive-t-il à l'intervalle (6.18) ? Qu'arrive-t-il à l'intervalle (6.19) ? Si n est très très grand, alors on aura

$$\begin{aligned} b_0 + b_1 x_* &\approx \beta_0 + \beta_1 x_* \\ t_{n-2,\frac{\alpha}{2}} &\approx z_{\alpha/2} \\ MSE &\approx \sigma^2 \\ 1/n &\approx 0 \\ \frac{(x_* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} &\approx 0. \end{aligned}$$

Donc, avec n super grand, l'intervalle (6.18) devient identique à l'intervalle (6.17) alors que l'intervalle (6.19) dégénère en un point, le point $\beta_0 + \beta_1 x_*$. Bref, si n est extrêmement grand, on estime $\beta_0 + \beta_1 x_*$ avec une précision extrême mais on ne pourra jamais prédire Y avec une précision extrême !

6.9 Le coefficient de corrélation

6.9.1 Introduction

Étant donné un ensemble de n points

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n),$$

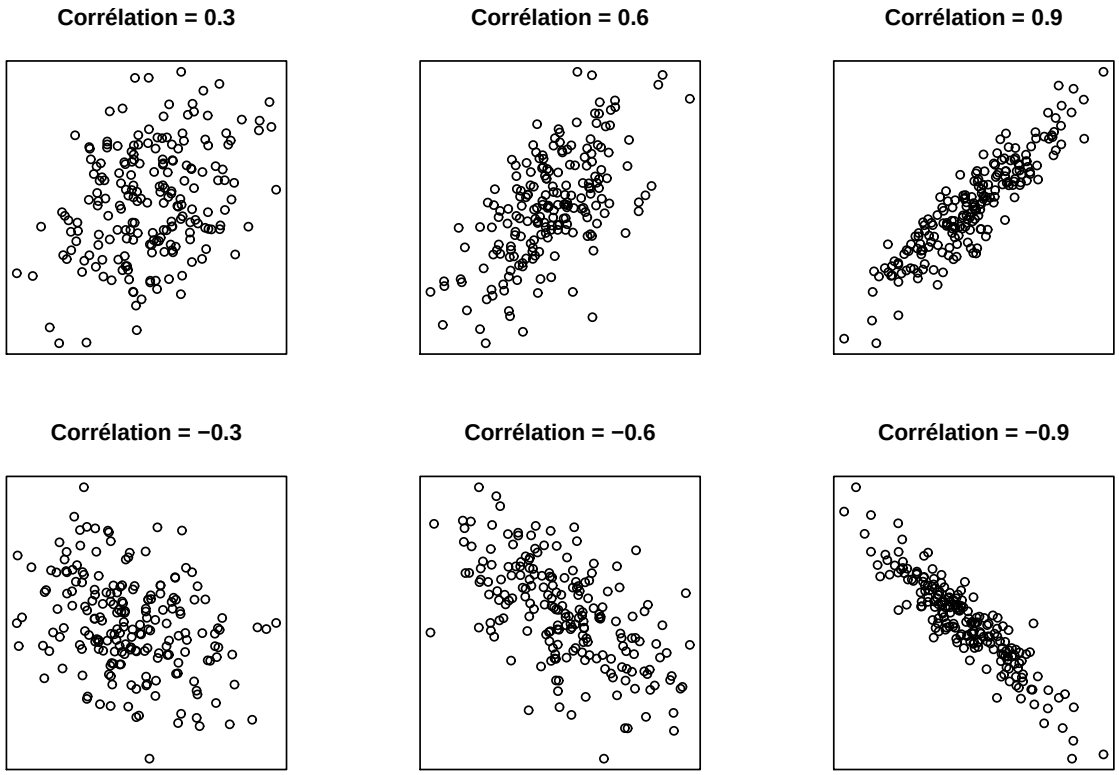
on définit le coefficient de corrélation r à l'aide de l'équation

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{6.20}$$

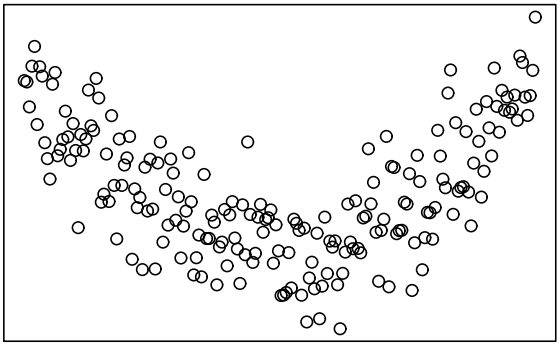
Parmi les principales propriétés du coefficient de corrélation, notons les suivantes :

1. Le coefficient de corrélation mesure le degré d'association linéaire entre la variable X et la variable Y.
2. On a toujours $-1 \leq r \leq 1$.
3. On obtient $r = 1$ si et seulement si les n points sont sur une droite de pente positive et on obtient $r = -1$ si et seulement si les n points sont sur une droite de pente négative.
4. Un coefficient de corrélation nul indique l'absence d'association linéaire.
5. Un coefficient de corrélation positif indique une association linéaire *positive* entre la variable X et la variable Y : plus X est grand et plus Y a tendance à être grand ; plus X est petit et plus Y a tendance à être petit. Cette association est d'autant plus forte que r est proche de 1.
6. Un coefficient de corrélation négatif indique une association linéaire *négative* entre la variable X et la variable Y : plus X est grand et plus Y a tendance à être petit ; plus X est petit et plus Y a tendance à être grand. Cette association est d'autant plus forte que r est proche de -1.

Voici quelques exemples de nuages de points et leurs coefficients de corrélation :



Lorsqu'il y a une association non linéaire entre la variable X et la variable Y, le coefficient de corrélation détecte mal cette association. Voici un exemple avec $r \approx 0$.



Bref, dans cet exemple, si on calcule r avec la formule (6.20), on obtient $r \approx 0$. Ceci indique qu'il n'y a pas d'association *linéaire* entre les deux variables. Or le graphe indique clairement qu'il y a une certaine association entre les deux variables : pour un x aux alentours de $\overline{x}$, la valeur y a tendance à être petite ; pour un x loin de $\overline{x}$, la valeur y a tendance à être grande.

6.9.2 Le coefficient de corrélation et la droite de régression

Il y a un lien entre le coefficient de régression r et la pente b_1 de la droite de régression. À partir de l'équation (6.20) on obtient

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \times \frac{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \times \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2}} \\ &= b_1 \, \frac{s_x}{s_y}. \end{aligned}$$

On a donc

$$b_1 = r \, \frac{s_y}{s_x}$$

et on obtient

$$\begin{aligned} y &= b_0 + b_1 x \\ &= (\overline{y} - b_1 \overline{x}) + b_1 x \\ &= \overline{y} + b_1 (x - \overline{x}) \\ &= \overline{y} + r \, \frac{s_y}{s_x} (x - \overline{x}) \\ &= \overline{y} + r s_y \left(\frac{x - \overline{x}}{s_x} \right). \end{aligned}$$

Notre droite de régression peut donc s'écrire sous la forme

$$y = \overline{y} + r s_y \left(\frac{x - \overline{x}}{s_x} \right). \tag{6.21}$$

Cette équation est extrêmement importante ! Elle permet d'expliquer ce qu'on appelle *l'effet de régression*.

6.9.3 L'effet de régression

Voici un bel exemple. Considérons papa et fiston. Si papa est très grand, doit-on s'attendre à ce que fiston soit très grand ? À la fin du 19-ième siècle, les scientifiques britanniques s'intéressaient beaucoup à ce genre de question. Deux d'entre eux, Pearson et Lee, examinèrent 1078 couples papa-fiston et mesurèrent (en pouces) la taille de papa et la taille de fiston (en âge adulte). Le graphe des 1078 points ainsi obtenus, avec la droite de régression, est présenté à la page suivante. Voici quelques statistiques calculées à partir de ces 1078 points :

$$\begin{array}{lll} \overline{x} = 67.687 & s_x = 2.745 & r = 0.501 \\ \overline{y} = 68.684 & s_y = 2.815 & \end{array}$$

Imaginez un papa dont la hauteur serait deux écarts-types en haut de la moyenne. Naïvement, on pourrait s'attendre à ce que son fils soit environ deux écarts-types en haut de la moyenne. Or ce n'est pas le cas ! Dénotons par x_* la grandeur de ce papa. Dire que papa est deux écarts-types en haut de la moyenne c'est dire que

$$x_* = \overline{x} + 2 s_x.$$

Pour prédire la grandeur de fiston, on insère ce x_* dans l'équation de notre droite de régression. Autrement dit on remplace x par $x_* = \overline{x} + 2 s_x$ dans l'équation (6.21). On obtient alors

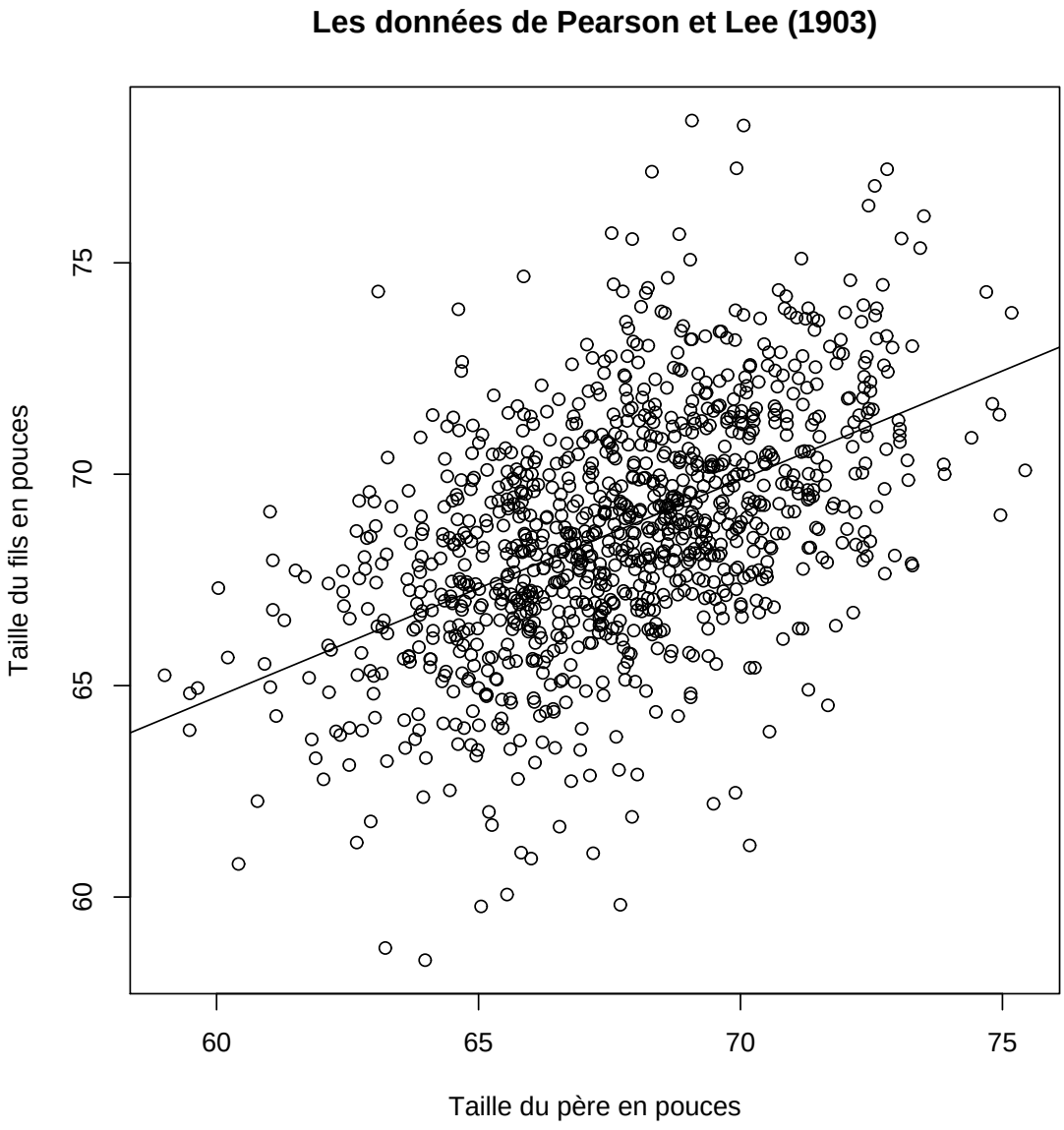
$$y = \overline{y} + r s_y \left(\frac{x_* - \overline{x}}{s_x} \right) = \overline{y} + r s_y \left(\frac{(\overline{x} + 2 s_x) - \overline{x}}{s_x} \right) = \overline{y} + 2 r s_y$$

Il faut donc s'attendre à ce que fiston soit non pas 2 écarts-types en haut de la moyenne mais bien $2r$ écarts-types en haut de la moyenne ! Dans notre exemple on a $r = 0.501$. Donc si papa est 2 écarts-types en haut de la moyenne, alors on s'attend à ce que fiston soit environ 1.002 écarts-types en haut de la moyenne. Voici le résultat général (avec « deux écarts-types » remplacé par « k écarts-types ») :

L'effet de régression

> Dans un modèle de régression linéaire simple, les individus dont la valeur de la variable X est k écarts-types en haut de la moyenne seront, pour la variable Y, en moyenne kr écarts-types en haut de la moyenne.

Cet *effet de régression* est la raison pour laquelle l'ensemble des méthodes statistiques étudiées dans le présent document s'appelle la *théorie de la régression*.



6.9.4 Le coefficient de corrélation théorique

Si les données

$$(x_1, y_1), \quad (x_2, y_2), \quad (x_3, y_3), \ \cdots, (x_x, y_n)$$

sont le résultat de n tirages à partir d'une distribution bivariée (comme dans l'exemple 2 de la section 6.1), alors le coefficient de corrélation r peut être vu comme étant une estimation du coefficient de corrélation théorique

$$\rho = \frac{\mathbb{C}\text{ov}[X, Y]}{\sigma_X \ \sigma_Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \ \sigma_Y}. \tag{6.22}$$

Ce coefficient de corrélation théorique possède des propriétés analogues au coefficient de corrélation échantillonnal r. Par exemple, on peut montrer qu'on a toujours $-1 \leq \rho \leq 1$. Le numérateur $\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ est appelé la *covariance* entre les variables X et Y. Notez que le coefficient de corrélation r défini à l'équation (6.20) peut aussi s'écrire sous la forme suivante :

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X \ s_Y}. \tag{6.23}$$

Le numérateur du terme de droite de l'équation (6.23) s'appelle la covariance échantillonnale. En comparant les équations (6.22) et (6.23), on voit bien que r est une bonne estimation pour ρ puisque

(a) la covariance échantillonnale $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ est une bonne estimation de la covariance théorique $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$,

(b) l'écart-type échantillonnal s_X est une bonne estimation pour l'écart-type σ_X,

(c) l'écart-type échantillonnal s_Y est une bonne estimation pour l'écart-type σ_Y.

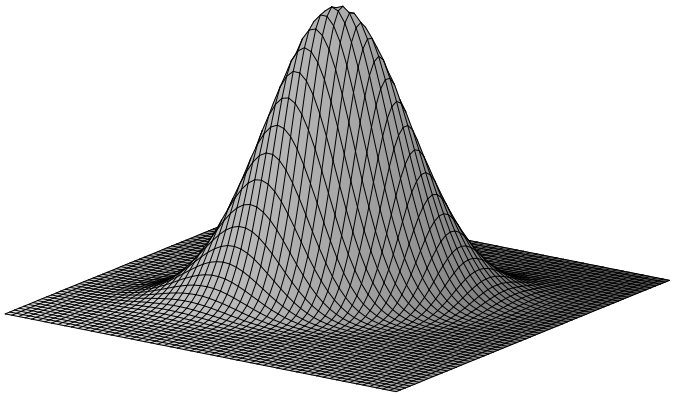
La distribution de la statistique r est très compliquée. Pour faire de l'inférence statistique (intervalle de confiance, test d'hypothèse,...), on utilise le résultat approximatif suivant : si n est assez grand, disons $n \geq 30$, alors on a

$$\frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{1/\sqrt{n-3}} \approx N(0, 1). \tag{6.24}$$

Ce résultat est valide lorsque l'échantillon aléatoire $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ provient d'une loi normale bivariée. Il n'est pas nécessaire d'avoir étudié la loi normale bivariée en détail pour évaluer si le résultat (6.24) s'applique. Voici la densité de la loi normale bivariée :

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \ \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}}.$$

Si on fait le graphe de cette densité conjointe, on obtient une belle surface en forme de cloche. En voici un exemple :



On peut reconnaître la loi normale bivariée en examinant ses courbes de niveaux, c'est-à-dire en examinant sa carte topographique. Celle-ci devrait présenter des ellipses concentriques centrées au point (μ_X, μ_Y). Lorsqu'on fait le graphe des n points d'un échantillon aléatoire issu d'une loi normale bivariée, on obtient un nuage de point semblable à celui de l'exemple des tailles des pères et des fils. Bref, pour l'exemple des tailles des pères et des fils, la loi normale bivariée est un très bon modèle. Il en est de même pour l'exemple des 50 hirondelles.

Intervalle de confiance pour ρ

À partir du résultat (6.24) on obtient

$$\mathbb{P}\left[-z_{\alpha/2} < \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{1/\sqrt{n-3}} < z_{\alpha/2}\right] = 1-\alpha$$

et on en déduit l'intervalle de confiance

$$\left(\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\,,\ \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

pour la quantité $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$. Or, dire que $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ est compris dans l'intervalle ci-dessus est équivalent à dire que ρ est compris dans l'intervalle

$$\left(\frac{\mathrm{e}^{2a}-1}{\mathrm{e}^{2a}+1}\,,\ \frac{\mathrm{e}^{2b}-1}{\mathrm{e}^{2b}+1}\right) \tag{6.25}$$

avec

$$a = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{z_{\alpha/2}}{\sqrt{n-3}} \qquad \text{et} \qquad b = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{z_{\alpha/2}}{\sqrt{n-3}}.$$

L'intervalle donné par l'équation (6.25) est donc un intervalle de confiance de niveau $1-\alpha$ pour ρ. Voici deux exemples.

L'exemple des 50 hirondelles : Nous avons obtenu $r = 0.903$. Pour l'intervalle de confiance de niveau 95%, on obtient d'abord

$$\begin{aligned} a &= \frac{1}{2}\ln\left(\frac{1+(0.903)}{1-(0.903)}\right) - \frac{1.96}{\sqrt{50-3}} = 1.2024 \\ b &= \frac{1}{2}\ln\left(\frac{1+(0.903)}{1-(0.903)}\right) + \frac{1.96}{\sqrt{50-3}} = 1.7742. \end{aligned}$$

On insère ces deux valeurs dans l'intervalle (6.25) et on obtient l'intervalle $(0.834, 0.944)$.

L'exemple de Pearson et Lee : Nous avons obtenu $r = 0.501$. Pour l'intervalle de confiance de niveau 99%, on obtient d'abord

$$\begin{aligned} a &= \frac{1}{2}\ln\left(\frac{1+(0.501)}{1-(0.501)}\right) - \frac{2.576}{\sqrt{1078-3}} = 0.4913 \\ b &= \frac{1}{2}\ln\left(\frac{1+(0.501)}{1-(0.501)}\right) + \frac{2.576}{\sqrt{1078-3}} = 0.6109. \end{aligned}$$

On insère ces deux valeurs dans l'intervalle (6.25) et on obtient l'intervalle $(0.455, 0.545)$.

Test d'hypothèses concernant ρ

Supposons qu'on veuille tester $H_0 : \rho = \rho_o$ contre $H_1 : \rho \neq \rho_o$. Ces hypothèses sont équivalentes aux suivantes :

$$\begin{aligned} H_0 : \ & \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_o}\right) \\ H_1 : \ & \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \neq \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_o}\right) \end{aligned}$$

Étant donné le résultat (6.24), on peut utiliser la règle de décision

$$\text{on rejette } H_0 \text{ si } \left|\frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_o}\right)}{1/\sqrt{n-3}}\right| \geq z_{\alpha/2}.$$

Pour les tests unilatéraux, on fait les modifications usuelles.

Le cas spécial $H_0 : \rho = 0$.

La règle de décision du paragraphe précédent est basée sur le résultat approximatif (6.24). Ce résultat est valide à condition que n soit suffisamment grand et que la loi normale bivariée soit un bon modèle pour la distribution conjointe de X et Y. Dans le cas où l'hypothèse nulle est $H_0 : \rho = 0$, il y a une approche alternative qui ne nécessite pas ces deux conditions. Pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$, on utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } \left|\frac{\sqrt{n-2}\ r}{\sqrt{1-r^2}}\right| \geq t_{n-2,\alpha/2}.$$

Ce résultat est basé sur le fait que si $\rho = 0$, alors on peut montrer que

$$\frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} \sim t_{n-2}. \tag{6.26}$$

6.10 Le lien entre r et MSE [2]

D'après les formules obtenues à la section 6.7, l'équation $SST = SSR + SSE$ peut s'écrire sous la forme

$$(n-1)s_Y^2 = b_1^2(n-1)s_X^2 + (n-2)MSE.$$

À la section 6.9 on a vu que $b_1 = rs_Y/s_X$. Si on insère cela dans l'équation précédente, on obtient

$$(n-1)s_Y^2 = r^2 s_Y^2(n-1) + (n-2)MSE.$$

Si on isole MSE on obtient, après simplification,

$$MSE = \frac{n-1}{n-2}(1-r^2)s_Y^2. \tag{6.27}$$

Remarque 1. Lorsque n est suffisamment grand, l'équation (6.27) nous donne

$$MSE \approx (1-r^2)s_Y^2. \tag{6.28}$$

Puisque $-1 \leq r \leq 1$, on a toujours $0 \leq 1 - r^2 \leq 1$. Rappelons que s_Y^2 est l'estimation de la variance théorique σ_Y^2 alors que MSE est l'estimation de la variance conditionnelle théorique $\sigma^2 = \sigma_{Y|X=x}^2$. L'équation (6.28) nous donne le lien entre ces deux estimations.

Remarque 2. L'équation (6.27) nous donne

$$(n-2)MSE = (1-r^2)(n-1)s_Y^2,$$

c'est-à-dire

$$SSE = (1-r^2)SST.$$

On a donc

$$\frac{SSE}{SST} = 1 - r^2,$$

c'est-à-dire

$$1 - \frac{SSR}{SST} = 1 - r^2,$$

c'est-à-dire

$$1 - R^2 = 1 - r^2$$

où R^2 est le coefficient de détermination introduit à la section 6.7. On a donc $R^2 = r^2$. Autrement dit, dans le modèle de régression linéaire simple, le coefficient de détermination est égal au carré du coefficient de corrélation.

[2] On peut omettre cette section, ou la laisser en lecture, si on manque de temps.

REMARQUE 3. L'équation (6.27) nous permet de démontrer le résultat (6.26). On sait que

$$\frac{b_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim t_{n-2}.$$

Or on obtient

$$\frac{b_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} = \frac{rs_Y/s_x}{\sqrt{\frac{\frac{n-1}{n-2}(1-r^2)s_Y^2}{(n-1)s_X^2}}} = \frac{\sqrt{n-2}\,r}{\sqrt{1-r^2}}.$$

Cela nous donne le résultat (6.26).

6.11 La vérification des hypothèses du modèle

Pour que les méthodes statistiques introduites dans le présent chapitre soient appropriées, il faut que les conditions énoncées à la section 6.2 soient satisfaites. Étant donné un ensemble de données bivariées $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, comment peut-on vérifier si ces conditions sont satisfaites ?

Notre principal outil est le diagramme de dispersion. Sur ce graphe, on devrait pouvoir constater les points suivants.

1. L'association entre les variables X et Y est linéaire. Autrement dit, l'espérance conditionnelle $\mathbb{E}[Y|X = x]$ est une fonction linéaire de x.
2. Pour une valeur x fixe, la variation en Y est sensiblement la même, peu importe le x qu'on choisit. Autrement dit, la variance conditionnelle $\mathbb{V}\text{ar}[Y|X = x]$ est la même pour tout x.
3. Pour une valeur x fixe, la distribution de Y est une loi normale, peu importe le x qu'on choisit.

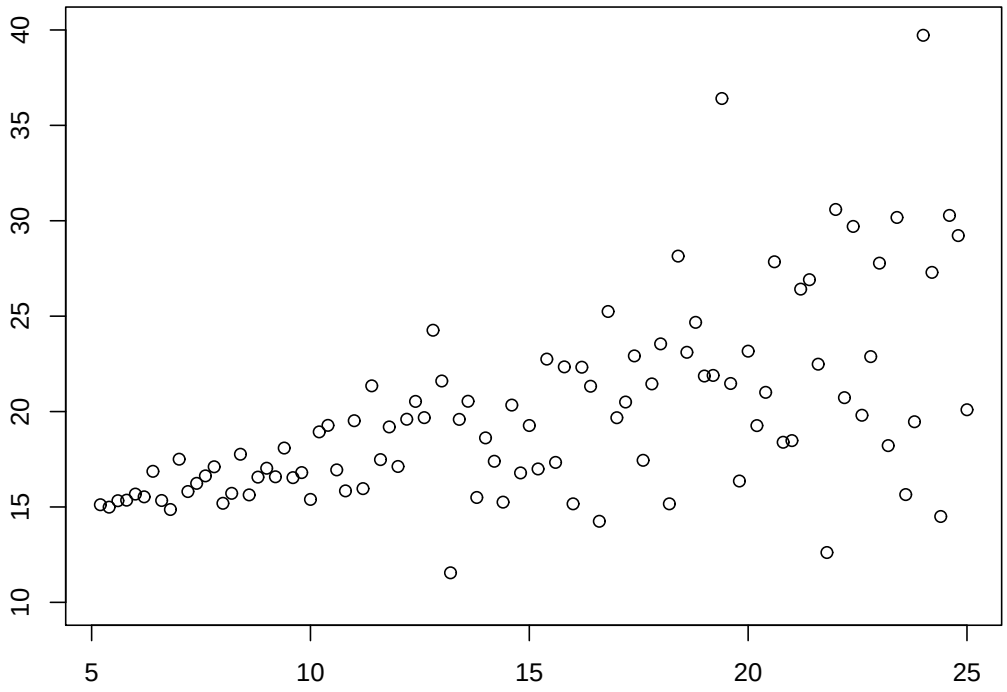
Prenez par exemple les 7 diagrammes de dispersion qui apparaissent à la section 6.9.1. Il est clair que l'hypothèse de linéarité est raisonnable dans les 6 premiers graphes mais pas dans le septième. Examinez maintenant le diagramme de dispersion de l'exemple de Pearson et Lee. Si on trace plusieurs petites bandes verticales de largeur 1 pouce et si pour chaque bande on détermine, à l'oeil, la moyenne des y pour les points appartenant à la bande, alors ces différentes moyennes formeront, grosso modo, une droite.

Si l'hypothèse de linéarité n'est pas satisfaite, on peut essayer une transformation de la variable Y (ou de la variable X). On essaie habituellement les transformations $1/y$, $\sqrt{y}$ et $\log(y)$, comme on a fait au chapitre sur l'analyse de la variance. On peut aussi essayer une modèle de régression non-linéaire. Avec le diagramme de dispersion qui apparaît à la fin de la section 6.9.1, il serait raisonnable d'essayer le modèle

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2.$$

La théorie présentée dans les sections précédentes peut être adaptée adéquatement.

Passons au point 2, l'homogénéité des variances conditionnelles. Notre premier outil est à nouveau le diagramme de dispersion. Tous les diagrammes de dispersion présentés jusqu'à maintenant étaient cohérents avec l'hypothèse d'homogénéité des variances. Voici un exemple d'un diagramme de dispersion pour lequel l'hypothèse d'homogénéité des variances ne tient pas. Dans ce diagramme de dispersion, on note que plus la variable X est grande, plus la variation en Y est grande. Pour remédier à la situation, il faudrait essayer une transformation.



Passons enfin au point 3, l'hypothèse de normalité. L'équation (6.1) nous donne

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i).$$

Si l'hypothèse de normalité est satisfaite, les *erreurs* $\epsilon_1, \epsilon_2, ..., \epsilon_n$ sont des variables aléatoires indépendantes avec distribution $N(0, \sigma^2)$. Pour vérifier l'hypothèse de normalité, il suffit d'analyser les résidus

$$e_i = y_i - (b_0 + b_1 x_i).$$

Si l'hypothèse de normalité est satisfaite, les résidus $e_1, e_2, ..., e_n$ devraient ressembler à un échantillon aléatoire de taille n issu d'une distribution normale. On peut donc vérifier l'hypothèse de normalité à l'aide des méthodes usuelles :

(a) Histogramme des résidus.

(b) Graphe quantile-quantile gaussien.

(c) Test de Shapiro et Wilk.

6.12 La régression linéaire multiple [3]

Dans le modèle de régression linéaire simple étudié dans le présent chapitre, la variable X est appelée la *variable explicative* et la variable Y est appelée la *variable réponse*. Dans certains problèmes, il y a une variable réponse et plusieurs variables explicatives. On peut alors utiliser le modèle de régression linéaire multiple.

Voici un exemple. Supposons qu'on veuille expliquer la taille de fiston à partir de la taille de papa et de la taille de maman. Posons

$$\begin{aligned} X &= \text{la taille du père} \\ Y &= \text{la taille de la mère} \\ Z &= \text{la taille du fils.} \end{aligned}$$

Les hypothèses du modèle de régression linéaire avec variable réponse Z et avec variables explicatives X et Y sont les suivantes.

(a) La moyenne de Z est une fonction linéaire de X et de Y. Autrement dit, on suppose que
$$\mathbb{E}[Z|(X,Y)=(x,y)] = \beta_0 + \beta_1 x + \beta_2 y.$$

(b) La variance conditionnelle $\mathbb{V}\text{ar}[Z|(X,Y)=(x,y)]$ est la même pour tout choix de (x,y). Cette variance conditionnelle est simplement dénotée σ^2.

(c) Pour chaque valeur possible (x,y) du couple de variables explicatives (X,Y), la distribution conditionnelle de la variable réponse Z sachant $(X,Y)=(x,y)$ est la loi normale (avec moyenne $\beta_0+\beta_1 x+\beta_2 y$ et avec variance σ^2).

Nous avons alors quatre paramètres à estimer : $\beta_0, \beta_1, \beta_2$ et σ^2.

Dans le formulation générale du modèle de régression linéaire multiple, on considère une variable réponse, qu'on note habituellement Y, et p variables explicatives, qu'on note habituellement $X_1, X_2, ..., X_p$. Les données sont alors de la forme

$$(x_{1,1}, x_{1,2}, ..., x_{1,p}, y_1), \quad (x_{2,1}, x_{2,2}, ..., x_{2,p}, y_2), ..., \quad (x_{n,1}, x_{n,2}, ..., x_{n,p}, y_n).$$

Les résultats des sections 6.3, 6.4, 6.7 et 6.8 peuvent être généralisés au cas de la régression linéaire multiple. Les calculs deviennent très compliqués. Pour s'en sortir, il faut avoir recours aux méthodes matricielles et il faut absolument utiliser un logiciel de statistique.

Pour en savoir plus sur la régression multiple, il est recommandé de suivre un cours de méthodes statistiques avancées.

6.13 Exercices

Numéro 1. Reprenez l'exemple 1 du chapitre 6 (les 60 plants de tomates) et calculez les intervalles de confiance suivants. Le fichier EXCEL est disponible sur le site web du cours.

(a) L'intervalle de confiance de niveau 90% pour σ.

[3] Cette section peut être omise, ou laissée en lecture, si on manque de temps.

(b) L'intervalle de confiance de niveau 90% pour β_0.

(c) L'intervalle de confiance de niveau 90% pour β_1.

(d) L'intervalle de confiance de niveau 90% pour $\beta_0 + \beta_1(5.4)$.

NUMÉRO 2. Reprenez l'exemple 2 du chapitre 6 (les 50 hirondelles) et calculez les erreurs types suivantes. Le fichier EXCEL est disponible sur le site web du cours.

(a) L'erreur type associée à l'estimation b_0.

(b) L'erreur type associée à l'estimation b_1.

(c) L'erreur type associée à l'estimation $\hat{\sigma}^2$.

NUMÉRO 3. Reprenez l'exemple 2 du chapitre 6 (les 50 hirondelles).

(a) Calculez une estimation pour l'étendue moyenne des hirondelles dont le poids est 53 grammes.

(b) Calculez l'erreur type associée à l'estimation obtenue en (a).

(c) Calculez un intervalle de confiance de niveau 95% pour l'étendue moyenne des hirondelles dont le poids est 53 grammes.

(d) Estimez l'écart-type de l'étendue des hirondelles dont le poids est 53 grammes.

(e) On obtient une 51^e hirondelle. Son poids est 42.5 grammes. Calculez un intervalle de prédiction de niveau 95% pour son étendue.

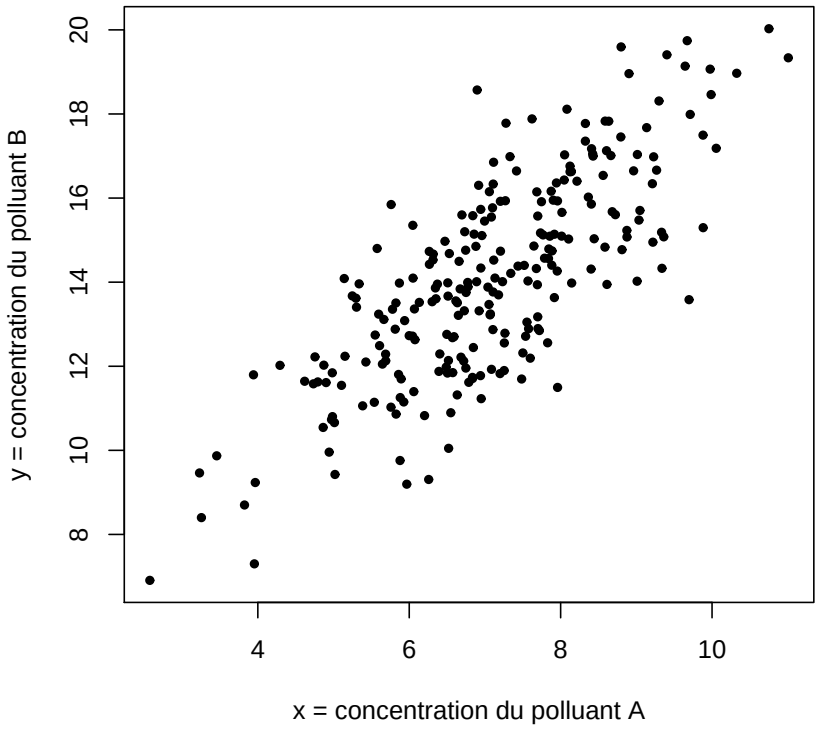
NUMÉRO 4. Les eaux du lac St-Pierre contiennent de grandes quantités de polluants industriels dont le polluant A et le polluant B. Ces deux polluants se retrouvent dans le foie des poissons du lac. On a capturé 250 perchaudes et on a mesuré, pour chaque perchaude, les variables

$$\begin{aligned} X &= \text{la concentration de polluant A (en ppm)} \\ Y &= \text{la concentration de polluant B (en ppm).} \end{aligned}$$

Voici quelques statistiques calculées à partir de ces observations :

$$\begin{array}{lll} \overline{x} = 7.083 \text{ ppm} & s_X = 1.467 \text{ ppm} & r = 0.775 \\ \overline{y} = 14.079 \text{ ppm} & s_Y = 2.425 \text{ ppm} & \end{array}$$

Voici le diagramme de dispersion :



(a) On capture une perchaude dans le lac St-Pierre. On s'attend à ce que la concentration de polluant B dans son foie soit environ _____ ppm,

plus ou moins environ _____ ppm.

(b) On capture une perchaude dans le lac St-Pierre. On mesure la concentration de polluant A dans son foie et on obtient $x = 9$ ppm. On s'attend à ce que la concentration de polluant B dans son foie soit environ _____ ppm,

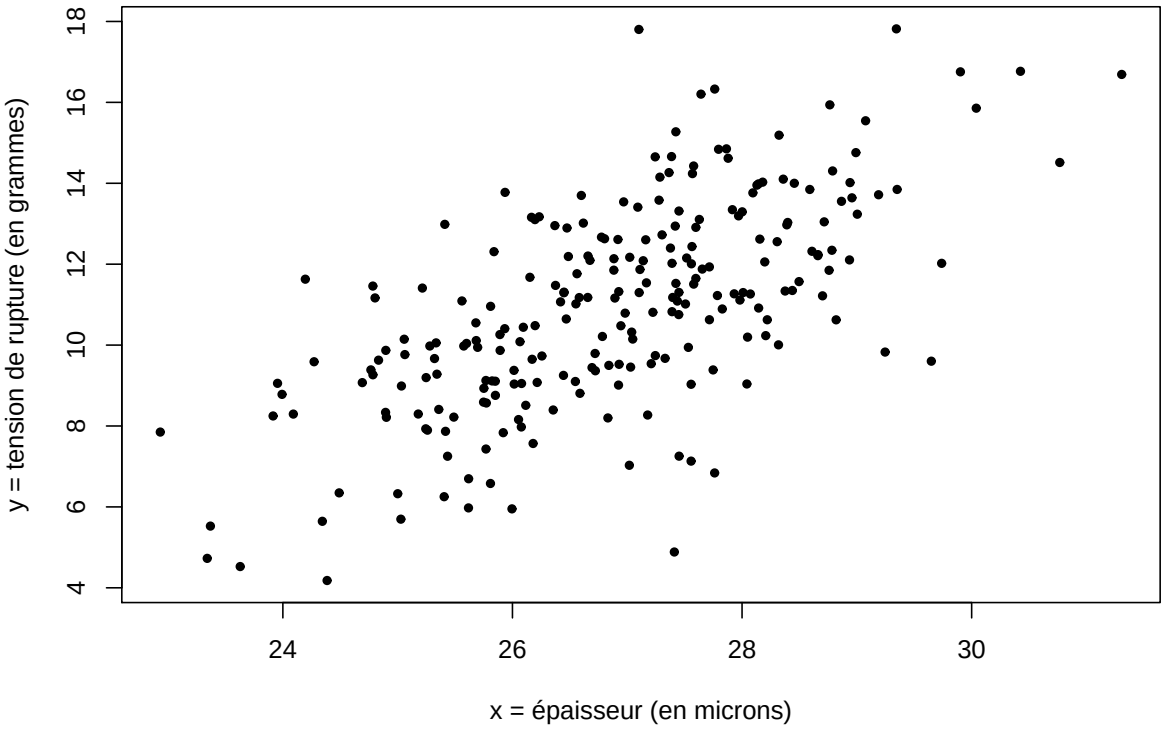
plus ou moins environ _____ ppm.

(c) La concentration de polluant B dans le foie varie beaucoup d'une perchaude à une autre. Il est raisonnable de dire qu'environ

☐ 20 ☐ 22.5 ☐ 40 ☐ 50 ☐ 60 ☐ 77.5 ☐ 80

pour cent de cette variation est expliquée par la concentration de polluant A dans le foie.

NUMÉRO 5. Nous avons mesuré l'épaisseur et la tension de rupture de 242 membranes de plastique. Voici le diagramme de dispersion :



(a) Le diagramme de dispersion suggère que le modèle de régression linéaire simple est un modèle approprié. Nous estimons les paramètres de ce modèle et nous obtenons les estimations suivantes :

Paramètre	Estimation	Erreur type
β_0	-20.3630 g	2.3506 g
β_1	1.1656 g/micron	0.0873 g/micron
σ^2	3.7745 g^2	0.3460 g^2

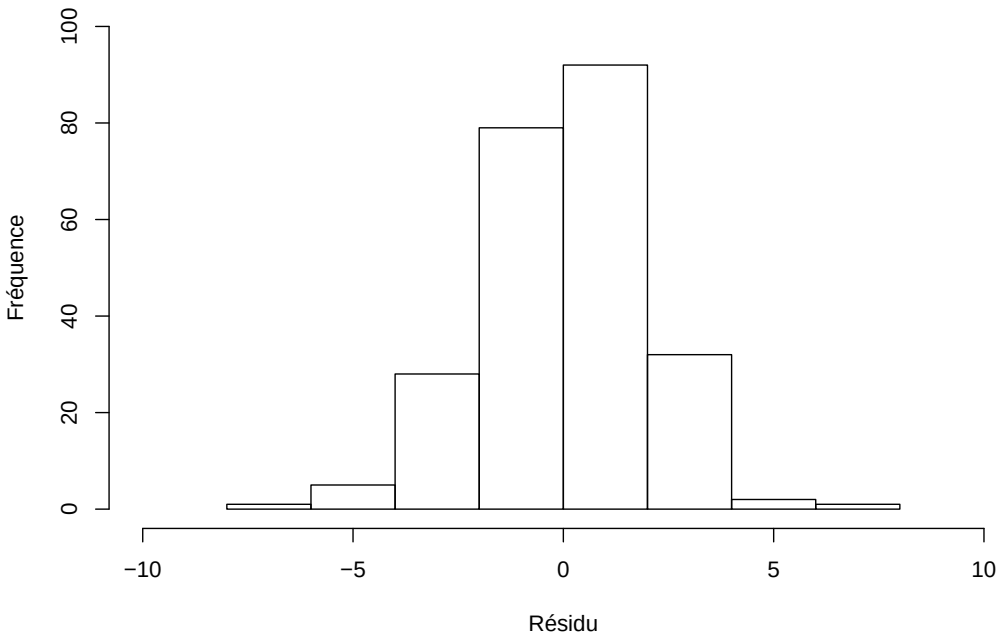
Obtenez un intervalle de confiance de niveau 95% pour le paramètre β_1.

(b) Nous avons également obtenu

$\overline{x} = 26.882$ micron $\quad s_X = 1.4393$ micron $\quad \overline{y} = 10.9692$ g $\quad s_Y = 2.5638$ g

Obtenez le coefficient de corrélation.

(c) Voici l'histogramme des résidus :



L'épaisseur de la membrane numéro 134 était 27.75 microns. Sa tension de rupture était 9.39 grammes. Quel est le résidu associé à la membrane numéro 134 ?

NUMÉRO 6. Voici les observations obtenues lors d'une expérience dont le but était de déterminer s'il y avait une association positive entre les variables X et Y. Le fichier EXCEL est disponible sur le site web du cours.

X	0.81	2.46	0.68	3.77	2.85	2.97	2.54	4.70	2.32	3.30
Y	5.56	5.69	4.64	5.01	5.76	5.16	5.05	5.58	4.84	5.54

X	1.70	4.13	1.06	3.47	2.39	2.63	1.26	0.76	3.37	0.55
Y	4.52	5.75	5.07	5.55	5.37	4.86	4.66	4.97	5.57	4.31

On veut donc tester

H_0 : Il n'y a pas d'association entre X et Y
H_1 : Il y a une association positive entre X et Y.

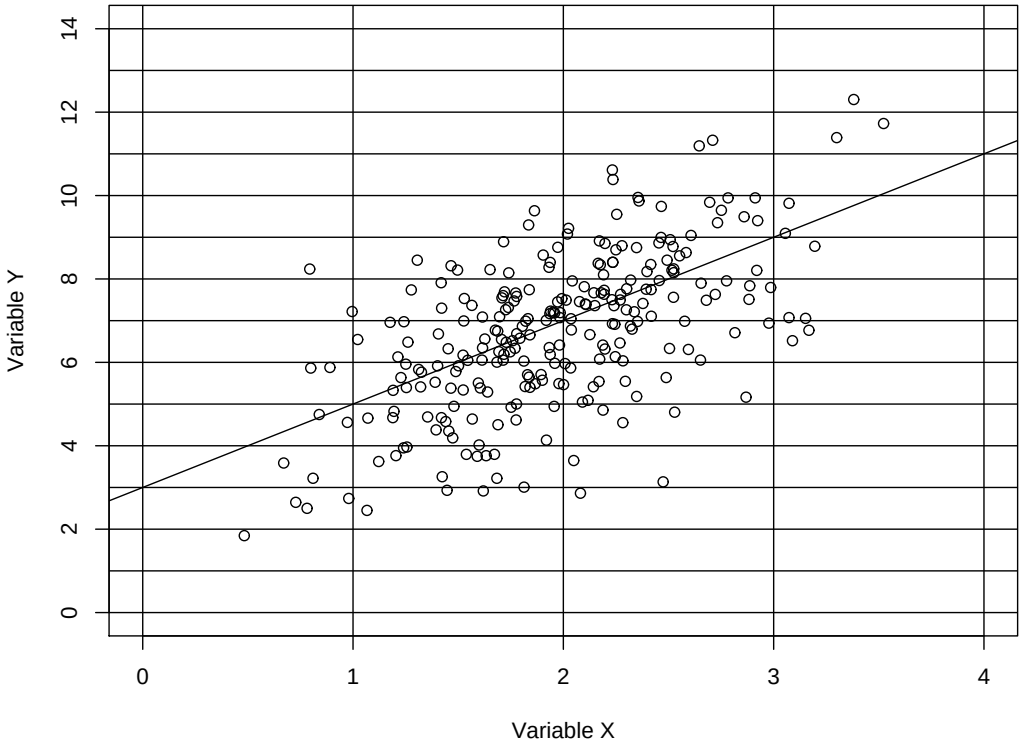
(a) Tracez le diagramme de dispersion.

(b) Calculez et tracez la droite de régression.

(c) Faites le test d'hypothèse en vous basant sur la statistique b_1.

(d) Faites le test d'hypothèse en vous basant sur la statistique r.

Numéro 7. Voici un diagramme de dispersion avec sa droite de régression.



L'équation de cette droite est $y = b_0 + b_1 x$. D'après le graphe, on a

$$b_0 = \text{\_\_\_\_\_} \qquad \text{et} \qquad b_1 = \text{\_\_\_\_\_}.$$

Numéro 8. À l'hiver 2006, 185 étudiants ont fait les deux premiers examens du cours STT-10400. Le diagramme de dispersion suggère que les hypothèses du modèle de régression linéaire simple sont satisfaites. Le coefficient de corrélation est 0.72. À l'examen 1, la note de Jean-Louis était deux écarts-types en haut de la moyenne. À l'examen 2, on s'attend à ce qu'il soit combien d'écarts-types en haut de la moyenne ?

Numéro 9. Reprenez les données du numéro 6. Calculez les 20 résidus. Dessinez l'histogramme de ces 20 résidus. Dessinez le graphe quantile-quantile gaussien. Faites le test de Shapiro et Wilk. Quelles sont vos conclusions ?

Chapitre 7

Tableaux de fréquences et tests du khi-deux

7.1 Test d'adéquation pour une variable discrète

Imaginez une population avec une variable statistique possédant un nombre fini de valeurs possibles. Il peut s'agir d'une variable quantitative discrète, d'une variable qualitative ordinale ou d'une variable qualitative nominale. Supposons que cette variable possède J valeurs possibles, disons les valeurs $v_1, v_2, ..., v_J$. Les proportions associées à ces valeurs possibles seront dénotées $p_1, p_2, ..., p_J$. Autrement dit, si on choisit un individu au hasard à partir de cette population et si Y dénote la variable d'intérêt, alors on a

$$\mathbb{P}[Y = v_1] = p_1, \qquad \mathbb{P}[Y = v_2] = p_2, \qquad \mathbb{P}[Y = v_3] = p_3, \quad \cdots, \quad \mathbb{P}[Y = v_J] = p_J.$$

Exemple 1. On considère l'ensemble des maisons unifamiliales dans la région de Québec. La variable d'intérêt est le nombre de chambres à coucher. Il s'agit d'une variable quantitative discrète. L'ensemble des valeurs possibles est peut-être l'ensemble $\{1, 2, 3, 4, 5, 6\}$. Les probabilités associées à ces valeurs possibles sont peut-être les suivantes :

$$\begin{array}{lll} \mathbb{P}[Y = 1] = 0.10 & \mathbb{P}[Y = 2] = 0.25 & \mathbb{P}[Y = 3] = 0.30 \\ \mathbb{P}[Y = 4] = 0.20 & \mathbb{P}[Y = 5] = 0.10 & \mathbb{P}[Y = 6] = 0.05 \end{array}$$

Exemple 2. On considère la population des plus de 30 ans au Québec. La variable d'intérêt est le niveau d'éducation atteint. Les valeurs possibles de cette variable, et les probabilités associées à ces valeurs possibles, sont peut-être les suivantes :

Valeur	Probabilité
$v_1 =$ Pas de secondaire	0.05
$v_2 =$ Diplôme secondaire	0.25
$v_3 =$ Diplôme de cégep	0.50
$v_4 =$ Baccalauréat	0.15
$v_5 =$ Maîtrise ou plus	0.05

Il s'agit ici d'une variable qualitative ordinale.

EXEMPLE 3. On considère la population de tous les élèves de la maternelle dans la province de Québec. La variable d'intérêt est la saveur de crème glacée préférée de l'enfant. Il y a trois valeurs possibles : Chocolat, Fraise, Vanille. Les probabilités associées à ces valeurs possibles sont peut-être les suivantes :

Valeur	Probabilité
$v_1 =$ Chocolat	0.50
$v_2 =$ Fraise	0.25
$v_3 =$ Vanille	0.25

Il s'agit ici d'une variable qualitative nominale.

Il arrive qu'on veuille tester

$$H_0 : (p_1, p_2, ..., p_J) = (p_1^*, p_2^*, ..., p_J^*)$$
$$H_1 : (p_1, p_2, ..., p_J) \neq (p_1^*, p_2^*, ..., p_J^*)$$

On parle alors de *test d'adéquation* ou de *test d'ajustement* car on cherche à déterminer si la distribution $(p_1^*, p_2^*, ..., p_J^*)$ est adéquate pour notre variable d'intérêt. En d'autres mots, on cherche à déterminer si la distribution $(p_1^*, p_2^*, ..., p_J^*)$ est bien ajusté à nos données. Ici le vecteur de proportions $(p_1^*, p_2^*, ...p_J^*)$ représente une distribution hypothétique pour la variable d'intérêt. Les valeurs $p_1^*, p_2^*, ...p_J^*$ sont spécifiées. Bien sûr, ce sont des nombres satisfaisant les deux conditions suivantes :

(a) $0 \leq p_j \leq 1$ pour $j = 1, 2, ..., J$.

(b) $\sum_{j=1}^{J} p_j = 1$.

Afin d'arriver à une décision, nous allons obtenir un échantillon aléatoire de taille n, disons $Y_1, Y_2, ..., Y_n$, à partir de notre population. Nous allons ensuite calculer nos *fréquences observées* :

$$\begin{aligned} O_1 &= \text{le nombre de fois qu'on obtient la valeur } v_1 \\ O_2 &= \text{le nombre de fois qu'on obtient la valeur } v_2 \\ &\vdots \\ O_J &= \text{le nombre de fois qu'on obtient la valeur } v_J \end{aligned}$$

Nous allons ensuite comparer ces fréquences observées avec les *fréquences espérées sous* H_0, c'est-à-dire avec les fréquences auxquelles on devrait s'attendre si l'hypothèse nulle était vraie :

$$\begin{aligned} E_1 &= \mathbb{E}_{H_0}[O_1] = np_1^* \\ E_2 &= \mathbb{E}_{H_0}[O_2] = np_2^* \\ &\vdots \\ E_J &= \mathbb{E}_{H_0}[O_J] = np_J^* \end{aligned}$$

Explication : les espérances ci-dessus viennent du fait que sous H_0 on a

$$O_j \sim \text{ binomiale}(n, p_j^*).$$

Enfin, nous allons mesurer la *distance* entre nos fréquences observées et nos fréquences espérées sous H_0. Cette distance se mesure de la façon suivante :

$$\text{Distance} \quad = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} \quad = \sum_{j=1}^{J} \frac{(O_j - np_j^*)^2}{np_j^*}$$

Notre règle de décision sera donc de la forme suivante :

$$\text{On rejette } H_0 \text{ si } \sum_{j=1}^{J} \frac{(O_j - np_j^*)^2}{np_j^*} \text{ est trop grand.}$$

En statistique mathématique, on peut montrer que si H_0 est vraie et si n est suffisamment grand, alors notre statistique de test suit à peu près la loi du khi-deux avec $J-1$ degrés de liberté, c'est-à-dire

$$\text{si } H_0 \text{ est vraie, alors } \sum_{j=1}^{J} \frac{(O_j - np_j^*)^2}{np_j^*} \approx \chi^2_{J-1}. \tag{7.1}$$

Notre règle de décision peut donc s'écrire sous la forme suivante :

$$\text{On rejette } H_0 \text{ si } \sum_{j=1}^{J} \frac{(O_j - np_j^*)^2}{np_j^*} \geq \chi^2_{J-1,\alpha}. \tag{7.2}$$

Que veut-on dire par n suffisamment grand ? Il n'y a pas de réponse parfaite à cette question. Plus n est grand et plus l'approximation (7.1) est bonne. L'important c'est que les fréquences espérées $np_1^*, np_2^*, ..., np_J^*$ soient toutes suffisamment grandes. La plupart des statisticiens considèrent que l'approximation (7.1) est bonne si les fréquences espérées np_j^* sont toutes plus grandes ou égales à 5.

La statistique de test utilisée à l'équation (7.2) est souvent appelée le khi-deux de Pearson.

Retour à l'exemple 2 :

Supposons que la distribution de probabilités présentée à l'exemple 2 soit un bon modèle pour le niveau d'éducation des plus de 30 ans au Québec. On se demande si cette même distribution est un bon modèle pour le niveau d'éducation des plus de 30 ans dans la ville de Québec. On veut donc tester

$$H_0 : (p_1, p_2, p_3, p_4, p_5) = (0.05, 0.25, 0.50, 0.15, 0.05)$$
$$H_1 : (p_1, p_2, p_3, p_4, p_5) \neq (0.05, 0.25, 0.50, 0.15, 0.05)$$

où p_1, p_2, p_3, p_4 et p_5 dénotent les proportions de chacune des 5 classes pour la ville de Québec. On obtient un échantillon aléatoire de 300 adultes de plus de 30 ans vivant à Québec. Parmi ces 300 personnes, il y en a 11 qui n'ont pas terminé le secondaire, 62 qui ont terminé le secondaire mais ne sont pas allées plus loin, 142 dont le plus haut diplôme est un DEC, 64 dont le plus haut diplôme est un baccalauréat et 21 qui ont une maîtrise ou un doctorat. Le tableau suivant résume nos données et nos calculs :

Valeur	O_j	E_j	$\frac{(O_j-E_j)^2}{E_j}$
Pas de secondaire	11	15	1.07
Diplôme de secondaire	62	75	2.25
Diplôme de cégep	142	150	0.43
Baccalauréat	64	45	8.02
Maîtrise ou plus	21	15	2.40
Total	300	300	14.17

Les fréquences espérées sous H_0 ont été calculées à l'aide de la formule $E_j = np_j^*$. Donc, $E_1 = np_1^* = 300 \times 0.05 = 15$, $E_2 = np_2^* = 300 \times 0.25 = 75$, etc. La valeur observée de notre statistique de test est donc 14.17. On doit comparer cette valeur avec les quantiles de la loi du khi-deux avec 4 degrés de liberté. La table de la loi du khi-deux nous permet de conclure que le *p-value* est entre 0.005 et 0.01. Le logiciel R nous donne un *p-value* de 0.0068. On rejette donc H_0 !

REMARQUE : Dans cet exemple, les fréquences observées suggèrent que les gens de la ville de Québec ont tendance à être plus instruits que ceux de l'ensemble du Québec. Ce n'est pas surprenant.

7.2 Test d'homogénéité de I populations

On s'intéresse ici à I populations. La variable d'intérêt Y est une variable discrète, la même variable pour chacune des I populations, avec J valeurs possibles, disons les valeurs $v_1, v_2, ..., v_J$. On écrit $p_{i,j}$ pour dénoter la proportion d'individus dans la population i pour lesquels la variable Y prend la valeur v_j. On a donc :

$$\begin{aligned}
&\text{Distribution de la variable } Y \text{ dans la population 1 :} && (p_{1,1}, p_{1,2}, ..., p_{1,J}) \\
&\text{Distribution de la variable } Y \text{ dans la population 2 :} && (p_{2,1}, p_{2,2}, ..., p_{2,J}) \\
&\text{Distribution de la variable } Y \text{ dans la population 3 :} && (p_{3,1}, p_{3,2}, ..., p_{3,J}) \\
&\qquad\qquad\vdots && \qquad\vdots \\
&\text{Distribution de la variable } Y \text{ dans la population I :} && (p_{I,1}, p_{I,2}, ..., p_{I,J})
\end{aligned}$$

On veut tester

H_0 : Les I populations sont homogènes

H_1 : Les I populations ne sont pas homogènes

Autrement dit, on veut tester

H_0 : Les I populations ont la même distribution (pour la variable Y)

H_1 : Les I populations n'ont pas la même distribution (pour la variable Y)

Autrement dit, on veut tester

$H_0 : (p_{1,1}, p_{1,2}, ..., p_{1,J}) = (p_{2,1}, p_{2,2}, ..., p_{2,J}) = \cdots = (p_{I,1}, p_{I,2}, ..., p_{I,J})$

H_1 : Ces I distributions ne sont pas toutes égales

On obtient

- un échantillon aléatoire de taille n_1 à partir de la population 1,
- un échantillon aléatoire de taille n_2 à partir de la population 2,

  $\vdots$
- un échantillon aléatoire de taille n_I à partir de la population I.

On a en tout $n = n_1 + n_2 + \cdots + n_I$ observations. Pour chaque couple (i, j), on calcule la fréquence observée O_{ij}, c'est-à-dire

$$O_{ij} = \text{ le nombre de } v_j \text{ parmi les } n_i \text{ observations issues de la population } i.$$

On résume nos données à l'aide d'un *tableau de fréquences*. Par exemple, si notre variable possède $J = 4$ valeurs possibles et si on compare $I = 3$ populations, alors notre tableau de fréquences aura l'allure suivante :

Valeurs de la variable Y

Population	v_1	v_2	v_3	v_4	Taille d'échantillon
1	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,4}$	n_1
2	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,4}$	n_2
3	$O_{3,1}$	$O_{3,2}$	$O_{3,3}$	$O_{3,4}$	n_3
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot 3}$	$O_{\cdot 4}$	n

On procède comme à la section 7.1 : notre règle de décision sera basée sur la statistique

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

avec

$$E_{ij} = \mathbb{E}_{H_0}[O_{ij}].$$

Comment calcule-t-on les E_{ij} ? L'hypothèse nulle H_0 nous dit que les I populations ont toutes la même distribution mais ne spécifie pas cette distribution. Dénotons cette distribution par $(p_1, p_2, ..., p_J)$. Sous H_0 on a

$$O_{ij} \sim \text{ binomiale}(n_i, p_j)$$

et donc

$$E_{ij} = \mathbb{E}_{H_0}[O_{ij}] = n_i p_j$$

On connaît n_i mais on ne connaît pas p_j. On utilise l'estimation suivante :

$$\hat{p}_j = \frac{O_{\cdot j}}{n}$$

où $O_{\cdot j}$ dénote le nombre total de v_j parmi les I échantillons, c'est-à-dire $O_{\cdot j} = \sum_{i=1}^{I} O_{i,j}$. C'est bien naturel : On a en tout n observations et parmi ces n observations, il y en a $O_{\cdot j}$ qui prennent la valeur v_j. On a donc

$$E_{ij} = \mathbb{E}_{H_0}[O_{ij}] \approx \hat{E}_{ij} = \hat{\mathbb{E}}_{H_0}[O_{ij}] = n_i \hat{p}_j = n_i \frac{O_{\cdot j}}{n} = \frac{n_i O_{\cdot j}}{n}$$

Notez comment on peut faire ces calculs à partir du tableau de fréquences. Pour la cellule (i, j) ont fait le produit de la somme de la ligne i et de la somme de la colonne j, puis on divise par le grand total n. Cette méthode est parfois appelée la *méthode du produit croisé*. En pratique, on laisse tomber les petits *chapeaux* et on écrit simplement $E_{ij} = \frac{n_i O_{\cdot j}}{n}$. Notre statistique de test sera donc

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - \frac{n_i O_{\cdot j}}{n}\right)^2}{\frac{n_i O_{\cdot j}}{n}}.$$

En statistique mathématique, on montre que sous H_0 cette statistique de test suit la loi du khi-deux avec $(I-1)(J-1)$ degrés de liberté. Notre règle de décision au seuil α est donc

$$\text{on rejette } H_0 \text{ si } \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - \frac{n_i O_{\cdot j}}{n}\right)^2}{\frac{n_i O_{\cdot j}}{n}} \geq \chi^2_{(I-1)(J-1),\alpha}. \tag{7.3}$$

L'étudiant devrait comparer l'équation (7.3) avec l'équation (7.2).

Exemple 4. Parmi les 332 étudiants qui ont suivi le cours STT-10400 à l'automne 1998, il y en a 113 qui étaient dans la section du professeur Aubin, 98 qui étaient dans la section du professeur Beaudoin et 121 qui étaient dans la section du professeur Claveau. Voici les distributions de cotes pour chacune de ces trois sections :

	A	B	C	D	E	$Total$
Aubin	14	29	40	21	9	113
Beaudoin	9	24	26	25	14	98
Claveau	16	30	39	26	10	121

Ces trois distributions de cotes ne sont pas identiques. Mais la différence est-elle significative ? Si on imagine que

- les 113 étudiants de Aubin constituent un échantillon aléatoire issu de la population potentielle de tous les étudiants (passé, présent et futur) prenant le cours avec Aubin,
- les 98 étudiants de Beaudoin constituent un échantillon aléatoire issu de la population potentielle de tous les étudiants (passé, présent et futur) prenant le cours avec Beaudoin,
- les 121 étudiants de Claveau constituent un échantillon aléatoire issu de la population potentielle de tous les étudiants (passé, présent et futur) prenant le cours avec Claveau,

alors on peut utiliser le test du khi-deux de Pearson présenté ci-dessus. D'abord on réécrit le tableau ci-dessus en y ajoutant les sommes de lignes (c'est-à-dire les n_i) et les sommes de colonnes (c'est-à-dire les $O_{\cdot j}$). On obtient le tableau suivant :

Professeur	A	B	C	D	E	Total
Aubin	14	29	40	21	9	113
Beaudoin	9	24	26	25	14	98
Claveau	16	30	39	26	10	121
Total	39	83	105	72	33	332

Ensuite on calcule nos fréquences espérées. Pour la cellule $(2, 4)$, on obtient

$$E_{2,4} = \frac{n_2 O_{\cdot 4}}{n} = \frac{98 \times 72}{332} = 21.253.$$

Avec le présent tableau, on a $3 \times 5 = 15$ fréquences espérées à calculer ! On fait tous ces calculs et on écrit nos E_{ij} entre parenthèses, en dessous des O_{ij} correspondants :

Professeur	A	B	C	D	E	Total
Aubin	14 (13.274)	29 (28.250)	40 (35.738)	21 (24.506)	9 (11.232)	113
Beaudoin	9 (11.512)	24 (24.500)	26 (30.994)	25 (21.253)	14 (9.741)	98
Claveau	16 (14.214)	30 (30.250)	39 (38.268)	26 (26.241)	10 (12.027)	121
Total	39	83	105	72	33	332

Notre statistique de test est donc

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(O_{ij} - \frac{n_i O_{\cdot j}}{n}\right)^2}{\frac{n_i O_{\cdot j}}{n}} &= \sum_{i=1}^{3}\sum_{j=1}^{5} \frac{\left(O_{ij} - \frac{n_i O_{\cdot j}}{332}\right)^2}{\frac{n_i O_{\cdot j}}{332}} \\
&= \frac{(14 - 13.274)^2}{13.274} + \frac{(29 - 28.250)^2}{28.250} + \cdots + \frac{(10 - 12.027)^2}{12.027} \\
&= 5.983.
\end{aligned}$$

Le nombre de degrés de liberté est

$$(I - 1)(J - 1) = (3 - 1)(5 - 1) = 8.$$

Donc, sous H_0 on s'attend à ce que la statistique de test soit environ 8, plus ou moins environ 4. On a obtenu la valeur 5.983. Il n'y a pas lieu de rejeter H_0.

Exemple 5. On considère un certain projet de loi. On a obtenu un échantillon aléatoire de 604 hommes. Parmi eux, il y en a 329 qui sont en faveur de ce projet de loi et 275 qui sont contre. On a aussi obtenu un échantillon aléatoire de 510 femmes. Parmi elles, il y en a 319 qui sont en faveur de ce projet de loi et 191 qui sont contre. Le tableau suivant résume nos données :

	Pour	Contre	
Homme	329	275	604
Femme	319	191	510
	648	466	1114

On s'intéresse aux distributions des pour et des contre chez les hommes et chez les femmes. On veut tester

H_0 : les deux distributions sont identiques
H_1 : les deux distributions ne sont pas identiques

À l'aide de la méthode du produit croisé, on calcule les fréquences espérées sous H_0 :

$$\begin{aligned}
\text{Homme-pour} &= E_{hp} = \frac{604 \times 648}{1114} = 351.34 \\
\text{Homme-contre} &= E_{hc} = \frac{604 \times 466}{1114} = 252.66 \\
\text{Femme-pour} &= E_{fp} = \frac{510 \times 648}{1114} = 296.66 \\
\text{Femme-contre} &= E_{fc} = \frac{510 \times 466}{1114} = 213.34
\end{aligned}$$

Voici donc le tableau des fréquences espérées (les E_{ij}) :

	Pour	Contre	
Homme	351.34	252.66	604
Femme	296.66	213.34	510
	648	466	1114

On calcule ensuite notre khi-deux de Pearson :

$$\frac{(329-351.34)^2}{351.34} + \frac{(275-252.66)^2}{252.66} + \frac{(319-296.66)^2}{296.66} + \frac{(191-213.34)^2}{213.34} = 7.417.$$

Le nombre de degrés de liberté est égal à

$$(I-1) \times (J-1) = (2-1) \times (2-1) = 1.$$

Notre *p-value* est donc la surface à droite de 7.417 sous la densité de la loi du khi-deux à 1 degré de liberté. Cette surface est égale à 0.0065. C'est très petit. C'est inférieur à 1%. On rejette H_0.

REMARQUE. Dans l'exemple précédent, nous aurions pu utiliser la technique de la section 4.6. Puisque la variable est dichotomique, les hypothèses peuvent s'écrire sous la forme suivante :

$$\begin{aligned} H_0 : &\quad p_H = p_F \\ H_1 : &\quad p_H \neq p_F \end{aligned}$$

où p_H et p_F dénotent, respectivement, la proportion d'hommes en faveur du projet de loi et la proportion de femmes en faveur du projet de loi.

EXERCICE. Faites le test suggéré ci-dessus. Quelles sont vos conclusions ?

7.3 Test d'indépendance de deux variables discrètes

Imaginez une population avec deux variables d'intérêt, disons la variable X et la variable Y. La variable X possède I valeurs possibles, disons les valeurs $w_1, w_2, ..., w_I$. La variable Y possède J valeurs possibles, disons les valeurs $v_1, v_2, ..., v_J$. On se demande si ces deux variables sont indépendantes l'une de l'autre. On veut tester

$$\begin{aligned} H_0 : &\quad \text{les variables } X \text{ et } Y \text{ sont indépendantes} \\ H_1 : &\quad \text{les variables } X \text{ et } Y \text{ ne sont pas indépendantes} \end{aligned}$$

On obtient un échantillon aléatoire de taille n à partir de cette population. Voici nos observations :

$$(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3),\ (x_4, y_4),\ ...,\ (x_n, y_n)$$

Pour chaque valeur $1 \leq i \leq I$ et chaque valeur $1 \leq j \leq J$, on pose

$$\begin{aligned} O_{ij} &= \text{le nombre de } (w_i, v_j) \text{ parmi nos } n \text{ observations.} \\ O_{i\cdot} &= \sum_{j=1}^{J} O_{ij} \\ O_{\cdot j} &= \sum_{i=1}^{I} O_{ij} \end{aligned}$$

Les O_{ij} sont les fréquences observées. On présente ces fréquences observées sous forme d'un tableau. Par exemple, si $I = 3$ et $J = 4$, on obtient le tableau suivant :

Valeurs de la variable Y

Valeurs de la variable X	v_1	v_2	v_3	v_4	Total
w_1	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,4}$	$O_{1\cdot}$
w_2	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,4}$	$O_{2\cdot}$
w_3	$O_{3,1}$	$O_{3,2}$	$O_{3,3}$	$O_{3,4}$	$O_{3\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot 3}$	$O_{\cdot 4}$	n

On procède comme à la section 7.1 et à la section 7.2 : notre règle de décision sera basée sur la statistique

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

avec

$$E_{ij}=\mathbb{E}_{H_0}[O_{ij}]=np_{ij}$$

avec $p_{ij}=\mathbb{P}_{H_0}[(X,Y)=(w_i,v_j)]$. L'hypothèse nulle H_0 nous dit que les variables X et Y sont indépendantes. Cela signifie que pour tout i et tout j on a

$$p_{ij}=p_{i\cdot}\ p_{\cdot j}$$

où $p_{i\cdot}=\mathbb{P}_{H_0}[X=w_i]$ et $p_{\cdot j}=\mathbb{P}_{H_0}[Y=v_j]$. Pour estimer $p_{i\cdot}$, on utilise $\hat{p}_{i\cdot}=O_{i\cdot}/n$ et pour estimer $p_{\cdot j}$, on utilise $\hat{p}_{\cdot j}=O_{\cdot j}/n$. On obtient donc

$$\hat{E}_{ij}=\hat{\mathbb{E}}_{H_0}[O_{ij}]=n\hat{p}_{ij}=n\hat{p}_{i\cdot}\hat{p}_{\cdot j}=n\frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n}=\frac{O_{i\cdot}O_{\cdot j}}{n}$$

Généralement, on omet le petit chapeau et on écrit simplement E_{ij} plutôt que $\hat{E}_{ij}$. On calcule donc nos fréquences espérées à l'aide de la méthode du produit croisé, comme à la section 7.2. Notre statistique de test sera donc la statistique

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij}-\frac{O_{i\cdot}O_{\cdot j}}{n}\right)^2}{\frac{O_{i\cdot}O_{\cdot j}}{n}}.$$

En statistique mathématique, on montre que sous H_0 cette statistique de test suit la loi du khi-deux avec $(I-1)(J-1)$ degrés de liberté. Notre règle de décision au seuil α est donc

$$\text{on rejette } H_0 \text{ si } \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij}-\frac{O_{i\cdot}O_{\cdot j}}{n}\right)^2}{\frac{O_{i\cdot}O_{\cdot j}}{n}}\geq\chi^2_{(I-1)(J-1),\alpha}. \tag{7.4}$$

L'étudiant devrait comparer l'équation (7.4) avec les équations (7.2) et (7.3).

Exemple 6. En 1976, on a obtenu un échantillon aléatoire de 100 couples mariés dans la région de Québec. Parmi les 100 couples, il y en avait 19 pour lesquels homme et femme étaient tous les deux des fumeurs, 53 pour lesquels ni l'homme ni la femme n'était fumeur, 20 pour lesquels seul l'homme était fumeur et 8 pour lesquels seule la femme était fumeuse. Le tableau suivant résume les données :

		Femme		
		Fumeuse	Non-fumeuse	Total
Homme	Fumeur	19	20	39
	Non-fumeur	8	53	61
	Total	27	73	100

Analysez ces données.

SOLUTION. Nous avons 100 couples mariés. Les variables d'intérêt sont les variables

$$\begin{aligned} X &= \begin{cases} F & \text{si l'homme est fumeur} \\ N & \text{si l'homme est non-fumeur} \end{cases} \\ Y &= \begin{cases} F & \text{si la femme est fumeuse} \\ N & \text{si la femme est non-fumeuse} \end{cases} \end{aligned}$$

Nos données brutes ont l'allure suivante :

$$(F,N),\ (N,N),\ (N,N),\ (F,F),\ (F,N),\ (N,F),\ (N,N),\ ...,\ (F,F).$$

La valeur (F,F) apparaît 19 fois, la valeur (F,N) apparaît 20 fois, la valeur (N,F) apparaît 8 fois et la valeur (N,N) apparaît 53 fois. On veut tester

$$\begin{aligned} H_0 &: \quad \text{les variables } X \text{ et } Y \text{ sont indépendantes} \\ H_1 &: \quad \text{les variables } X \text{ et } Y \text{ ne sont pas indépendantes} \end{aligned}$$

Nos fréquences espérées sous H_0 sont

$$\begin{aligned} E_{FF} &= \frac{27 \times 39}{100} = 10.53 \qquad E_{FN} = \frac{73 \times 39}{100} = 28.47 \\ E_{NF} &= \frac{27 \times 61}{100} = 16.47 \qquad E_{NN} = \frac{73 \times 61}{100} = 44.53 \end{aligned}$$

La valeur observée de notre statistique de test est donc

$$\frac{(19-10.53)^2}{10.53} + \frac{(20-28.47)^2}{28.47} + \frac{(8-16.47)^2}{16.47} + \frac{(53-44.53)^2}{44.53} = 15.230.$$

Le nombre de degrés de liberté est $(I-1)(J-1) = (2-1)(2-1) = 1$. Le *p-value* est extrêmement petit. C'est la surface à droite de 15.230 sous la densité de la loi du khi-deux avec 1 degré de liberté. Selon la table, c'est plus petit que 0.005. Selon le logiciel R, c'est environ 0.0001. On rejette H_0 (et on conclut que les semblables s'attirent).

7.4 Quelques remarques [1]

7.4.1 Le pourquoi du khi-deux

Dans chacune des trois sections précédentes, nous obtenons une statistique de test dont la distribution, sous H_0, est la loi du khi-deux. Voici un argument qui explique en partie pourquoi on arrive à cette loi du khi-deux. Pour fixer les idées, considérons le scénario de la section 7.1. Si H_0 est vraie, alors on obtient

$$\begin{aligned} O_1 &\sim \text{binomiale}(n, p_1^*) \\ O_2 &\sim \text{binomiale}(n, p_2^*) \\ &\vdots \\ O_J &\sim \text{binomiale}(n, p_J^*) \end{aligned}$$

[1] On peut omettre cette section, ou la laisser en lecture, si on manque de temps.

D'après le théorème limite central (Section 2.6.6), on obtient

$$\begin{aligned}
\frac{O_1 - np_1^*}{\sqrt{np_1^*(1-p_1^*)}} &\approx N(0,1) \\
\frac{O_2 - np_2^*}{\sqrt{np_2^*(1-p_2^*)}} &\approx N(0,1) \\
&\vdots \\
\frac{O_J - np_J^*}{\sqrt{np_J^*(1-p_J^*)}} &\approx N(0,1)
\end{aligned}$$

Donc, si les variables aléatoires $O_1, O_2, ..., O_J$ étaient indépendantes, alors le théorème 3.3 de la section 3.6 nous donnerait

$$\sum_{j=1}^{J} \frac{(O_j - np_j^*)^2}{np_j^*(1-p_j^*)} \approx \chi^2_J. \tag{7.5}$$

Or les O_j ne sont pas indépendantes puisque $O_1+O_2+\cdots+O_J = n$. Néanmoins, l'équation (7.5) n'est pas trop loin de la vérité. Avec des arguments mathématiques plus poussés, on arrive à démontrer l'équation (7.1), c'est-à-dire

$$\sum_{j=1}^{J} \frac{(O_j - np_j^*)^2}{np_j^*} \approx \chi^2_{J-1}. \tag{7.6}$$

7.4.2 Test d'homogénéité ou test d'indépendance ?

Le test d'homogénéité de la section 7.2 et le test d'indépendance de la section 7.3 sont, d'un point de vue pratique, essentiellement identiques. Voici les principales différences :

1. Pour le test d'homogénéité, on considère I populations à partir desquelles on tire des échantillons de tailles n_1, n_2,..., n_I. Ces tailles d'échantillons sont fixées d'avance. Autrement dit, les sommes-lignes n_i de notre tableau de fréquences sont fixées d'avance. Pour le test d'indépendance, on considère une seule population. On obtient un échantillon aléatoire de taille n à partir de cette population. Avec chaque tirage on observe une valeur X et une valeur Y. Les sommes-lignes $O_{i\cdot}$ ne sont pas fixées d'avance.
2. Les p_{ij} du test d'homogénéité satisfont les conditions suivantes :

$$\sum_{j=1}^{J} p_{1j} = 1, \quad \sum_{j=1}^{J} p_{2j} = 1, \quad \sum_{j=1}^{J} p_{3j} = 1, \quad \cdots, \quad \sum_{j=1}^{J} p_{Ij} = 1.$$

   Les p_{ij} du test d'indépendance satisfont la condition suivante :

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1.$$

7.4.3 Test d'adéquation avec un paramètre à estimer

Avec le test d'adéquation de la section 7.1, l'hypothèse nulle spécifiait complètement la distribution de la variable Y. Dans certains problèmes il peut arriver que l'hypothèse nulle ne spécifie la distribution de la variable Y qu'à un paramètre près. La méthode présentée à la section 7.1 peut être adaptée pour traiter cette situation. Voici un exemple illustratif.

On considère une espèce animale dont la portée comprend toujours exactement deux petits. On s'intéresse à la variable

$$Y = \text{ le nombre de petits dans la portée qui sont porteurs d'un certain virus.}$$

Cette variable Y possède 3 valeurs possibles : 0, 1 et 2. La distribution de cette variable peut donc s'écrire sous la forme suivante : (p_0, p_1, p_2). Un biologiste soupçonne que la loi binomiale$(2, p)$ est un bon modèle pour la variable Y. Il arrive à cette conclusion en raisonnant de la façon suivante. Il imagine que p dénote la proportion de petits qui naissent avec le virus. Il se dit que chacun des deux petits d'une portée a donc probabilité p d'être porteur du virus. Le nombre de porteurs de virus dans une portée est donc le nombre de *succès* parmi deux répétitions indépendantes d'une expérience avec probabilité de succès égal à p, d'où la loi binomiale. On veut donc tester

$$\begin{aligned} H_0 : &\quad \text{la distribution de } Y \text{ est une loi binomiale} \\ H_1 : &\quad \text{la distribution de } Y \text{ n'est pas une loi binomiale} \end{aligned}$$

Autrement dit, on veut tester

$$\begin{aligned} H_0 : &\quad (p_0, p_1, p_2) \text{ est de la forme } \left((1-p)^2, 2p(1-p), p^2\right) \\ H_1 : &\quad (p_0, p_1, p_2) \text{ n'est pas de la forme } \left((1-p)^2, 2p(1-p), p^2\right) \end{aligned}$$

Parmi 64 portées observées par notre biologiste, il y a 34 portées avec aucun porteur de virus, 19 portées avec un seul petit qui porte le virus et 11 portées pour lesquelles les petits sont tous les deux porteurs du virus. Est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Quel est le *p-value* ?

Nous procédons comme à la section 7.1. Cette fois-ci les fréquences espérées E_j sont données par

$$\begin{aligned} E_0 &= np_0 = n \times (1-p)^2 \\ E_1 &= np_1 = n \times 2p(1-p) \\ E_2 &= np_2 = n \times p^2. \end{aligned}$$

Il faut estimer la proportion p. Ce p représente la proportion de petits qui sont porteurs du virus. On a observé en tout 128 petits (64 portées de 2 petits). Parmi ces 128 petits, il y en a $19 + (2 \times 11) = 41$ qui sont porteurs du virus. Notre estimation de p est donc $\hat{p} = 41/128 \approx 0.32$. On peut maintenant calculer nos fréquences espérées :

$$\begin{aligned} E_0 &= np_0 = n \times (1-\hat{p})^2 = 29.566 \\ E_1 &= np_1 = n \times 2\hat{p}(1-\hat{p}) = 27.867 \\ E_2 &= np_2 = n \times \hat{p}^2 = 6.566. \end{aligned}$$

Notre statistique du khi-deux est donc

$$\frac{(34-29.566)^2}{29.566}+\frac{(19-27.867)^2}{27.867}+\frac{(11-6.566)^2}{6.566}=6.48.$$

Selon la section 7.1, le nombre de degrés de liberté devrait être 3 - 1 = 2. Or pour calculer nos E_j, on a d'abord dû estimer le paramètre p à partir de nos données. Conséquence : on perd 1 degré de liberté ! Il faut donc comparer la valeur observée de notre statistique de test, 6.48, non pas avec la loi du khi-deux avec 2 degrés de liberté mais plutôt avec la loi du khi-deux avec 1 seul degré de liberté. D'après le logiciel R, le *p-value* est 0.011.

7.4.4 Degrés de liberté

Voici la règle générale pour calculer le nombre de degrés de liberté lorsqu'on fait un test du khi-deux à partir d'un tableau de fréquences :

$$\text{d.l. } = c-k-\ell \tag{7.7}$$

avec

c = le nombre de cellules dans notre tableau de fréquences

k = le nombre de contraintes sur les fréquences observées

ℓ = le nombre de paramètres à estimer afin de pouvoir calculer les fréquences espérées.

Dans le cas de la section 7.1, l'hypothèse H_0 spécifiait complètement la distribution de la variable. Il n'y avait donc aucun paramètre à estimer. La somme des fréquences observées était nécessairement égale à n, la taille de l'échantillon. On avait donc une contrainte sur les O_j. La formule (7.7) nous donne donc

$$\text{d.l. } = c-k-\ell = J-1-0 = J-1.$$

Dans le cas de la section 7.2, on avait $c = IJ$ cellules, on avait $k = I$ contraintes sur les fréquences observées (chaque somme-ligne était égale à la taille de l'échantillon correspondant) et on avait $\ell = J-1$ paramètres à estimer (il fallait estimer $p_1, p_2, p_3, ..., p_{J-1}, p_J$, mais comme la somme de ces probabilités est 1, dès qu'on en connaît $J-1$, on connaît automatiquement l'autre). La formule (7.7) nous donne donc

$$\text{d.l. } = c-k-\ell = IJ-I-(J-1) = (I-1)(J-1).$$

Dans le cas de la section 7.3, on avait $c = IJ$ cellules, on avait $k = 1$ contrainte sur les fréquences observées (la somme de toutes les fréquences observées devait être égale à n) et on avait $\ell = (I-1)+(J-1)$ paramètres à estimer (il fallait estimer la loi marginale de X, donc $I-1$ paramètres, et la loi marginale de Y, donc $J-1$ paramètres). La formule (7.7) nous donne donc

$$\text{d.l. } = c-k-\ell = IJ-1-((I-1)+(J-1)) = (I-1)(J-1).$$

7.5 Exercices

NUMÉRO 1. Voici les pourcentages pour chacun des 4 groupes sanguins au Canada :

Groupe	O	A	B	AB
Proportion	0.46	0.39	0.11	0.04

On a déterminé le groupe sanguin de 200 personnes choisies au hasard parmi la population de Chicoutimi. Voici les résultats :

Groupe	O	A	B	AB
Fréquence observée	85	75	27	13

(a) Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette l'hypothèse nulle selon laquelle la distribution des groupes sanguins à Chicoutimi est la même qu'au Canada en général ? Énoncez vos hypothèses. Énoncez votre règle de décision au seuil 5%. Avec les données présentées ci-dessus, quelle est votre décision au seuil 5% ? Quel est votre *p-value* ?

(b) Si votre H_0 était vraie, quelle serait la distribution de votre statistique de test ?

(c) Si votre H_0 était vraie, quelle serait la distribution du nombre de personnes de type sanguin AB dans un échantillon de 200 personnes ?

NUMÉRO 2. Les fleurs d'une certaine espèce sont ou bien rouges, ou bien blanches, ou bien roses. D'un point de vue génétique, les deux premiers types sont purs alors que les fleurs roses sont des hybrides obtenues en croisant une rouge et une blanche. Selon Mendel, si on croise deux fleurs roses, on obtient des rouges 25% du temps, des roses 50% du temps et des blanches 25% du temps. On a fait une expérience. Parmi 564 fleurs obtenues en croisant deux fleurs roses, on a observé 141 fleurs rouges, 291 fleurs roses et 132 fleurs blanches. Est-ce que ces observations sont cohérentes avec la théorie de Mendel ? Expliquez.

NUMÉRO 3. Une expérience a été réalisée afin de comparer 4 insecticides, disons les insecticides A, B, C et D. Chaque insecticide a été utilisé sur un groupe de mouches. Dans chaque cas, on a noté combien de mouches sont mortes et combien de mouches ont survécu. Voici les résultats :

Insecticide	Mouches mortes	Mouches vivantes	Nombre de mouches
A	58	57	115
B	43	77	120
C	56	42	98
D	45	75	120

Y a-t-il une différence significative entre les taux de succès de ces différents insecticides ? Expliquez.

NUMÉRO 4. On considère deux traitements pour le mal de mer, disons le traitement A et le traitement B. On réalise une étude avec 90 volontaires. Parmi eux, 45 recoivent le

traitement A et 45 recoivent le traitement B. Les 90 volontaires font ensuite un voyage en mer d'une durée de 4 heures dans des conditions difficiles. Voici les résultats :

Type de nausée

	Aucune	Faible	Moyenne	Forte	Total
Traitement A	18	17	6	4	45
Traitement B	11	14	14	6	45

Y a-t-il lieu de conclure que ces deux traitements sont différents ?

NUMÉRO 5. Deux cents plants ont été classés selon l'apparence du plant au moment de la floraison (inférieur, moyen, supérieur) et selon la qualité du fruit au moment de la récolte (inférieur, moyen, supérieur). Voici les résultats :

		Qualité du fruit			
		Inférieur	Moyen	Supérieur	Total
Apparence du plant	Inférieur	18	16	10	44
	Moyen	27	61	17	105
	Supérieur	12	16	23	51
	Total	57	93	50	200

Y a-t-il lieu de conclure que les variables $X =$ APPARENCE DU PLANT et $Y =$QUALITÉ DU FRUIT sont dépendantes ?

Annexe A

Tables de distributions

A.1 La loi normale

La table qui apparaît à la page suivante nous permet de trouver la surface à gauche d'une valeur donnée sous la densité de la loi normale de moyenne 0 et de variance 1, aussi appelée la *loi normale standard* ou la *loi normale centrée et réduite.*. Voici quelques exemples illustratifs.

Exemple 1. On suppose que Z suit la loi $N(0,1)$ et on veut trouver $\mathbb{P}[Z \leq 1.26]$. Puisque 1.26 peut s'écrire sous la forme $1.26 = 1.20 + 0.06$, on trouve $\mathbb{P}[Z \leq 1.26]$ à l'intersection de la ligne « 1.2 » et de la colonne « 0.06 » de la table. On obtient $\mathbb{P}[Z \leq 1.26] = \Phi(1.26) = 0.8962$. Bref, la surface à gauche de 1.26 sous la densité de la loi $N(0,1)$ est égale à 0.8962.

Exemple 2. On suppose que Z suit la loi $N(0,1)$ et on veut trouver $\mathbb{P}[Z \leq -0.94]$. En utilisant le fait que la densité de la loi normale est symétrique et en procédant comme à l'exemle 1, on obtient

$$\begin{aligned}\mathbb{P}[Z \leq -0.94] &= \text{ surface à gauche de -0.94 } = \text{ surface à droite de 0.94} \\ &= 1 - \text{ surface à gauche de 0.94 } = 1 - 0.8264 \ = \ 0.1736.\end{aligned}$$

Exemple 3. On suppose que X suit la loi $N(18,4)$, c'est-à-dire la loi normale avec moyenne 18 et avec varance 4, donc écart-type 2, et on veut trouver $\mathbb{P}[16.72 \leq X \leq 18.94]$. D'abord on se ramène à la loi $N(0,1)$, puis on procède comme aux exemples 1 et 2. On obtient

$$\mathbb{P}[16.72 \leq X \leq 18.94] = \mathbb{P}\left[\frac{16.72-18}{\sqrt{4}} \leq Z \leq \frac{18.94-18}{\sqrt{4}}\right] = 0.6808 \ - \ 0.2611 = 0.4197$$

Exemple 4. Supposons qu'on veuille trouver le 99^e centile de la loi $N(0,1)$. En fouillant dans la table principale, on voit que ce 99^e centile est entre 2.32 et 2.33. En utilisant le petit tableau situé au dessous de la grande table, on note que ce 99^e centile est 2.326. Autrement dit, si Z suit la loi normale standard, alors $\mathbb{P}[Z \leq 2.326] = 0.99$. Rappelons que le 99^e centile de la loi normale standard est dénoté $z_{0.01}$. On a donc $z_{0.01} = 2.326$.

Exemple 5. Le quantile d'ordre $1-\gamma$ de la loi $N(\mu, \sigma^2)$ est donnée par la formule $\mu + z_\gamma \, \sigma$. Par exemple, le 95^e centile de la loi $N(200, 400)$ est égal à $200 + z_{0.05} \times 20 = 200 + 1.645 \times 20 = 232.9$.

Fonction de répartition de la loi normale standard

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-x^2/2} \, dx$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

z	0.841	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
$\Phi(z)$	0.8000	0.9000	0.9500	0.9750	0.9800	0.9900	0.9950	0.9975	0.9990	0.9995

A.2 La loi de Student

La table qui apparaît à la page suivante nous donne certains quantiles de la loi de Student. Voici quelques exemples illustratifs.

Exemple 1. Trouvons le quantile d'ordre 0.975 de la loi de Student avec 18 degrés de liberté. On pose $1-\gamma = 0.975$. On a donc $\gamma = 1-0.975 = 0.025$. Dans la table, le quantile d'ordre 0.975 de la loi de Student avec 18 degrés de liberté se trouve donc à l'intersection de la ligne « $k = 18$ » avec la colonne « $\gamma = 0.025$ ». On obtient la valeur 2.101. Ce quantile est habituellement dénoté $t_{18,0.025}$. On a donc $t_{18,0.025} = 2.101$.

Exemple 2. Trouvons le 99^e centile de la loi de Student avec 15 degrés de liberté. Il s'agit donc du quantile d'ordre 0.99. Ce quantile est souvent dénoté $t_{15,0.01}$. On le trouve à l'intersection de la ligne « $k = 15$ » avec la colonne « $\gamma = 0.01$ ». On obtient $t_{15,0.01} = 2.602$.

Exemple 3. Trouvons le 20^e centile de la loi de Student avec 23 degrés de liberté. Il s'agit donc du quantile d'ordre 0.20. Ce quantile est souvent dénoté $t_{23,0.80}$. Puisque la loi de Student est symétrique par rapport à l'origine, on a $t_{23,0.80} = -t_{23,0.20}$. La table nous donne $t_{23,0.20} = 0.858$. On a donc $t_{23,0.80} = -0.858$. Le 20^e centile de la loi de Student avec 23 degrés de liberté est donc égal à -0.858.

Exemple 4. On suppose que T suit la loi de Student avec 9 degrés de liberté. Que vaut $\mathbb{P}[1.10 < T < 3.25]$? On cherche la surface sous la densité de la loi de Student avec 9 degrés de liberté entre l'abscisse $t = 1.10$ et l'abscisse $t = 3.25$. La table nous dit que la surface à gauche de 3.25 est 0.995 et que la surface à gauche de 1.10 est 0.85. La surface recherchée est donc 0.995 - 0.850 = 0.145. On a donc $\mathbb{P}[1.10 < T < 3.25] = 0.145$.

Exemple 5. On suppose que T suit la loi de Student avec 9 degrés de liberté. Que vaut $\mathbb{P}[T \geq 2.4]$? On cherche la surface sous la densité de la loi de Student avec 9 degrés de liberté à droite de l'abscisse $t = 2.4$. La table nous dit que la surface à droite de 2.262 est 0.025 et que la surface à droite de 2.821 est 0.01. La surface recherchée est donc quelque part entre 0.01 et 0.025. Autrement dit, si T suit la loi de Student avec 9 degrés de liberté, alors $0.01 < \mathbb{P}[T \geq 2.4] < 0.025$. Si on fait une interpolation linéaire, on obtient $\mathbb{P}[T \geq 2.4] \approx 0.021$. (D'après le logiciel R, la valeur exacte est 0.01995).

Exemple 6. Trouvons le 95^e centile de la loi de Student avec 45 degrés de liberté. Ce quantile est dénoté $t_{45,0.05}$. La table nous donne $t_{40,0.05} = 1.684$ et $t_{50,0.05} = 1.676$. On peut donc conclure que $1.676 < t_{45,0.05} < 1.684$. Si on fait une interpolation linéaire, on obtient $t_{45,0.05} \approx 1.680$. (D'après le logiciel R, la valeur exacte est 1.6794).

Exemple 7. Nous avons vu au chapitre 3 que lorsque le nombre de degrés de liberté k *s'approche de l'infini*, la loi de Student s'approche de la loi $N(0,1)$. Ainsi, la ligne « $k = \infty$ » de la table de la loi de Student nous donne les quantiles de la loi normale standard. Par exemple, à l'intersection de la ligne « $k = \infty$ » et de la colonne « 0.010 » on trouve la valeur 2.326. Il s'agit du 99^e centile de la loi normale standard. Autrement dit, la table de la loi de Student nous donne $z_{0.01} = 2.326$.

Loi de Student avec k degrés de liberté
Quantiles d'ordre $1-\gamma$

	γ										
k	0.25	0.20	0.15	0.10	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

A.3 La loi du khi-deux

La table qui apparaît à la page suivante nous donne certains quantiles de la loi du khi-deux. Voici quelques exemples illustratifs.

Exemple 1. Trouvons le quantile d'ordre 0.975 de la loi du khi-deux avec 18 degrés de liberté. On pose $1-\gamma = 0.975$. On a donc $\gamma = 1-0.975 = 0.025$. Dans la table, le quantile d'ordre 0.975 de la loi du khi-deux avec 18 degrés de liberté se trouve donc à l'intersection de la ligne « $k = 18$ » avec la colonne « $\gamma = 0.025$ ». On obtient la valeur 31.53. Ce quantile est habituellement dénoté $\chi^2_{18,0.025}$. On a donc $\chi^2_{18,0.025} = 31.53$.

Exemple 2. Trouvons le 99^e centile de la loi du khi-deux avec 15 degrés de liberté. Il s'agit donc du quantile d'ordre 0.99. Ce quantile est souvent dénoté $\chi^2_{15,0.01}$. On le trouve à l'intersection de la ligne « $k = 15$ » avec la colonne « $\gamma = 0.01$ ». On obtient $\chi^2_{15,0.01} = 30.58$.

Exemple 3. Trouvons la médiane de la loi du khi-deux avec 23 degrés de liberté. Il s'agit donc du quantile d'ordre 0.50. Ce quantile est souvent dénoté $\chi^2_{23,0.50}$. La table nous donne $\chi^2_{23,0.50} = 22.34$. La médiane de la loi du khi-deux avec 23 degrés de liberté est donc 22.34.

Exemple 4. On suppose que U suit la loi du khi-deux avec 15 degrés de liberté. Que vaut $\mathbb{P}[8.55 < U < 25.0]$? On cherche la surface sous la densité de la loi du khi-deux avec 15 degrés de liberté entre l'abscisse $u = 8.55$ et l'abscisse $u = 25.0$. La table nous dit que la surface à gauche de 25.0 est 0.95 et que la surface à gauche de 8.55 est 0.10. La surface recherchée est donc 0.95 - 0.10 = 0.85. On a donc $\mathbb{P}[8.55 < U < 25.0] = 0.85$.

Exemple 5. On suppose que U suit la loi du khi-deux avec 7 degrés de liberté. Que vaut $\mathbb{P}[U \geq 12.4]$? On cherche la surface sous la densité de la loi du khi-deux avec 7 degrés de liberté à droite de l'abscisse $u = 12.4$. La table nous dit que la surface à droite de 12.02 est 0.10 et que la surface à droite de 14.07 est 0.05. La surface recherchée est donc quelque part entre 0.05 et 0.10. Autrement dit, si U suit la loi du khi-deux avec 7 degrés de liberté, alors $0.05 < \mathbb{P}[U \geq 12.4] < 0.10$. Si on fait une interpolation linéaire, on obtient $\mathbb{P}[U \geq 12.4] \approx 0.091$. (D'après le logiciel R, la valeur exacte est 0.08815).

Exemple 6. Trouvons le 95^e centile de la loi du khi-deux avec 45 degrés de liberté. Ce quantile est dénoté $\chi^2_{45,0.05}$. La table nous donne $\chi^2_{40,0.05} = 55.76$ et $t_{50,0.05} = 67.50$. On peut donc conclure que $55.76 < \chi^2_{45,0.05} < 67.50$. Si on fait une interpolation linéaire, on obtient $\chi^2_{45,0.05} \approx 61.63$. (D'après le logiciel R, la valeur exacte est 61.6562).

Exemple 7. Supposons qu'on veuille trouver le 95^e centile de la loi du khi-deux avec 200 degrés de liberté. La valeur « $k = 200$ » est hors table. La remarque au bas de la table nous dit que le 95^e centile de la loi du khi-deux avec 200 degrés de liberté peut-être approximé par le 95^e centile de la loi $N(200, 400)$. Ce centile est égal à

$$200 + z_{0.05}\sqrt{400} = 200 + 1.645 \times 20 = 232.9.$$

On a donc $\chi^2_{200,0.05} \approx 232.9$. (D'après le logiciel R, la valeur exacte est 233.9943).

Loi du khi-deux avec k degrés de liberté
Quantiles d'ordre $1-\gamma$

	γ										
k	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.94	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.81	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Si k est entre 30 et 100 mais n'est pas un multiple de 10, on utilise la table ci-haut et on fait une interpolation linéaire. Si $k > 100$ on peut, grâce au théorème limite central, approximer la loi $\chi^2(k)$ par la loi $N(k, 2k)$.

A.4 La loi de Fisher

La table qui apparaît dans les pages suivantes nous donne le 95^e centile de la loi de Fisher avec k degrés de liberté au numérateur et ℓ degrés de liberté au dénominateur. Ce quantile est dénoté $F_{k,\ell,0.05}$. Voici quelques exemples illustratifs.

EXEMPLE 1. Quel est le 95^e centile de la loi de Fisher avec 10 degrés de liberté au numérateur et 15 degrés de liberté au dénominateur ? Ce quantile est dénoté $F_{10,15,0.05}$. On le trouve à l'intersection de la ligne « $\ell = 15$ » avec la colonne « $k = 10$ ». On obtient $F_{10,15,0.05} = 2.544$.

EXEMPLE 2. Quel est le 5^e centile de la loi de Fisher avec 10 degrés de liberté au numérateur et 15 degrés de liberté au dénominateur ? Ce quantile est dénoté $F_{10,15,0.95}$. On utilise la propriété

$$F_{k,\ell,0.95} = \frac{1}{F_{\ell,k,0.05}}.$$

On a donc $F_{10,15,0.95} = 1/F_{15,10,0.05}$. Dans la table, on trouve $F_{15,10,0.05} = 2.845$. On obtient donc $F_{10,15,0.95} = 1/2.845 = 0.3515$.

EXEMPLE 3. On suppose que la variable aléatoire F suit la loi de Fisher avec 5 degrés de liberté au numérateur et 23 degrés de libertés au dénominateur. Que peut-on dire de $\mathbb{P}[F \geq 2.64]$? La table nous dit que la valeur 2.64 est précisément le 95^e centile le la loi de Fisher avec 5 degrés de liberté au numérateur et 23 degrés de libertés au dénominateur. On a donc $\mathbb{P}[F \geq 2.64] = 0.05$.

EXEMPLE 4. On suppose que la variable aléatoire F suit la loi de Fisher avec 8 degrés de liberté au numérateur et 13 degrés de libertés au dénominateur. Que peut-on dire de $\mathbb{P}[F \geq 2.64]$? Selon la table, on a $\mathbb{P}[F \geq 2.767] = 0.05$. On peut donc conclure que $\mathbb{P}[F \geq 2.64]$ est un peu plus grand que 0.05. C'est le mieux qu'on puisse faire avec la table. (Avec l'aide du logiciel R, on obtient $\mathbb{P}[F \geq 2.64] = 0.05805$).

EXEMPLE 5. On suppose que la variable aléatoire F suit la loi de Fisher avec 8 degrés de liberté au numérateur et 13 degrés de libertés au dénominateur. On cherche des nombres a et b pour lesquels on aura $\mathbb{P}[a < F < b] = 0.90$. Il suffit de prendre $a =$ *le* 5^e *centile* et $b =$ *le* 95^e *centile* de la loi de Fisher avec 8 degrés de liberté au numérateur et 13 degrés de libertés au dénominateur. Avec l'aide de la table on obtient

$$\begin{aligned} a &= F_{8,13,0.95} = \frac{1}{F_{13,5,0.05}} = \frac{1}{4.655} = 0.2148, \\ b &= F_{8,13,0.05} = 2.767. \end{aligned}$$

On a donc $\mathbb{P}[0.2148 < F < 2.767] = 0.90$.

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001	1.943	1.894
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
400	3.865	3.018	2.627	2.394	2.237	2.121	2.032	1.962	1.903	1.854

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	11	12	13	14	15	16	17	18	19	20
1	243.0	243.9	244.7	245.4	245.9	246.5	246.9	247.3	247.7	248.0
2	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
3	8.763	8.745	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.660
4	5.936	5.912	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803
5	4.704	4.678	4.655	4.636	4.619	4.604	4.590	4.579	4.568	4.558
6	4.027	4.000	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874
7	3.603	3.575	3.550	3.529	3.511	3.494	3.480	3.467	3.455	3.445
8	3.313	3.284	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.150
9	3.102	3.073	3.048	3.025	3.006	2.989	2.974	2.960	2.948	2.936
10	2.943	2.913	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774
11	2.818	2.788	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646
12	2.717	2.687	2.660	2.637	2.617	2.599	2.583	2.568	2.555	2.544
13	2.635	2.604	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459
14	2.565	2.534	2.507	2.484	2.463	2.445	2.428	2.413	2.400	2.388
15	2.507	2.475	2.448	2.424	2.403	2.385	2.368	2.353	2.340	2.328
16	2.456	2.425	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276
17	2.413	2.381	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.230
18	2.374	2.342	2.314	2.290	2.269	2.250	2.233	2.217	2.203	2.191
19	2.340	2.308	2.280	2.256	2.234	2.215	2.198	2.182	2.168	2.155
20	2.310	2.278	2.250	2.225	2.203	2.184	2.167	2.151	2.137	2.124
21	2.283	2.250	2.222	2.197	2.176	2.156	2.139	2.123	2.109	2.096
22	2.259	2.226	2.198	2.173	2.151	2.131	2.114	2.098	2.084	2.071
23	2.236	2.204	2.175	2.150	2.128	2.109	2.091	2.075	2.061	2.048
24	2.216	2.183	2.155	2.130	2.108	2.088	2.070	2.054	2.040	2.027
25	2.198	2.165	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007
26	2.181	2.148	2.119	2.094	2.072	2.052	2.034	2.018	2.003	1.990
27	2.166	2.132	2.103	2.078	2.056	2.036	2.018	2.002	1.987	1.974
28	2.151	2.118	2.089	2.064	2.041	2.021	2.003	1.987	1.972	1.959
29	2.138	2.104	2.075	2.050	2.027	2.007	1.989	1.973	1.958	1.945
30	2.126	2.092	2.063	2.037	2.015	1.995	1.976	1.960	1.945	1.932
40	2.038	2.003	1.974	1.948	1.924	1.904	1.885	1.868	1.853	1.839
50	1.986	1.952	1.921	1.895	1.871	1.850	1.831	1.814	1.798	1.784
60	1.952	1.917	1.887	1.860	1.836	1.815	1.796	1.778	1.763	1.748
70	1.928	1.893	1.863	1.836	1.812	1.790	1.771	1.753	1.737	1.722
80	1.910	1.875	1.845	1.817	1.793	1.772	1.752	1.734	1.718	1.703
90	1.897	1.861	1.830	1.803	1.779	1.757	1.737	1.720	1.703	1.688
100	1.886	1.850	1.819	1.792	1.768	1.746	1.726	1.708	1.691	1.676
150	1.853	1.817	1.786	1.758	1.734	1.711	1.691	1.673	1.656	1.641
200	1.837	1.801	1.769	1.742	1.717	1.694	1.674	1.656	1.639	1.623
400	1.813	1.776	1.745	1.717	1.691	1.669	1.648	1.630	1.613	1.597

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	21	22	23	24	25	26	27	28	29	30
1	248.3	248.6	248.8	249.1	249.3	249.5	249.6	249.8	250.0	250.1
2	19.45	19.45	19.45	19.45	19.46	19.46	19.46	19.46	19.46	19.46
3	8.654	8.648	8.643	8.639	8.634	8.630	8.626	8.623	8.620	8.617
4	5.795	5.787	5.781	5.774	5.769	5.763	5.759	5.754	5.750	5.746
5	4.549	4.541	4.534	4.527	4.521	4.515	4.510	4.505	4.500	4.496
6	3.865	3.856	3.849	3.841	3.835	3.829	3.823	3.818	3.813	3.808
7	3.435	3.426	3.418	3.410	3.404	3.397	3.391	3.386	3.381	3.376
8	3.140	3.131	3.123	3.115	3.108	3.102	3.095	3.090	3.084	3.079
9	2.926	2.917	2.908	2.900	2.893	2.886	2.880	2.874	2.869	2.864
10	2.764	2.754	2.745	2.737	2.730	2.723	2.716	2.710	2.705	2.700
11	2.636	2.626	2.617	2.609	2.601	2.594	2.588	2.582	2.576	2.570
12	2.533	2.523	2.514	2.505	2.498	2.491	2.484	2.478	2.472	2.466
13	2.448	2.438	2.429	2.420	2.412	2.405	2.398	2.392	2.386	2.380
14	2.377	2.367	2.357	2.349	2.341	2.333	2.326	2.320	2.314	2.308
15	2.316	2.306	2.297	2.288	2.280	2.272	2.265	2.259	2.253	2.247
16	2.264	2.254	2.244	2.235	2.227	2.220	2.212	2.206	2.200	2.194
17	2.219	2.208	2.199	2.190	2.181	2.174	2.167	2.160	2.154	2.148
18	2.179	2.168	2.159	2.150	2.141	2.134	2.126	2.119	2.113	2.107
19	2.144	2.133	2.123	2.114	2.106	2.098	2.090	2.084	2.077	2.071
20	2.112	2.102	2.092	2.082	2.074	2.066	2.059	2.052	2.045	2.039
21	2.084	2.073	2.063	2.054	2.045	2.037	2.030	2.023	2.016	2.010
22	2.059	2.048	2.038	2.028	2.020	2.012	2.004	1.997	1.990	1.984
23	2.036	2.025	2.014	2.005	1.996	1.988	1.981	1.973	1.967	1.961
24	2.015	2.003	1.993	1.984	1.975	1.967	1.959	1.952	1.945	1.939
25	1.995	1.984	1.974	1.964	1.955	1.947	1.939	1.932	1.926	1.919
26	1.978	1.966	1.956	1.946	1.938	1.929	1.921	1.914	1.907	1.901
27	1.961	1.950	1.940	1.930	1.921	1.913	1.905	1.898	1.891	1.884
28	1.946	1.935	1.924	1.915	1.906	1.897	1.889	1.882	1.875	1.869
29	1.932	1.921	1.910	1.901	1.891	1.883	1.875	1.868	1.861	1.854
30	1.919	1.908	1.897	1.887	1.878	1.870	1.862	1.854	1.847	1.841
40	1.826	1.814	1.803	1.793	1.783	1.775	1.766	1.759	1.751	1.744
50	1.771	1.759	1.748	1.737	1.727	1.718	1.710	1.702	1.694	1.687
60	1.735	1.722	1.711	1.700	1.690	1.681	1.672	1.664	1.656	1.649
70	1.709	1.696	1.685	1.674	1.664	1.654	1.646	1.637	1.629	1.622
80	1.689	1.677	1.665	1.654	1.644	1.634	1.626	1.617	1.609	1.602
90	1.675	1.662	1.650	1.639	1.629	1.619	1.610	1.601	1.593	1.586
100	1.663	1.650	1.638	1.627	1.616	1.607	1.598	1.589	1.581	1.573
150	1.627	1.614	1.602	1.590	1.580	1.570	1.560	1.552	1.543	1.535
200	1.609	1.596	1.583	1.572	1.561	1.551	1.542	1.533	1.524	1.516
400	1.582	1.569	1.556	1.545	1.534	1.523	1.514	1.505	1.496	1.488

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	40	50	60	70	80	90	100	150	200	400
1	251.1	251.8	252.2	252.5	252.7	252.9	253.0	253.5	253.7	253.8
2	19.47	19.48	19.48	19.48	19.48	19.48	19.49	19.49	19.49	19.49
3	8.594	8.581	8.572	8.566	8.561	8.557	8.554	8.545	8.540	8.537
4	5.717	5.699	5.688	5.679	5.673	5.668	5.664	5.652	5.646	5.643
5	4.464	4.444	4.431	4.422	4.415	4.409	4.405	4.392	4.385	4.381
6	3.774	3.754	3.740	3.730	3.722	3.716	3.712	3.698	3.690	3.686
7	3.340	3.319	3.304	3.294	3.286	3.280	3.275	3.260	3.252	3.248
8	3.043	3.020	3.005	2.994	2.986	2.980	2.975	2.959	2.951	2.947
9	2.826	2.803	2.787	2.776	2.768	2.761	2.756	2.739	2.731	2.726
10	2.661	2.637	2.621	2.610	2.601	2.594	2.588	2.572	2.563	2.558
11	2.531	2.507	2.490	2.478	2.469	2.462	2.457	2.439	2.431	2.426
12	2.426	2.401	2.384	2.372	2.363	2.356	2.350	2.332	2.323	2.318
13	2.339	2.314	2.297	2.284	2.275	2.267	2.261	2.243	2.234	2.229
14	2.266	2.241	2.223	2.210	2.201	2.193	2.187	2.169	2.159	2.154
15	2.204	2.178	2.160	2.147	2.137	2.130	2.123	2.105	2.095	2.089
16	2.151	2.124	2.106	2.093	2.083	2.075	2.068	2.049	2.039	2.034
17	2.104	2.077	2.058	2.045	2.035	2.027	2.020	2.001	1.991	1.985
18	2.063	2.035	2.017	2.003	1.993	1.985	1.978	1.958	1.948	1.942
19	2.026	1.999	1.980	1.966	1.955	1.947	1.940	1.920	1.910	1.903
20	1.994	1.966	1.946	1.932	1.922	1.913	1.907	1.886	1.875	1.869
21	1.965	1.936	1.916	1.902	1.891	1.883	1.876	1.855	1.845	1.838
22	1.938	1.909	1.889	1.875	1.864	1.856	1.849	1.827	1.817	1.810
23	1.914	1.885	1.865	1.850	1.839	1.830	1.823	1.802	1.791	1.784
24	1.892	1.863	1.842	1.828	1.816	1.808	1.800	1.779	1.768	1.761
25	1.872	1.842	1.822	1.807	1.796	1.787	1.779	1.757	1.746	1.739
26	1.853	1.823	1.803	1.788	1.776	1.767	1.760	1.738	1.726	1.719
27	1.836	1.806	1.785	1.770	1.758	1.749	1.742	1.719	1.708	1.701
28	1.820	1.790	1.769	1.754	1.742	1.733	1.725	1.702	1.691	1.683
29	1.806	1.775	1.754	1.738	1.726	1.717	1.710	1.686	1.675	1.667
30	1.792	1.761	1.740	1.724	1.712	1.703	1.695	1.672	1.660	1.652
40	1.693	1.660	1.637	1.621	1.608	1.597	1.589	1.564	1.551	1.542
50	1.634	1.599	1.576	1.558	1.544	1.534	1.525	1.498	1.484	1.475
60	1.594	1.559	1.534	1.516	1.502	1.491	1.481	1.453	1.438	1.428
70	1.566	1.530	1.505	1.486	1.471	1.459	1.450	1.420	1.404	1.394
80	1.545	1.508	1.482	1.463	1.448	1.436	1.426	1.395	1.379	1.368
90	1.528	1.491	1.465	1.445	1.429	1.417	1.407	1.375	1.358	1.348
100	1.515	1.477	1.450	1.430	1.415	1.402	1.392	1.359	1.342	1.331
150	1.475	1.436	1.407	1.386	1.369	1.356	1.345	1.309	1.290	1.278
200	1.455	1.415	1.386	1.364	1.346	1.332	1.321	1.283	1.263	1.249
400	1.425	1.383	1.352	1.329	1.311	1.296	1.283	1.242	1.219	1.204