HAL open science

Résumé du Cours de Statistique Descriptive avec une Introduction au Calcul de Probabilités

Yves Tillé

▶ **To cite this version:**

Yves Tillé. Résumé du Cours de Statistique Descriptive avec une Introduction au Calcul de Probabilités. Licence. Statistique Descriptive, Suisse. 2023, pp.226. ⟨hal-04132484v2⟩

HAL Id: hal-04132484

https://hal.science/hal-04132484v2

Submitted on 27 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

Résumé du Cours de

Statistique Descriptive

avec une Introduction au Calcul de Probabilités

Yves Tillé

avec la collaboration de

Caren Hasler, Matti Langel, Lionel Qualité, Vittoria Sacco et Audrey-Anne Vallée

Seconde édition avec exercices corrigés

27 janvier 2025

Préface

Ce document contient les notes du cours de "statistique descriptive" donné à l'Université de Neuchâtel jusqu'en 2015. Ce cours contient une introduction à la statistique descriptive de base. Outre la définition des paramètres pour la statistique univariée et bivariée, une introduction élémentaire aux indices d'inégalité et aux indices de prix est développée. Quelques éléments sont également présentés pour le traitement des séries temporelles, notamment concernant les méthodes de désaisonnalisation. Le fascicule se termine par une petite introduction au calcul des probabilités.

Ce manuel contient également un ensemble d'exercices qui ont été réalisés par les collaboratrices et collaborateurs qui ont préparé ou animé les séances de travaux pratiques de ce cours : Caren Hasler, Matti Langel, Lionel Qualité, Vittoria Sacco et Audrey-Anne Vallée. Un ensemble de questions ayant servi aux différentes évaluations sont aussi présentées. Toutes les solutions se trouvent dans le fascicule. Cette version est donc plus complète que les anciennes versions qui ont été disséminées sur internet.

Après avoir décidé de mettre mon matériel pédagogique en ligne, j'ai reçu un grand nombre de courriers électroniques qui montrent que le cours a été utile pour beaucoup d'étudiants. Un manuel est probablement le meilleur outil pour un étudiant désireux d'apprendre.

Ce fascicule comprend un ensemble de codes en langage R. Pour une introduction bien faite en français, le manuel en libre accès de Paradis (2002) est très bien réalisé. On peut le télécharger facilement depuis le site de développement du langage R.

Yves Tillé, le 27 janvier 2025

Table des matières

I Cours 11

1 Variables, données statistiques, tableaux, effectifs 13
1.1 Définitions fondamentales . . . 13
1.1.1 Mesure et variable . . . 13
1.1.2 Typologie des variables . . . 13
1.1.3 Série statistique . . . 13
1.2 Variable qualitative nominale . . . 14
1.2.1 Effectifs, fréquences et tableau statistique . . . 14
1.2.2 Diagramme en secteurs et diagramme en barres . . . 15
1.3 Variable qualitative ordinale . . . 16
1.3.1 Le tableau statistique . . . 16
1.3.2 Diagramme en secteurs . . . 17
1.3.3 Diagramme en barres des effectifs et des effectifs cumulés . . . 17
1.4 Variable quantitative discrète . . . 17
1.4.1 Le tableau statistique . . . 17
1.4.2 Diagramme en bâtonnets des effectifs . . . 19
1.4.3 Fonction de répartition . . . 19
1.5 Variable quantitative continue . . . 20
1.5.1 Le tableau statistique . . . 20
1.5.2 Histogramme . . . 21
1.5.3 La fonction de répartition . . . 22

2 Statistique descriptive univariée 25
2.1 Paramètres de position . . . 25
2.1.1 Le mode . . . 25
2.1.2 La moyenne arithmétique . . . 25
2.1.3 Remarques sur le signe de sommation $\sum$ . . . 26
2.1.4 Moyenne géométrique . . . 28
2.1.5 Moyenne harmonique . . . 28
2.1.6 Moyenne pondérée . . . 29
2.1.7 La médiane . . . 29
2.1.8 Quantiles . . . 31
2.2 Paramètres de dispersion . . . 32
2.2.1 L'étendue . . . 32
2.2.2 La distance interquartile . . . 32
2.2.3 La variance . . . 32
2.2.4 L'écart-type . . . 33
2.2.5 L'écart moyen absolu . . . 34
2.2.6 L'écart médian absolu . . . 34
2.3 Moments . . . 34
2.4 Paramètres de forme . . . 35
2.4.1 Coefficient d'asymétrie de Fisher (skewness) . . . 35
2.4.2 Coefficient d'asymétrie de Yule . . . 35
2.4.3 Coefficient d'asymétrie de Pearson . . . 35
2.5 Paramètre d'aplatissement (kurtosis) . . . 36
2.6 Changement d'origine et d'unité . . . 36
2.7 Moyennes et variances dans des groupes . . . 37
2.8 Diagramme en tiges et feuilles . . . 38
2.9 La boîte à moustaches . . . 39

3 Statistique descriptive bivariée **41**
3.1 Série statistique bivariée . . . 41
3.2 Deux variables quantitatives . . . 41
3.2.1 Représentation graphique de deux variables . . . 41
3.2.2 Analyse des variables . . . 42
3.2.3 Covariance . . . 42
3.2.4 Corrélation . . . 43
3.2.5 Droite de régression . . . 43
3.2.6 Résidus et valeurs ajustées . . . 45
3.2.7 Sommes de carrés et variances . . . 46
3.2.8 Décomposition de la variance . . . 47
3.3 Deux variables qualitatives . . . 48
3.3.1 Données observées . . . 48
3.3.2 Tableau de contingence . . . 48
3.3.3 Tableau des fréquences . . . 49
3.3.4 Profils lignes et profils colonnes . . . 50
3.3.5 Effectifs théoriques et khi-carré . . . 50

4 Théorie des indices, mesures d'inégalité **55**
4.1 Nombres indices . . . 55
4.1.1 Définition . . . 55
4.1.2 Propriétés des indices . . . 55
4.1.3 Indices synthétiques . . . 56
4.1.4 Indice de Laspeyres . . . 56
4.1.5 Indice de Paasche . . . 57
4.1.6 L'indice de Fisher . . . 57
4.1.7 L'indice de Sidgwick . . . 58
4.1.8 Indices chaînes . . . 58
4.2 Mesures de l'inégalité . . . 58
4.2.1 Introduction . . . 58
4.2.2 Courbe de Lorenz . . . 58
4.2.3 Indice de Gini . . . 59
4.2.4 Robin Hood index . . . 60
4.2.5 Quintile et Decile share ratio . . . 60
4.2.6 Indice de pauvreté . . . 60
4.2.7 Indices selon les pays . . . 60

5 Séries temporelles **63**
5.1 Définitions générales et exemples . . . 63
5.1.1 Définitions . . . 63
5.1.2 Traitement des séries temporelles . . . 63
5.1.3 Exemples . . . 63
5.2 Description de la tendance . . . 68
5.2.1 Les principaux modèles . . . 68
5.2.2 Tendance linéaire . . . 69
5.2.3 Tendance quadratique . . . 69
5.2.4 Tendance polynomiale d'ordre q . . . 69
5.2.5 Tendance logistique . . . 69
5.3 Opérateurs de décalage et de différence . . . 70
5.3.1 Opérateurs de décalage . . . 70
5.3.2 Opérateur différence . . . 70
5.3.3 Différence saisonnière . . . 71
5.4 Filtres linéaires et moyennes mobiles . . . 74
5.4.1 Filtres linéaires . . . 74
5.4.2 Moyennes mobiles : définition . . . 74
5.4.3 Moyenne mobile et composante saisonnière . . . 74
5.5 Moyennes mobiles particulières . . . 75
5.5.1 Moyenne mobile de Van Hann . . . 75
5.5.2 Moyenne mobile de Spencer . . . 75
5.5.3 Moyenne mobile de Henderson . . . 75
5.5.4 Médianes mobiles . . . 76

5.6 Désaisonnalisation . . . 76
5.6.1 Méthode additive . . . 76
5.6.2 Méthode multiplicative . . . 77
5.7 Lissage exponentiel . . . 78
5.7.1 Lissage exponentiel simple . . . 78
5.7.2 Lissage exponentiel double . . . 80

6 Calcul des probabilités et variables aléatoires 85
6.1 Probabilités . . . 85
6.1.1 Événement . . . 85
6.1.2 Opérations sur les événements . . . 85
6.1.3 Relations entre les événements . . . 86
6.1.4 Ensemble des parties d'un ensemble et système complet . . . 86
6.1.5 Axiomatique des Probabilités . . . 86
6.1.6 Probabilités conditionnelles et indépendance . . . 88
6.1.7 Théorème des probabilités totales et théorème de Bayes . . . 89
6.2 Analyse combinatoire . . . 90
6.2.1 Introduction . . . 90
6.2.2 Permutations (sans répétition) . . . 90
6.2.3 Permutations avec répétition . . . 90
6.2.4 Arrangements (sans répétition) . . . 90
6.2.5 Combinaisons . . . 91
6.3 Variables aléatoires . . . 91
6.4 Variables aléatoires discrètes . . . 91
6.4.1 Définition, espérance et variance . . . 91
6.4.2 Variable indicatrice ou bernoullienne . . . 92
6.4.3 Variable binomiale . . . 92
6.4.4 Variable de Poisson . . . 94
6.5 Variable aléatoire continue . . . 95
6.5.1 Définition, espérance et variance . . . 95
6.5.2 Variable uniforme . . . 96
6.5.3 Variable normale . . . 98
6.5.4 Variable normale centrée réduite . . . 99
6.5.5 Distribution exponentielle . . . 99
6.6 Distribution bivariée . . . 100
6.6.1 Cas continu . . . 100
6.6.2 Cas discret . . . 101
6.6.3 Remarques . . . 102
6.6.4 Indépendance de deux variables aléatoires . . . 102
6.7 Propriétés des espérances et des variances . . . 102
6.8 Autres variables aléatoires . . . 104
6.8.1 Variable khi-carrée . . . 104
6.8.2 Variable de Student . . . 104
6.8.3 Variable de Fisher . . . 105
6.8.4 Distribution normale bivariée . . . 105

II Exercices 111

7 Exercices : Variables, données statistiques, tableaux, effectifs 113
7.1 Types de variables . . . 113
7.2 Séries statistiques et graphiques . . . 114
7.3 Poids d'élèves . . . 118
7.4 Variables et graphiques . . . 120
7.5 Histogrammes . . . 125
7.6 Histogramme et classes . . . 127
7.7 Variables, types et graphiques . . . 129

8 Exercices : Statistique descriptive univariée 133
8.1 Opérateur de sommation 1 . . . 133
8.2 Opérateur de sommation 2 . . . 136

8.3 Variance avec une double somme . . . 137
8.4 Classes d'élèves . . . 138
8.5 Moyennes arithmétique, géometrique et harmonique . . . 141
8.6 Changement d'origine . . . 142
8.7 Quelle moyenne ? . . . 143
8.8 Calcul de paramètres . . . 144
8.9 Salaires hommes et femmes . . . 145
8.10 Nombre d'enfants . . . 146
8.11 Moyennes géométrique, harmonique ou arithmétique . . . 147
8.12 Primes d'assurance . . . 149
8.13 Distribution de salaires . . . 151
8.14 Boxplot . . . 153
8.15 Variances . . . 154
8.16 Âges dans les familles . . . 155
8.17 Spectateurs dans les stades . . . 156
8.18 Paramètres dans une distribution . . . 158
8.19 Séries et calcul de paramètres . . . 160

9 Exercices : Statistique descriptive bivariée 167
9.1 Années d'études et revenus . . . 167
9.2 Ancienneté et absence . . . 169
9.3 Crèmes glacées . . . 171
9.4 Poids et tailles . . . 172
9.5 Cours du dollar . . . 174
9.6 Avis pédagogiques . . . 177
9.7 Consommation de médicaments . . . 179
9.8 Daltonie . . . 181
9.9 Loyers et cantons . . . 182
9.10 Bières . . . 183
9.11 État civil et nationalité . . . 184
9.12 Dépendance entre variables dichotomiques . . . 186
9.13 Manque de sommeil . . . 188
9.14 Habitudes alimenaires selon le sexe . . . 190

10 Exercices : Théorie des indices, mesures d'inégalité 191
10.1 Indices simples . . . 191
10.2 Indice de Laspeyres . . . 193
10.3 Indice de Laspeyres et de Paasche . . . 194
10.4 Indices composites . . . 196
10.5 Indices de prix et montres . . . 198
10.6 Indice et lait . . . 200
10.7 Indice chaîne . . . 202
10.8 Courbe de Lorenz et inégalités . . . 203
10.9 Indice de Gini . . . 205
10.10 Indices d'inégalité et courbe de Lorenz . . . 207
10.11 Comparaison de courbes de Lorenz . . . 208
10.12 Calcul d'indices d'inégalité . . . 209
10.13 Revenus dans les États . . . 211

11 Exercices : Séries temporelles 213
11.1 Opérateurs de décalage 1 . . . 213
11.2 Opérateurs de décalage 2 . . . 214
11.3 Villas vendues . . . 215
11.4 Désaisonnalisation . . . 218
11.5 Décomposition et désaisonnalisation . . . 219

12 Exercices : Calcul des probabilités et variables aléatoires 223
12.1 Cluedo . . . 223
12.2 Cadenas . . . 223
12.3 Cartes . . . 224
12.4 Opinion . . . 225

12.5 Tirage de jetons . . . 227
12.6 Cartes et évènements . . . 228
12.7 Séquence d'enfants . . . 229
12.8 Activités après le bachelor . . . 230
12.9 Boîtes de chocolats . . . 231
12.10 Jeu avec des cartes . . . 232
12.11 Boules de Noël . . . 233
12.12 Lancer de dés . . . 235
12.13 Chaussures et probabilités . . . 236
12.14 Pratiques culturelles . . . 237
12.15 Qualité d'ampoules . . . 238
12.16 Urne et boules . . . 239
12.17 Probabiliés d'accident . . . 240
12.18 Théorème des probabilités totales . . . 241
12.19 Monstres . . . 241
12.20 Temps de travail . . . 242
12.21 Tabagisme et cancer . . . 243
12.22 Théorème des probabilités totales et de Bayes . . . 244
12.23 Test de grossesse . . . 245
12.24 Test pharmaceutique . . . 246
12.25 Anagrammes . . . 247
12.26 Course de chevaux . . . 248
12.27 Avion . . . 249
12.28 Places à table . . . 249
12.29 Notation binaire . . . 250
12.30 Loterie . . . 251
12.31 Euro Millions . . . 251
12.32 Poker . . . 252
12.33 Choix de films . . . 253
12.34 Cartes et mains . . . 254
12.35 Espérance et écart-type . . . 255
12.36 Espérance et variance . . . 255
12.37 Vol dans les magasins . . . 256
12.38 Jeu de dés . . . 257
12.39 Répartition et espérance . . . 258
12.40 Assurance et prime . . . 259
12.41 Urne et jetons . . . 260
12.42 Classe et échantillon . . . 261
12.43 Germination . . . 262
12.44 Pièce de monnaie . . . 263
12.45 Taux de réussite . . . 264
12.46 Excès de vitesse et loi de Poisson . . . 265
12.47 Employés absents pour cause de maladie et loi de Poisson . . . 266
12.48 File d'attente et loi de Poisson . . . 267
12.49 Table de la loi normale 1 . . . 268
12.50 Table de la loi normale 2 . . . 269
12.51 Table de la loi normale 3 . . . 269
12.52 Table de la loi normale 4 . . . 270
12.53 Lecture inverse de la table de la loi normale . . . 270
12.54 Résultat et loi normale . . . 271
12.55 Variable normale standardisation . . . 272
12.56 Tailles et loi normale . . . 273
12.57 Temps et loi normale . . . 273
12.58 Vitesse et loi normale . . . 274
12.59 Consommation d'eau et loi normale . . . 275
12.60 Boîte de conserve et loi normale . . . 276
12.61 Cylindres et loi normale . . . 277
12.62 Théorème des probabilités totales . . . 278
12.63 Théorème de Bayes . . . 279
12.64 Lecture des tables statistiques . . . 279

III Matériel pour l'évaluation **281**

13 Questions à choix multiples **283**

14 Questions ouvertes **323**

IV Annexes **403**

Tables statistiques **405**

Liste des tableaux **413**

Liste des figures **417**

Références **419**

Index **421**

Première partie

Cours

Chapitre 1

Variables, données statistiques, tableaux, effectifs

1.1 Définitions fondamentales

1.1.1 Mesure et variable

- On s'intéresse à des *unités statistiques* ou *unités d'observation* : par exemple des individus, des entreprises, des ménages. En sciences humaines, on s'intéresse dans la plupart des cas à un nombre fini d'unités.
- Sur ces unités, on mesure un caractère ou une *variable*, le chiffre d'affaires de l'entreprise, le revenu du ménage, l'âge de la personne, la catégorie socioprofessionnelle d'une personne. On suppose que la variable prend toujours une seule valeur sur chaque unité. Les variables sont désignées par simplicité par une lettre (X, Y, Z).
- Les *valeurs possibles* de la variable, sont appelées *modalités*.
- L'ensemble des valeurs possibles ou des modalités est appelé le *domaine* de la variable.

1.1.2 Typologie des variables

- *Variable qualitative* : La variable est dite qualitative quand les modalités sont des catégories.
  - *Variable qualitative nominale* : La variable est dite qualitative nominale quand les modalités ne peuvent pas être ordonnées.
  - *Variable qualitative ordinale* : La variable est dite qualitative ordinale quand les modalités peuvent être ordonnées. Le fait de pouvoir ou non ordonner les modalités est parfois discutable. Par exemple : dans les catégories socioprofessionnelles, on admet d'ordonner les modalités : 'ouvriers', 'employés', 'cadres'. Si on ajoute les modalités 'sans profession', 'enseignant', 'artisan', l'ordre devient beaucoup plus discutable.
- *Variable quantitative* : Une variable est dite quantitative si toute ses valeurs possibles sont numériques.
  - *Variable quantitative discrète* : Une variable est dite discrète, si l'ensemble des valeurs possibles est dénombrable.
  - *Variable quantitative continue* : Une variable est dite continue, si l'ensemble des valeurs possibles est continu.

Remarque 1.1 Ces définitions sont à relativiser, l'âge est théoriquement une variable quantitative continue, mais en pratique, l'âge est mesuré dans le meilleur des cas au jour près. Toute mesure est limitée en précision !

Exemple 1.1 Les modalités de la variable *sexe* sont *masculin* (codé M) et *féminin* (codé F). Le domaine de la variable est $\{M, F\}$.

Exemple 1.2 Les modalités de la variable nombre d'enfants par famille sont $0, 1, 2, 3, 4, 5, \ldots$ C'est une variable quantitative discrète.

1.1.3 Série statistique

On appelle *série statistique* la suite des valeurs prises par une variable X sur les unités d'observation.
Le nombre d'unités d'observation est noté n.

Les valeurs de la variable X sont notées

$$x_1, \ldots, x_i, \ldots, x_n.$$

Exemple 1.3 On s'intéresse à la variable 'état-civil' notée X et à la série statistique des valeurs prises par X sur 20 personnes. La codification est présentée dans le Tableau 1.1.

TABLE 1.1 – Domaine de la variable

C :	célibataire,
M :	marié(e),
V :	veuf(ve),
D :	divorcée.

Le domaine de la variable X est $\{C, M, V, D\}$. Considérons la série statistique présentée dans le Tableau 1.2.

TABLE 1.2 – Série statistique

M	M	D	C	C	M	C	C	C	M	C	M	V	M	V	D	C	C	C	M

Ici, $n = 20$,

$$x_1 = M, x_2 = M, x_3 = D, x_4 = C, x_5 = C, \ldots, x_{20} = M.$$

1.2 Variable qualitative nominale

1.2.1 Effectifs, fréquences et tableau statistique

Une variable qualitative nominale a des valeurs distinctes qui ne peuvent pas être ordonnées. On note J le nombre de valeurs distinctes ou modalités. Les valeurs distinctes sont notées $x_1, \ldots, x_j, \ldots, x_J$. On appelle *effectif* d'une modalité ou d'une valeur distincte, le nombre de fois que cette modalité (ou valeur distincte) apparaît. On note n_j l'effectif de la modalité x_j. La fréquence d'une modalité est l'effectif divisé par le nombre d'unités d'observation.

$$f_j = \frac{n_j}{n}, j = 1, \ldots, J.$$

Exemple 1.4 Avec la série de l'exemple précédent, on obtient le Tableau 1.3.

TABLE 1.3 – Tableau statistique

x_j	n_j	f_j
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10
	$n = 20$	1

En langage R

```
>X=c('Marié(e)','Marié(e)','Divorcé(e)','Célibataire','Célibataire',
'Marié(e)','Célibataire','Célibataire','Célibataire','Marié(e)',
'Célibataire','Marié(e)','Veuf(ve)','Marié(e)','Veuf(ve)','Divorcé(e)',
'Célibataire','Célibataire','Célibataire','Marié(e)')
> T1=table(X)
```

```
> V1=c(T1)
> data.frame(Eff=V1,Freq=V1/sum(V1))
Eff Freq
Célibataire 9 0.45
Divorcé(e)  2 0.10
Marié(e)    7 0.35
Veuf(ve)    2 0.10
```

1.2.2 Diagramme en secteurs et diagramme en barres

Le tableau statistique d'une variable qualitative nominale peut être représenté par deux types de graphique. Les effectifs sont représentés par un diagramme en barres et les fréquences par un diagramme en secteurs (ou camembert ou *piechart* en anglais) (voir Figures 1.1 et 1.2).

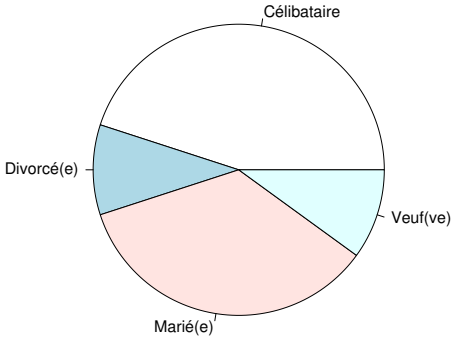


FIGURE 1.1 – Diagramme en secteurs des fréquences

En langage R

```
> pie(T1,radius=1.0)
```

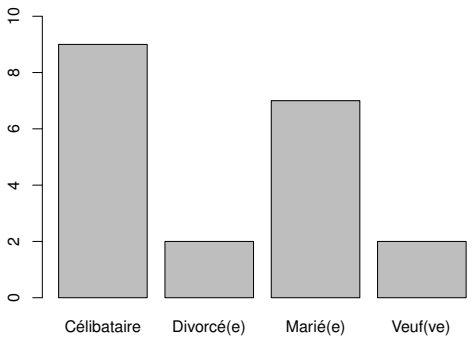


FIGURE 1.2 – Diagramme en barres des effectifs

En langage R

```
>m=max(V1)
>barplot(T1, ylim=c(0,m+1))
```

1.3 Variable qualitative ordinale

1.3.1 Le tableau statistique

Les valeurs distinctes d'une variable ordinale peuvent être ordonnées, ce qu'on écrit

$$x_1 \prec x_2 \prec \cdots \prec x_{j-1} \prec x_j \prec \cdots \prec x_{J-1} \prec x_J.$$

La notation $x_1 \prec x_2$ se lit x_1 précède x_2.

Si la variable est ordinale, on peut calculer les effectifs cumulés :

$$N_j = \sum_{k=1}^{j} n_k, j = 1, \ldots, J.$$

On a $N_1 = n_1$ et $N_J = n$. On peut également calculer les fréquences cumulées

$$F_j = \frac{N_j}{n} = \sum_{k=1}^{j} f_k, j = 1, \ldots, J.$$

Exemple 1.5 On interroge 50 personnes sur leur dernier diplôme obtenu (variable Y). La codification a été faite selon le Tableau 1.4. On a obtenu la série statistique présentée dans le Tableau 1.5. Finalement, on obtient le tableau statistique complet présenté dans le Tableau 1.6.

Table 1.4 – Codification de la variable Y

Dernier diplôme obtenu	x_j
Sans diplôme	Sd
Primaire	P
Secondaire	Se
Supérieur non-universitaire	Su
Universitaire	U

Table 1.5 – Série statistique de la variable Y

Sd	Sd	Sd	Sd	P	P	P	P	P	P	P	P	P	P	P	Se	Se
Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Su	Su	Su	Su	Su
Su	Su	Su	Su	U	U	U	U	U	U	U	U	U	U	U	U	

Table 1.6 – Tableau statistique complet

x_j	n_j	N_j	f_j	F_j
Sd	4	4	0.08	0.08
P	11	15	0.22	0.30
Se	14	29	0.28	0.58
Su	9	38	0.18	0.76
U	12	50	0.24	1.00
	50		1.00	

En langage R

```
> YY=c("Sd","Sd","Sd","Sd","P","P","P","P","P","P","P","P","P","P","P",
"Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se",
"Su","Su","Su","Su","Su","Su","Su","Su","Su",
"U","U","U","U","U","U","U","U","U","U","U","U")
YF=factor(YY,levels=c("Sd","P","Se","Su","U"))
T2=table(YF)
V2=c(T2)
> data.frame(Eff=V2,EffCum=cumsum(V2),Freq=V2/sum(V2),
FreqCum=cumsum(V2/sum(V2)))
Eff EffCum Freq FreqCum
Sd  4   4 0.08  0.08
P  11   15 0.22  0.30
Se 14   29 0.28  0.58
Su  9   38 0.18  0.76
U  12   50 0.24  1.00
```

1.3.2 Diagramme en secteurs

Les fréquences d'une variable qualitative ordinale sont représentées au moyen d'un diagramme en secteurs (voir Figure 1.3).

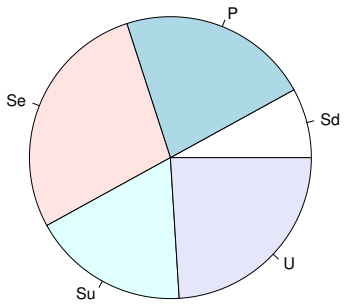


FIGURE 1.3 – Diagramme en secteurs des fréquences

En langage R

```
> pie(T2,radius=1)
```

1.3.3 Diagramme en barres des effectifs et des effectifs cumulés

Les effectifs et les effectifs umulés d'une variable qualitative ordinale sont représentés au moyen d'un diagramme en barres (voir Figure 1.4).

En langage R

```
barplot(T2)
T3=cumsum(T2)
barplot(T3)
```

1.4 Variable quantitative discrète

1.4.1 Le tableau statistique

Une variable discrète a un domaine dénombrable.

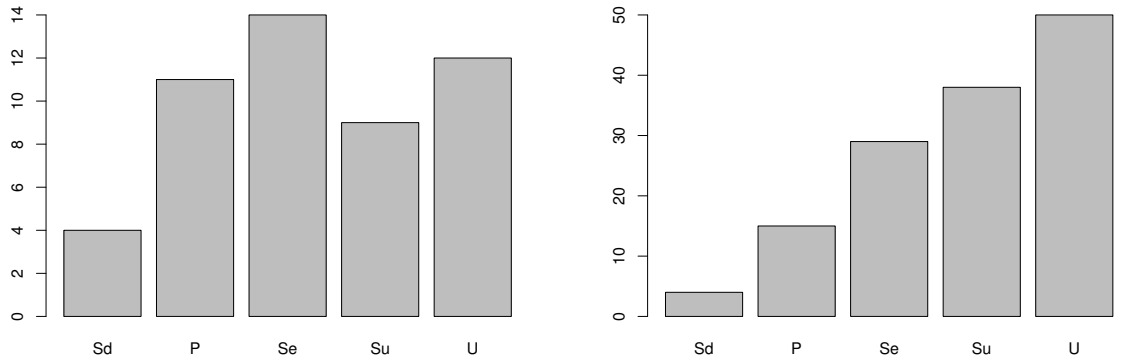


FIGURE 1.4 – Diagrammes en barres des effectifs et des effectifs cumulés

Exemple 1.6 Un quartier est composé de 50 ménages et la variable Z représente le nombre de personnes par ménage. Les valeurs de la variable sont dans le Tableau 1.7.

TABLE 1.7 – Nombre de personnes par ménage

1	1	1	1	1	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	6	6	6	8	8

Comme pour les variables qualitatives ordinales, on a calculé, dans le Tableau 1.8, les effectifs, les effectifs cumulés, les fréquences, les fréquences cumulées.

TABLE 1.8 – Tableau statistique du nombre de personnes par ménage

x_j	n_j	N_j	f_j	F_j
1	5	5	0.10	0.10
2	9	14	0.18	0.28
3	15	29	0.30	0.58
4	10	39	0.20	0.78
5	6	45	0.12	0.90
6	3	48	0.06	0.96
8	2	50	0.04	1.00
	50		1.0	

En langage R

```
> Z=c(1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,
+  4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,6,6,6,8,8)
> T4=table(Z)
> T4c=c(T4)
> data.frame(Eff=T4c,EffCum=cumsum(T4c),Freq=T4c/sum(T4c),
FreqCum=cumsum(T4c/sum(T4c)))
Eff EffCum Freq FreqCum
1  5   5  0.10  0.10
2  9   14 0.18  0.28
3 15   29 0.30  0.58
4 10   39 0.20  0.78
5  6   45 0.12  0.90
6  3   48 0.06  0.96
```

```
8  2   50 0.04  1.00
```

1.4.2 Diagramme en bâtonnets des effectifs

Quand la variable est discrète, les effectifs sont représentés par des bâtonnets (voir Figure 1.5).

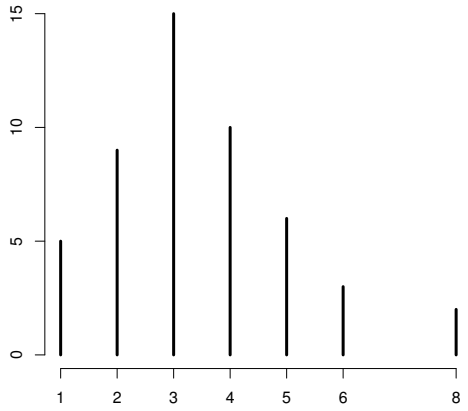


FIGURE 1.5 – Diagramme en bâtonnets des effectifs pour une variable quantitative discrète

En langage R

```
> plot(T4,type="h",xlab="",ylab="",main="",frame=0,lwd=3)
```

1.4.3 Fonction de répartition

Les fréquences cumulées sont représentées au moyen de la fonction de répartition. Cette fonction, présentée en Figure 1.6,est définie de $\mathbb{R}$ dans $[0,1]$ et vaut :

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_j & x_j \leq x < x_{j+1} \\ 1 & x_J \leq x. \end{cases}$$

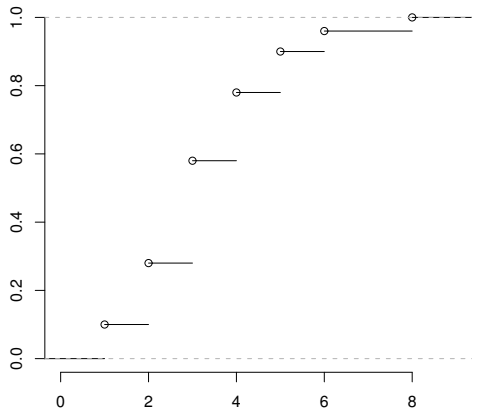


FIGURE 1.6 – Fonction de répartition d'une variable quantitative discrète

En langage R

```
> plot(ecdf(Z),xlab="",ylab="",main="",frame=0)
```

1.5 Variable quantitative continue

1.5.1 Le tableau statistique

Une variable quantitative continue peut prendre une infinité de valeurs possibles. Le domaine de la variable est alors $\mathbb{R}$ ou un intervalle de $\mathbb{R}$. En pratique, une mesure est limitée en précision. La taille peut être mesurée en centimètres, voire en millimètres. On peut alors traiter les variables continues comme des variables discrètes. Cependant, pour faire des représentations graphiques et construire le tableau statistique, il faut procéder à des regroupements en classes. Le tableau regroupé en classe est souvent appelé *distribution groupée*. Si $[c_j^-, c_j^+[$ designe la classe j, on note, de manière générale :

— c_j^- la borne inférieure de la classe j,
— c_j^+ la borne supérieure de la classe j,
— $c_j = (c_j^+ + c_j^-)/2$ le centre de la classe j,
— $a_j = c_j^+ - c_j^-$ la longueur de la classe j,
— n_j l'effectif de la classe j,
— N_j l'effectif cumulé de la classe j,
— f_j la fréquence de la classe j,
— F_j la fréquence cumulée de la classe j.

La répartition en classes des données nécessite la définition *a priori* du nombre de classes J et donc la longueur de chaque classe. En règle générale, on choisit au moins cinq classes de même longueur. Cependant, il existe des formules qui nous permettent d'établir le nombre de classes et l'intervalle de classe (la longueur) pour une série statistique de n observations.

— La règle de Sturge : $J = 1 + (3.3 \log_{10}(n))$.
— La règle de Yule : $J = 2.5\sqrt[4]{n}$.

L'intervalle de classe est obtenue ensuite de la manière suivante : longueur de l'intervalle = $(x_{max} - x_{min})/J$, où x_{max} (resp. x_{min}) désigne la plus grande (resp. la plus petite) valeur observée.

Remarque 1.2 Il faut arrondir le nombre de classe J à l'entier le plus proche. Par commodité, on peut aussi arrondir la valeur obtenue de l'intervalle de classe.

À partir de la plus petite valeur observée, on obtient les bornes de classes en additionnant successivement l'intervalle de classe (la longueur).

Exemple 1.7 Le Tableau 1.9 contient la taille en centimètres de 50 élèves d'une classe.

Table 1.9 – Taille en centimètres de 50 élèves

152	152	152	153	153	154	154	154	155	155
156	156	156	156	156	157	157	157	158	158
159	159	160	160	160	161	160	160	161	162
162	162	163	164	164	164	164	165	166	167
168	168	168	169	169	170	171	171	171	171

Les classes de taille sont données dans le Tableau 1.10.

Table 1.10 – Bornes des classes

$[151.5, 155.5[$
$[155.5, 159.5[$
$[159.5, 163.5[$
$[163.5, 167.5[$
$[167.5, 171.5[$

Le Tableau 1.11 contient la distribution groupée en classes.

En langage R

Table 1.11 – Distribution groupée

$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[151.5, 155.5[$	10	10	0.20	0.20
$[155.5, 159.5[$	12	22	0.24	0.44
$[159.5, 163.5[$	11	33	0.22	0.66
$[163.5, 167.5[$	7	40	0.14	0.80
$[167.5, 171.5[$	10	50	0.20	1.00
	50		1.00	

```
> S=c(152,152,152,153,153,154,154,154,155,155,156,156,156,156,156,
+ 157,157,157,158,158,159,159,160,160,160,161,160,160,161,162, +
162,162,163,164,164,164,164,165,166,167,168,168,168,169,169, +
170,171,171,171,171)
> T5=table(cut(S, breaks=c(151,155,159,163,167,171)))
> T5c=c(T5)
> data.frame(Eff=T5c,EffCum=cumsum(T5c),Freq=T5c/sum(T5c),
FreqCum=cumsum(T5c/sum(T5c)))
Eff EffCum Freq FreqCum
(151,155] 10   10 0.20  0.20 (155,159] 12   22 0.24 0.44
(159,163] 11   33 0.22  0.66 (163,167]  7   40 0.14 0.80
(167,171] 10   50 0.20  1.00
```

1.5.2 Histogramme

L'histogramme consiste à représenter les effectifs (resp. les fréquences) des classes par des rectangles contigus dont la surface (et non la hauteur) représente l'effectif (resp. la fréquence). Pour un histogramme des effectifs, la hauteur du rectangle correspondant à la classe j est donc donnée par :

$$h_j = \frac{n_j}{a_j}$$

— On appelle h_j la densité d'effectif.
— L'aire de l'histogramme est égale à l'effectif total n, puisque l'aire de chaque rectangle est égale à l'effectif de la classe j : $a_j \times h_j = n_j$.

Pour un histogramme des fréquences, on a

$$d_j = \frac{f_j}{a_j}$$

— On appelle d_j la densité de fréquence.
— L'aire de l'histogramme est égale à 1, puisque l'aire de chaque rectangle est égale à la fréquence de la classe j : $a_j \times d_j = f_j$.

Figure 1.7 représente l'histogramme des fréquences de l'exemple précedent :

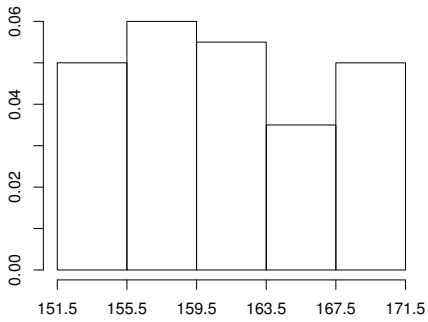


Figure 1.7 – Histogramme des fréquences

En langage R

```
> hist(S,breaks=c(151.5,155.5,159.5,163.5,167.5,171.5), freq=FALSE,
xlab="",ylab="",main="",xaxt = "n")
> axis(1, c(151.5,155.5,159.5,163.5,167.5,171.5))
```

Si les deux dernières classes sont agrégées, comme dans la Figure 1.8, la surface du dernier rectangle est égale à la surface des deux derniers rectangles de l'histogramme de la Figure 1.7.

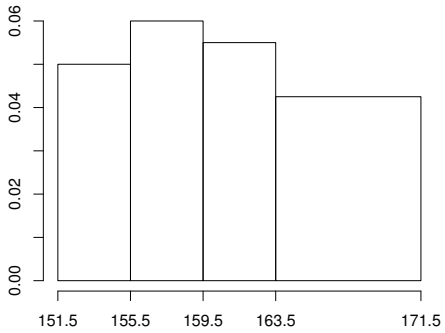


FIGURE 1.8 – Histogramme des fréquences avec les deux dernières classes agrégées

En langage R

```
> hist(S,breaks=c(151.5,155.5,159.5,163.5,171.5),
xlab="",ylab="",main="",xaxt = "n")
> axis(1, c(151.5,155.5,159.5,163.5,171.5))
```

Remarque 1.3 Dans le cas de classes de mêmes longueurs, certains auteurs et logiciels représentent l'histogramme avec les effectifs (resp. les fréquences) reportés en ordonnée, l'aire de chaque rectangle étant proportionnelle à l'effectif (resp. la fréquence) de la classe.

1.5.3 La fonction de répartition

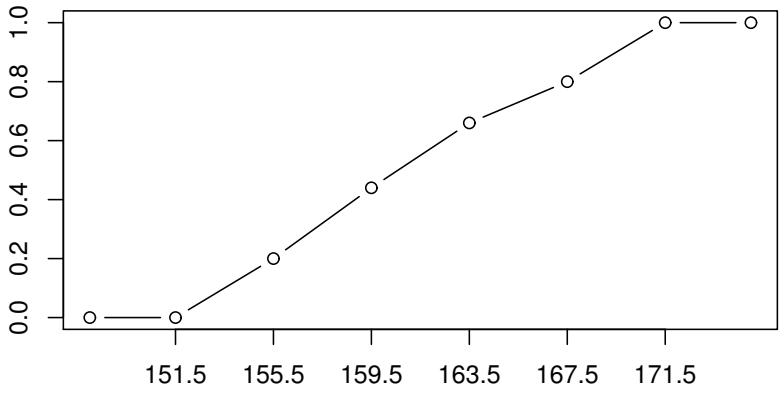
La fonction de répartition $F(x)$ est une fonction de $\mathbb{R}$ dans $[0,1]$, qui est définie par

$$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_J^+ \leq x \end{cases}$$

En langage R

```
> y=c(0,0,cumsum(T5c/sum(T5c)),1)
> x=c(148,151.5,155.5,159.5,163.5,167.5,171.5,175)
> plot(x,y,type="b",xlab="",ylab="",xaxt = "n")
> axis(1, c(151.5,155.5,159.5,163.5,167.5,171.5))
```

FIGURE 1.9 – Fonction de répartition d'une distribution groupée



Chapitre 2

Statistique descriptive univariée

2.1 Paramètres de position

2.1.1 Le mode

Le mode est la valeur distincte correspondant à l'effectif le plus élevé. Il est noté x_M. Si on reprend la variable 'État civil', dont la distribution est donnée dans le Tableau 2.1, le mode est C : célibataire.

Table 2.1 – Tableau statistique

x_j	n_j	f_j
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10
	$n = 20$	1

Remarque 2.1
- Le mode peut être calculé pour tous les types de variable, quantitative et qualitative.
- Le mode n'est pas nécessairement unique.
- Quand une variable continue est découpée en classes, on peut définir une classe modale (classe correspondant à l'effectif le plus élevé).

2.1.2 La moyenne arithmétique

La *moyenne* ne peut être définie que sur une variable *quantitative*.
La moyenne est la somme des valeurs observées divisée par leur nombre, elle est notée $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La moyenne peut être calculée à partir des valeurs distinctes et des effectifs

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j.$$

Exemple 2.1 Les nombres d'enfants de 8 familles sont les suivants $0, 0, 1, 1, 1, 2, 3, 4$. La moyenne est

$$\bar{x} = \frac{0+0+1+1+1+2+3+4}{8} = \frac{12}{8} = 1.5.$$

On peut aussi faire les calculs avec les valeurs distinctes et les effectifs. On considère le Tableau 2.2.

TABLE 2.2 – Tableau des valeurs distinctes

x_j	n_j
0	2
1	3
2	1
3	1
4	1
	8

$$\bar{x} = \frac{2\times 0+3\times 1+1\times 2+1\times 3+1\times 4}{8} = \frac{3+2+3+4}{8} = 1.5.$$

Remarque 2.2 La moyenne n'est pas nécessairement une valeur possible.

En langage R

```
E=c(0,0,1,1,1,2,3,4)
n=length(E)
xb=sum(E)/n
xb
xb=mean(E)
xb
```

2.1.3 Remarques sur le signe de sommation $\sum$

Définition 2.1.

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

1. En statistique les x_i sont souvent les valeurs observées.
2. L'indice est muet : $\displaystyle\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$.
3. Quand il n'y a pas de confusion possible, on peut écrire $\sum_{i=1}^{n} x_i$.

Exemple 2.2

1. $\displaystyle\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$.
2. $\displaystyle\sum_{i=3}^{5} x_{i2} = x_{32} + x_{42} + x_{52}$.
3. $\displaystyle\sum_{i=1}^{3} i = 1 + 2 + 3 = 6$.
4. On peut utiliser plusieurs sommations emboîtées, mais il faut bien distinguer les indices :

$$\begin{aligned}\sum_{i=1}^{3}\sum_{j=1}^{2} x_{ij} &= x_{11} + x_{12} && (i=1)\\ &+ x_{21} + x_{22} && (i=2)\\ &+ x_{31} + x_{32} && (i=3)\end{aligned}$$

5. On peut exclure une valeur de l'indice :

$$\sum_{\substack{i=1\\ i\neq 3}}^{5} x_i = x_1 + x_2 + x_4 + x_5.$$

Propriété 2.1.

1. Somme d'une constante

$$\sum_{i=1}^{n} a = \underbrace{a + a + \cdots + a}_{n \text{ fois}} = na \qquad \text{(a constante).}$$

   Exemple

$$\sum_{i=1}^{5} 3 = 3 + 3 + 3 + 3 + 3 = 5 \times 3 = 15.$$

2. Mise en évidence

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i \qquad \text{(a constante).}$$

   Exemple

$$\sum_{i=1}^{3} 2 \times i = 2(1 + 2 + 3) = 2 \times 6 = 12.$$

3. Somme des n premiers entiers

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

4. Distribution

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

5. Distribution

$$\sum_{i=1}^{n} (x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i.$$

   Exemple (avec $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$)

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n \frac{1}{n} \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

6. Somme de carrés

$$\sum_{i=1}^{n} (x_i - y_i)^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i y_i + y_i^2) = \sum_{i=1}^{n} x_i^2 - 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$$

   C'est une application de la formule

$$(a - b)^2 = a^2 - 2ab + b^2.$$

2.1.4 Moyenne géométrique

Si $x_i \geq 0$, on appelle moyenne géométrique la quantité

$$G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (x_1 \times x_2 \times \cdots \times x_n)^{1/n}.$$

On peut écrire la moyenne géométrique comme l'exponentielle de la moyenne arithmétique des logarithmes des valeurs observées

$$G = \exp \log G = \exp \log \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \exp \frac{1}{n} \log \prod_{i=1}^{n} x_i = \exp \frac{1}{n} \sum_{i=1}^{n} \log x_i.$$

La moyenne géométrique s'utilise, par exemple, quand on veut calculer la moyenne de taux d'intérêt.

Exemple 2.3 Supposons que les taux d'intérêt pour 4 années consécutives soient respectivement de 5, 10, 15, et 10%. Que va-t-on obtenir après 4 ans si je place 100 francs ?

— Après 1 an on a, $100 \times 1.05 = 105$ Fr.
— Après 2 ans on a, $100 \times 1.05 \times 1.1 = 115.5$ Fr.
— Après 3 ans on a, $100 \times 1.05 \times 1.1 \times 1.15 = 132.825$ Fr.
— Après 4 ans on a, $100 \times 1.05 \times 1.1 \times 1.15 \times 1.1 = 146.1075$ Fr.

Si on calcule la moyenne arithmétique des taux on obtient

$$\bar{x} = \frac{1.05 + 1.10 + 1.15 + 1.10}{4} = 1.10.$$

Si on calcule la moyenne géométrique des taux, on obtient

$$G = (1.05 \times 1.10 \times 1.15 \times 1.10)^{1/4} = 1.099431377.$$

Le bon taux moyen est bien G et non $\bar{x}$, car si on applique 4 fois le taux moyen G aux 100 francs, on obtient

$$100 \text{ Fr} \times G^4 = 100 \times 1.099431377^4 = 146.1075 \text{ Fr}.$$

2.1.5 Moyenne harmonique

Si $x_i \geq 0$, on appelle moyenne harmonique la quantité

$$H = \frac{n}{\sum_{i=1}^{n} 1/x_i}.$$

Il est judicieux d'appliquer la moyenne harmonique sur des vitesses.

Exemple 2.4 Un cycliste parcourt 4 étapes de 100km. Les vitesses respectives pour ces étapes sont de 10 km/h, 30 km/h, 40 km/h, 20 km/h. Quelle a été sa vitesse moyenne ?

— Un raisonnement simple nous dit qu'il a parcouru la première étape en 10h, la deuxième en 3h20 la troisième en 2h30 et la quatrième en 5h. Il a donc parcouru le total des 400km en

$$10 + 3h20 + 2h30 + 5h = 20h50 = 20.8333h,$$

sa vitesse moyenne est donc

$$\text{Moy} = \frac{400}{20.8333} = 19.2 \text{ km/h}.$$

— Si on calcule la moyenne arithmétique des vitesses, on obtient

$$\bar{x} = \frac{10 + 30 + 40 + 20}{4} = 25 \text{ km/h}.$$

— Si on calcule la moyenne harmonique des vitesses, on obtient

$$H = \frac{4}{\frac{1}{10} + \frac{1}{30} + \frac{1}{40} + \frac{1}{20}} = 19.2 \text{ km/h}.$$

La moyenne harmonique est donc la manière appropriée de calculer la vitesse moyenne.

Remarque 2.3 Il est possible de montrer que la moyenne harmonique est toujours inférieure ou égale à la moyenne géométrique qui est toujours inférieure ou égale à la moyenne arithmétique

$$H \leq G \leq \bar{x}.$$

2.1.6 Moyenne pondérée

Dans certains cas, on n'accorde pas le même poids à toutes les observations. Par exemple, si on calcule la moyenne des notes pour un programme d'étude, on peut pondérer les notes de l'étudiant par le nombre de crédits ou par le nombre d'heures de chaque cours. Si $w_i > 0, i = 1, \ldots, n$ sont les poids associés à chaque observation, alors la moyenne pondérée par w_i est définie par :

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

Exemple 2.5 Supposons que les notes soient pondérées par le nombre de crédits et que les notes de l'étudiant soient dans le Tableau 2.3. La moyenne pondérée des notes par les crédits est alors

Table 2.3 – Notes d'un étudiant

Note	5	4	3	6	5
Crédits	6	3	4	3	4

$$\bar{x}_w = \frac{6 \times 5 + 3 \times 4 + 4 \times 3 + 3 \times 6 + 4 \times 5}{6 + 3 + 4 + 3 + 4} = \frac{30 + 12 + 12 + 18 + 20}{20} = \frac{92}{20} = 4.6.$$

2.1.7 La médiane

La médiane, notée $x_{1/2}$, est une valeur centrale de la série statistique obtenue de la manière suivante :

— On trie la série statistique par ordre croissant des valeurs observées. Avec la série observée :

$$3 \quad 2 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2,$$

on obtient :

$$0 \quad 0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3.$$

— La médiane $x_{1/2}$ est la valeur qui se trouve au milieu de la série ordonnée :

$$\begin{array}{ccccccc} 0 & 0 & 1 & 1 & 2 & 2 & 3. \\ & & & \uparrow & & & \end{array}$$

On note alors $x_{1/2} = 1$.

Nous allons examiner une manière simple de calculer la médiane. Deux cas doivent être distingués.

— Si n est impair, il n'y a pas de problème (ici avec $n = 7$), alors $x_{1/2} = 1$:

$$\begin{array}{ccccccc} 0 & 0 & 1 & 1 & 2 & 2 & 3. \\ & & & \uparrow & & & \end{array}$$

La Figure 2.1 montre la fonction de répartition de la série. La médiane peut être définie comme l'inverse de la fonction de répartition pour la valeur $1/2$:

$$x_{1/2} = F^{-1}(0.5).$$

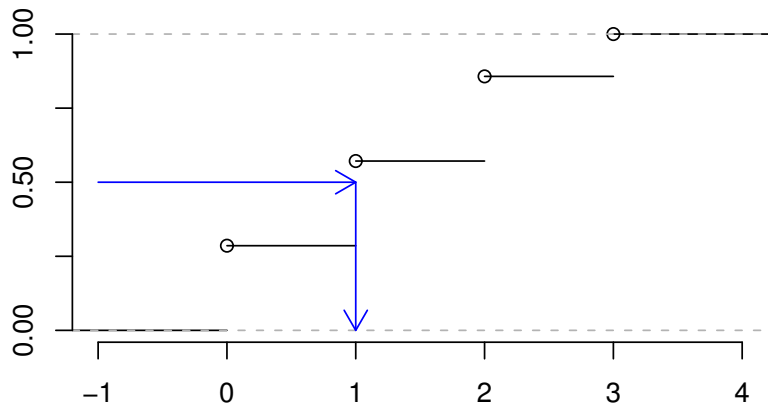


FIGURE 2.1 – Médiane quand n est impair

En langage R

```
x=c(0, 0, 1, 1, 2, 2, 3)
median(x)
plot(ecdf(x),xlab="",ylab="",main="",frame=FALSE,yaxt = "n")
axis(2, c(0.0,0.25,0.50,0.75,1.00))
arrows(-1,0.5,1,0.50,length=0.14,col="blue")
arrows(1,0.50,1,0,length=0.14,col="blue")
```

— Si n est pair, deux valeurs se trouvent au milieu de la série (ici avec $n = 8$)

$$\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 2 & 2 & 3 & 4 \\ & & & \uparrow & \uparrow & & & \end{array}$$

La médiane est alors la moyenne de ces deux valeurs :

$$x_{1/2} = \frac{1+2}{2} = 1.5.$$

La Figure 2.2 montre la fonction de répartition de la série de taille paire. La médiane peut toujours être définie comme l'inverse de la fonction de répartition pour la valeur $1/2$:

$$x_{1/2} = F^{-1}(0.5).$$

Cependant, la fonction de répartition est discontinue par 'palier'. L'inverse de la répartition correspond exactement à un 'palier'.

En langage R

```
x=c(0, 0, 1, 1, 2, 2, 3, 4)
median(x)
plot(ecdf(x),xlab="",ylab="",main="",frame=FALSE,yaxt = "n")
axis(2, c(0.0,0.25,0.50,0.75,1.00))
arrows(-1,0.5,1,0.50,length=0.14,col="blue")
arrows(1.5,0.50,1.5,0,,length=0.14,col="blue")
```

En général on note

$$x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(n)}$$

la série ordonnée par ordre croissant. On appelle cette série ordonnée la statistique d'ordre. Cette notation, très usuelle en statistique, permet de définir la médiane de manière très synthétique.

— Si n est impair

$$x_{1/2} = x_{(\frac{n+1}{2})}$$

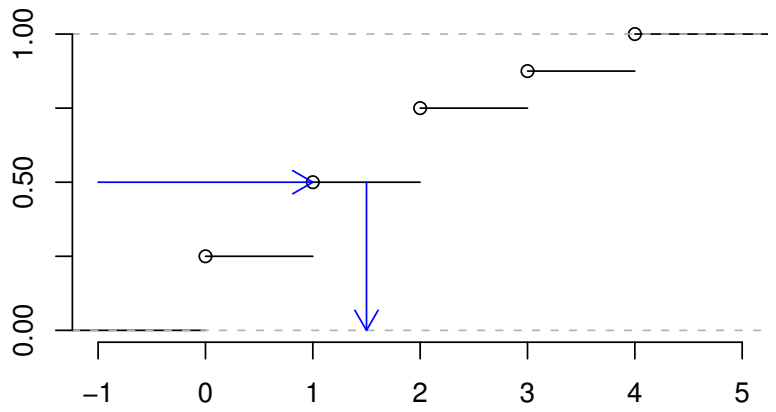


FIGURE 2.2 – Médiane quand n est pair

— Si n est pair

$$x_{1/2} = \frac{1}{2}\left\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right\}.$$

Remarque 2.4 La médiane peut être calculée sur des variables quantitatives et sur des variables qualitatives ordinales.

2.1.8 Quantiles

La notion de quantile d'ordre p (où $0 < p < 1$) généralise la médiane. Formellement un quantile est donné par l'inverse de la fonction de répartition :

$$x_p = F^{-1}(p).$$

Si la fonction de répartition était continue et strictement croissante, la définition du quantile serait sans équivoque. La fonction de répartition est cependant discontinue et "par palier". Quand la fonction de répartition est par palier, il existe au moins 9 manières différentes de définir les quantiles selon que l'on fasse ou non une interpolation de la fonction de répartition. Nous présentons une de ces méthodes, mais il ne faut pas s'étonner de voir les valeurs des quantiles différer légèrement d'un logiciel statistique à l'autre.

— Si np est un nombre entier, alors

$$x_p = \frac{1}{2}\left\{x_{(np)} + x_{(np+1)}\right\}.$$

— Si np n'est pas un nombre entier, alors

$$x_p = x_{(\lceil np \rceil)},$$

où $\lceil np \rceil$ représente le plus petit nombre entier supérieur ou égal à np.

Remarque 2.5

— La médiane est le quantile d'ordre $p = 1/2$.

— On utilise souvent

$x_{1/4}$	le premier quartile,
$x_{3/4}$	le troisième quartile,
$x_{1/10}$	le premier décile,
$x_{1/5}$	le premier quintile,
$x_{4/5}$	le quatrième quintile,
$x_{9/10}$	le neuvième décile,
$x_{0.05}$	le cinquième percentile,
$x_{0.95}$	le nonante-cinquième percentile.

— Si $F(x)$ est la fonction de répartition, alors $F(x_p) \geq p$.

Exemple 2.6 Soit la série statistique 12, 13, 15, 16, 18, 19, 22, 24, 25, 27, 28, 34 contenant 12 observations ($n = 12$).

— Le premier quartile : Comme $np = 0.25 \times 12 = 3$ est un nombre entier, on a

$$x_{1/4} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{15 + 16}{2} = 15.5.$$

— La médiane : Comme $np = 0.5 \times 12 = 6$ est un nombre entier, on a

$$x_{1/2} = \frac{1}{2}\left\{x_{(6)} + x_{(7)}\right\} = (19 + 22)/2 = 20.5.$$

— Le troisième quartile : Comme $np = 0.75 \times 12 = 9$ est un nombre entier, on a

$$x_{3/4} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{25 + 27}{2} = 26.$$

En langage R

```
x=c(12,13,15,16,18,19,22,24,25,27,28,34)
quantile(x,type=2)
```

Exemple 2.7 Soit la série statistique 12, 13, 15, 16, 18, 19, 22, 24, 25, 27 contenant 10 observations ($n = 10$).

— Le premier quartile : Comme $np = 0.25 \times 10 = 2.5$ n'est pas un nombre entier, on a

$$x_{1/4} = x_{(\lceil 2.5 \rceil)} = x_{(3)} = 15.$$

— La médiane : Comme $np = 0.5 \times 10 = 5$ est un nombre entier, on a

$$x_{1/2} = \frac{1}{2}\left\{x_{(5)} + x_{(6)}\right\} = (18 + 19)/2 = 18.5.$$

— Le troisième quartile : Comme $np = 0.75 \times 10 = 7.5$ n'est pas un nombre entier, on a

$$x_{3/4} = x_{(\lceil 7.5 \rceil)} = x_{(8)} = 24.$$

En langage R

```
x=c(12,13,15,16,18,19,22,24,25,27)
quantile(x,type=2)
```

2.2 Paramètres de dispersion

2.2.1 L'étendue

L'*étendue* est simplement la différence entre la plus grande et la plus petite valeur observée.

$$\mathrm{E} = x_{(n)} - x_{(1)}.$$

2.2.2 La distance interquartile

La distance interquartile est la différence entre le troisième et le premier quartile :

$$IQ = x_{3/4} - x_{1/4}.$$

2.2.3 La variance

La *variance* est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Théorème 2.1. *La variance peut aussi s'écrire*

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{2.1}$$

Démonstration.

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} x_i\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.
\end{aligned}$$

□

La variance peut également être définie à partir des effectifs et des valeurs distinctes :

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j(x_j - \bar{x})^2.$$

La variance peut aussi s'écrire

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2.$$

Quand on veut estimer une variance d'une variable X à partir d'un échantillon (une partie de la population sélectionnée au hasard) de taille n, on utilise la variance "corrigée" divisée par $n-1$.

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2\frac{n}{n-1}.$$

La plupart des logiciels statistiques calculent S_x^2 et non s_x^2.

2.2.4 L'écart-type

L'*écart-type* est la racine carrée de la variance :

$$s_x = \sqrt{s_x^2}.$$

Quand on veut estimer l'écart-type d'une variable X partir d'un échantillon de taille n, utilise la variance "corrigée" pour définir l'écart type

$$S_x = \sqrt{S_x^2} = s_x\sqrt{\frac{n}{n-1}}.$$

La plupart des logiciels statistiques calculent S_x et non s_x.

Exemple 2.8 Soit la série statistique $2, 3, 4, 4, 5, 6, 7, 9$ de taille 8. On a

$$\bar{x} = \frac{2+3+4+4+5+6+7+9}{8} = 5,$$

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{8}\left[(2-5)^2 + (3-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2 + (9-5)^2\right] \\
&= \frac{1}{8}[9+4+1+1+0+1+4+16] = \frac{36}{8} = 4.5.
\end{aligned}$$

On peut également utiliser la formule (2.1) de la variance, ce qui nécessite moins de calcul (surtout quand la moyenne n'est pas un nombre entier).

$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{8}(2^2+3^2+4^2+4^2+5^2+6^2+7^2+9^2) - 5^2 \\ &= \frac{1}{8}(4+9+16+16+25+36+49+81) - 25 = \frac{236}{8} - 25 = 29.5 - 25 = 4.5. \end{aligned}$$

En langage R

```
> x=c(2,3,4,4,5,6,7,9)
> n=length(x)
> s2=sum((x-mean(x))^2)/n
> s2
[1] 4.5
> S2=s2*n/(n-1)
> S2
[1] 5.142857
> S2=var(x)
> S2
[1] 5.142857
> s=sqrt(s2)
> s
[1] 2.121320
> S=sqrt(S2)
> S
[1] 2.267787
> S=sd(x)
> S
[1] 2.267787
> E=max(x)-min(x)
> E
[1] 7
```

2.2.5 L'écart moyen absolu

L'*écart moyen absolu* est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre d'observations :

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| .$$

2.2.6 L'écart médian absolu

L'*écart médian absolu* est la somme des valeurs absolues des écarts à la médiane divisée par le nombre d'observations :

$$e_{med} = \frac{1}{n}\sum_{i=1}^{n} \left|x_i - x_{1/2}\right| .$$

2.3 Moments

Définition 2.2. *On appelle moment à l'origine d'ordre $r \in \mathbb{N}$ le paramètre*

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r .$$

Définition 2.3. *On appelle moment centré d'ordre $r \in \mathbb{N}$ le paramètre*

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r.$$

Les moments généralisent la plupart des paramètres. On a en particulier
— $m'_1 = \bar{x}$,
— $m_1 = 0$,
— $m'_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = s_x^2 + \bar{x}^2$,
— $m_2 = s_x^2$.
Nous verrons plus loin que des moments d'ordres supérieurs ($r = 3, 4$) sont utilisés pour mesurer la symétrie et l'aplatissement.

2.4 Paramètres de forme

2.4.1 Coefficient d'asymétrie de Fisher (skewness)

Le *moment centré d'ordre trois* est défini par

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Il peut prendre des valeurs positives, négatives ou nulles. L'asymétrie se mesure au moyen du coefficient d'asymétrie de Fisher

$$g_1 = \frac{m_3}{s_x^3},$$

où s_x^3 est le cube de l'écart-type.

2.4.2 Coefficient d'asymétrie de Yule

Le coefficient d'asymétrie de Yule est basé sur les positions des 3 quartiles (1er quartile, médiane et troisième quartile) et est normalisé par la distance interquartile :

$$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}}.$$

2.4.3 Coefficient d'asymétrie de Pearson

Le coefficient d'asymétrie de Pearson est basé sur une comparaison de la moyenne et du mode et est standardisé par l'écart-type :

$$A_P = \frac{\bar{x} - x_M}{s_x}.$$

Tous les coefficients d'asymétrie ont les mêmes propriétés, ils sont nuls si la distribution est symétrique, négatifs si la distribution est allongée à gauche (left asymmetry) et positifs si la distribution est allongée à droite (right asymmetry) comme montré dans la Figure 2.3.

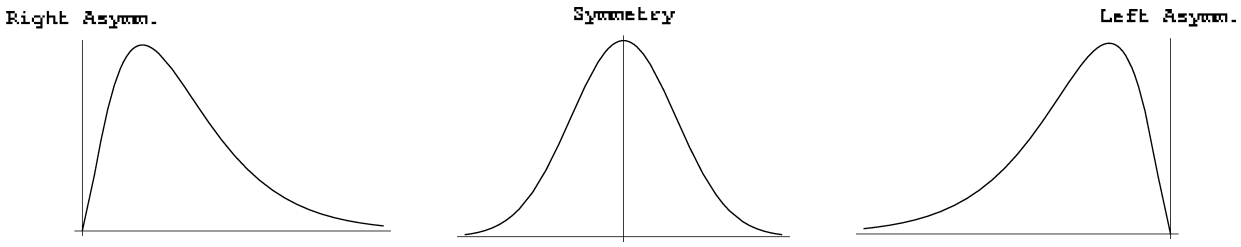


FIGURE 2.3 – Asymétrie d'une distribution

Remarque 2.6 Certaines variables sont toujours très asymétriques à droite, comme les revenus, les tailles des entreprises, ou des communes. Une méthode simple pour rendre une variable symétrique consiste alors à prendre le logarithme de cette variable.

2.5 Paramètre d'aplatissement (kurtosis)

L'aplatissement est mesuré par le coefficient d'aplatissement de Pearson

$$\beta_2 = \frac{m_4}{s_x^4},$$

ou le coefficient d'aplatissement de Fisher

$$g_2 = \beta_2 - 3 = \frac{m_4}{s_x^4} - 3,$$

où m_4 est le moment centré d'ordre 4 et s_x^4 est le carré de la variance.

— Une courbe mésokurtique si $g_2 \approx 0$.
— Une courbe leptokurtique si $g_2 > 0$. Elle est plus pointue et possède des queues plus longues.
— Une courbe platykurtique si $g_2 < 0$. Elle est plus arrondie et possède des queues plus courtes.

Dans la Figure 2.4, on présente un exemple de deux distributions de même moyenne et de même variance. La distribution plus pointue est leptokurtique, l'autre est mésokurtique. La distribution leptokurtique a une queue plus épaisse.

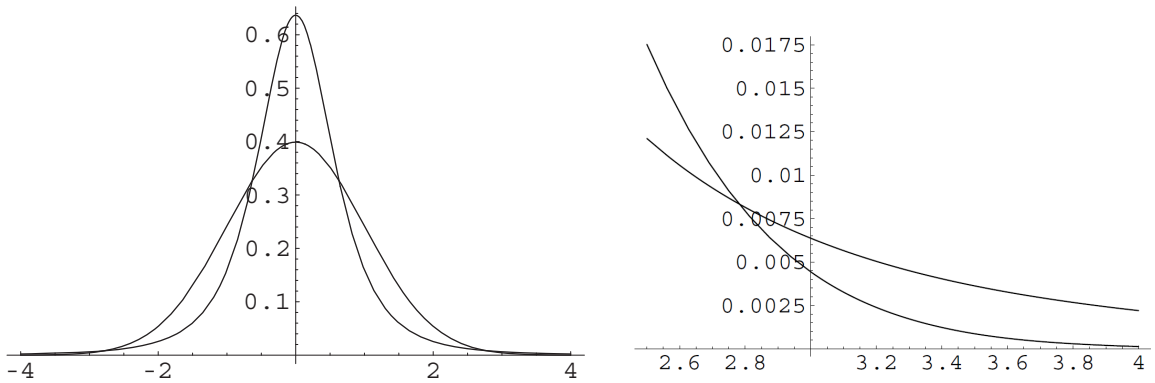


FIGURE 2.4 – Distributions mésokurtique et leptokurtique

2.6 Changement d'origine et d'unité

Définition 2.4. *On appelle changement d'origine l'opération consistant à ajouter (ou soustraire) la même quantité $a \in \mathbb{R}$ à toutes les observations*

$$y_i = a + x_i, i = 1, \ldots, n$$

Définition 2.5. *On appelle changement d'unité l'opération consistant à multiplier (ou diviser) par la même quantité $b \in \mathbb{R}$ toutes les observations*

$$y_i = b\ x_i, i = 1, \ldots, n.$$

Définition 2.6. *On appelle changement d'origine et d'unité l'opération consistant à multiplier toutes les observations par la même quantité $b \in \mathbb{R}$ puis à ajouter la même quantité $a \in \mathbb{R}$ à toutes les observations :*

$$y_i = a + b\ x_i, i = 1, \ldots, n.$$

Théorème 2.2. *Si on effectue un changement d'origine et d'unité sur une variable X, alors sa moyenne est affectée du même changement d'origine et d'unité.*

Démonstration. Si $y_i = a + b\ x_i$, alors

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(a + b\ x_i) = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\ \bar{x}.$$

□

Théorème 2.3. *Si on effectue un changement d'origine et d'unité sur une variable X, alors sa variance est affectée par le carré du changement d'unité et pas par le changement d'origine.*

Démonstration. Si $y_i = a + b\ x_i$, alors

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(a + b\ x_i - a - b\ \bar{x})^2 = b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 s_x^2.$$

□

Remarque 2.7

1. Les paramètres de position sont tous affectés par un changement d'origine et d'unité.
2. Les paramètres de dispersion sont tous affectés par un changement d'unité mais pas par un changement d'origine.
3. Les paramètres de forme et d'aplatissement ne sont affectés ni par un changement d'unité ni par un changement d'origine.

2.7 Moyennes et variances dans des groupes

Supposons que les n observations soient réparties dans deux groupes G_A et G_B. Les n_A premières observations sont dans le groupe G_A et les n_B dernières observations sont dans le groupe G_B, avec la relation

$$n_A + n_B = n.$$

On suppose que la série statistique contient d'abord les unités de G_A puis les unités de G_B :

$$\underbrace{x_1, x_2, \ldots, x_{n_A-1}, x_{n_A}}_{\text{observations de } G_A}, \underbrace{x_{n_A+1}, x_{n_A+2}, \ldots, x_{n-1}, x_n}_{\text{observations de } G_B}.$$

On définit les moyennes des deux groupes :

— la moyenne du premier groupe $\bar{x}_A = \dfrac{1}{n_A}\displaystyle\sum_{i=1}^{n_A} x_i$,

— la moyenne du deuxième groupe $\bar{x}_B = \dfrac{1}{n_B}\displaystyle\sum_{i=n_A+1}^{n} x_i$.

La moyenne générale est une moyenne pondérée par la taille des groupes des moyennes des deux groupes. En effet,

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n_A} x_i + \sum_{i=n_A+1}^{n} x_i\right) = \frac{1}{n}\left(n_A\bar{x}_A + n_B\bar{x}_B\right).$$

On peut également définir les variances des deux groupes :

— la variance du premier groupe $s_A^2 = \dfrac{1}{n_A}\displaystyle\sum_{i=1}^{n_A}(x_i - \bar{x}_A)^2$,

— la variance du deuxième groupe $s_B^2 = \dfrac{1}{n_B}\displaystyle\sum_{i=n_A+1}^{n}(x_i - \bar{x}_B)^2$.

Théorème 2.4. *(de la variance totale) La variance totale, définie par*

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

se décompose de la manière suivante :

$$s_x^2 = \underbrace{\frac{n_A s_A^2 + n_B s_B^2}{n}}_{\textit{variance intra-groupes}} + \underbrace{\frac{n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2}{n}}_{\textit{variance inter-groupes}}.$$

Démonstration.

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left\{\sum_{i=1}^{n_A}(x_i - \bar{x})^2 + \sum_{i=n_A+1}^{n}(x_i - \bar{x})^2\right\} \tag{2.2}$$

On note que

$$\begin{aligned}
&\sum_{i=1}^{n_A}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n_A}(x_i - \bar{x}_A + \bar{x}_A - \bar{x})^2 \\
&= \sum_{i=1}^{n_A}(x_i - \bar{x}_A)^2 + \sum_{i=1}^{n_A}(\bar{x}_A - \bar{x})^2 + \underbrace{2\sum_{i=1}^{n_A}(x_i - \bar{x}_A)(\bar{x}_A - \bar{x})}_{=0} \\
&= n_A s_A^2 + n_A(\bar{x}_A - \bar{x})^2.
\end{aligned}$$

On a évidemment la même relation dans le groupe G_B :

$$\sum_{i=n_A+1}^{n}(x_i - \bar{x})^2 = n_B s_B^2 + n_B(\bar{x}_B - \bar{x})^2.$$

En revenant à l'Expression (2.2), on obtient

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\left\{\sum_{i=1}^{n_A}(x_i - \bar{x})^2 + \sum_{i=n_A+1}^{n}(x_i - \bar{x})^2\right\} \\
&= \frac{1}{n}\left\{n_A s_A^2 + n_A(\bar{x}_A - \bar{x})^2 + n_B s_B^2 + n_B(\bar{x}_B - \bar{x})^2\right\} \\
&= \frac{n_A s_A^2 + n_B s_B^2}{n} + \frac{n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2}{n}.
\end{aligned}$$

□

2.8 Diagramme en tiges et feuilles

Le diagramme en tiges et feuilles ou *Stem and leaf diagram* est une manière rapide de présenter une variable quantitative. Par exemple, si l'on a la série statistique ordonnée suivante :

$$15, 15, 16, 17, 18, 20, 21, 22, 23, 23, 23, 24, 25, 25, 26, 26, 27, 28, 28, 29, 30, 30, 32, 34, 35, 36, 39, 40, 43, 44,$$

la tige du diagramme sera les dizaines et les feuilles seront les unités. On obtient le graphique suivant.

```
The decimal point is 1 digit(s) to the right of the |

1 | 55678
2 | 012333455667889
3 | 0024569
4 | 034
```

Ce diagramme permet d'avoir une vue synthétique de la distribution. Évidemment, les tiges peuvent être définies par les centaines, ou des millers, selon l'ordre de grandeur de la variable étudiée.

En langage R

```
#
# Diagramme en tige et feuilles
#
X=c(15,15,16,17,18,20,21,22,23,23,23,24,25,25,26,26,
27,28,28,29,30,30,32,34,35,36,39,40,43,44)
stem(X,0.5)
```

2.9 La boîte à moustaches

La boîte à moustaches, ou diagramme en boîte, ou encore *boxplot* en anglais, est un diagramme simple qui permet de représenter la distribution d'une variable. Ce diagramme est composé de :

- Un rectangle qui s'étend du premier au troisième quartile. Le rectangle est divisé par une ligne correspondant à la médiane.
- Ce rectangle est complété par deux segments de droites.
  - Pour les dessiner, on calcule d'abord les bornes

$$b^- = x_{1/4} - 1.5IQ \quad \text{et} \quad b^+ = x_{3/4} + 1.5IQ,$$

    où IQ est la distance interquartile.
  - On identifie ensuite la plus petite et la plus grande observation comprise entre ces bornes. Ces observations sont appelées "valeurs adjacentes".
  - On trace les segments de droites reliant ces observations au rectangle.
  - Les valeurs qui ne sont pas comprises entre les valeurs adjacentes, sont représentées par des points et sont appelées "valeurs extrêmes".

Exemple 2.9 On utilise une base de données de communes suisses de 2003 fournie par l'Office fédéral de la statistique (OFS) contenant un ensemble de variables concernant la population et l'aménagement du territoire. L'objectif est d'avoir un aperçu des superficies des communes du canton de Neuchâtel. On s'intéresse donc à la variable HApoly donnant la superficie en hectares des 62 communes neuchâteloises. La boîte à moustaches est présentée en Figure 2.5. L'examen du graphique indique directement une dissymétrie de la distribution, au sens où il y a beaucoup de petites communes et peu de grandes communes. Le graphique montre aussi que deux communes peuvent être considérées communes des points extrêmes, car elles ont plus de 3000 hectares. Il s'agit de la Brévine (4182ha) et de la Chaux-de-Fonds (5566ha).

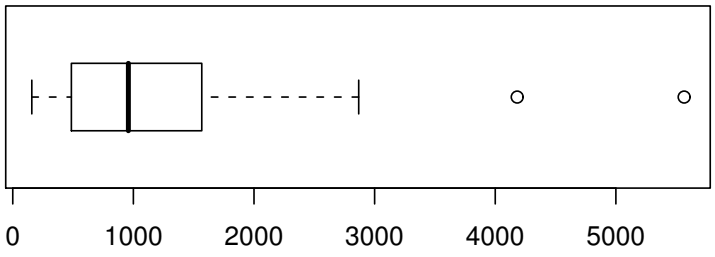


FIGURE 2.5 – Boîtes à moustaches pour la variable superficie en hectares (HApoly) des communes du canton de Neuchâtel

En langage R

```
# Étape 1 : installation du package sampling
#     dans lequel se trouve la base de données des communes belges
install.packages("sampling")
# Etape 2 : chargement du package sampling
library(sampling)
# Utilisation des données
data(swissmunicipalities)
attach(swissmunicipalities)
# boxplot de la sélection des communes neuchâteloises
# le numéro du canton est 24
boxplot(HApoly[CT==24],horizontal=TRUE)
# selection des communes neuchâteloises de plus de 3000 HA
data.frame(Nom=Nom[HApoly>3000 & CT==24],Superficie=HApoly[HApoly>3000 & CT==24])
```

Exemple 2.10 On utilise une base de données belges fournie par l'Institut National (belge) de Statistique contenant des informations sur la population et les revenus des personnes physiques dans les communes. On s'intéresse à la variable "revenu moyen en euros par habitant en 2004" pour chaque commune (variable averageincome) et l'on aimerait comparer les 9 provinces belges : Anvers, Brabant, Flandre occidentale, Flandre orientale, Hainaut, Liège, Limbourg, Luxembourg, Namur. La Figure 2.6 contient les boîtes à moustaches de chaque province. Les communes ont été triées selon les provinces belges. De ce graphique, on peut directement voir que la province du Brabant contient à la fois la commune la plus riche (Lasne) et la plus pauvre (Saint-Josse-ten-Noode). On voit également une dispersion plus importante dans la province du Brabant.

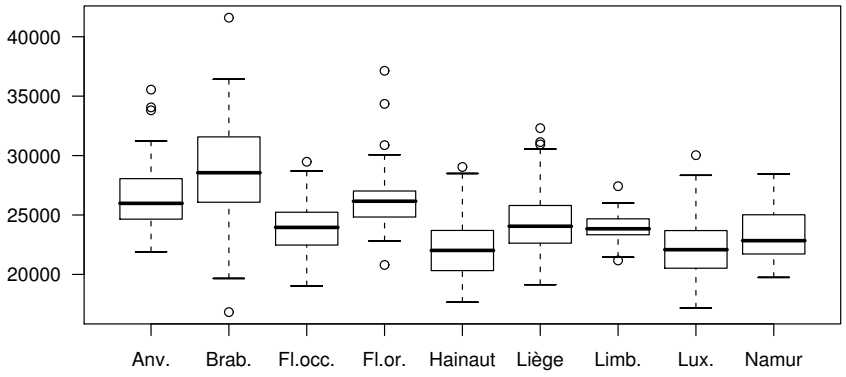


FIGURE 2.6 – Boîtes à moustaches du "revenu moyen des habitants" des communes selon les provinces belges

En langage R

```
# Utilisation des données
data(belgianmunicipalities)
attach(belgianmunicipalities)
# Construction d'une liste avec les noms des provinces
b=list("Anv."=averageincome[Province==1],
"Brab."=averageincome[Province==2],
"Fl.occ."=averageincome[Province==3],
"Fl.or."=averageincome[Province==4],
"Hainaut"=averageincome[Province==5],
"Liège"=averageincome[Province==6],
"Limb."=averageincome[Province==7],
"Lux."=averageincome[Province==8],
"Namur"=averageincome[Province==9])
boxplot(b)
```

Chapitre 3

Statistique descriptive bivariée

3.1 Série statistique bivariée

On s'intéresse à deux variables x et y. Ces deux variables sont mesurées sur les n unités d'observation. Pour chaque unité, on obtient donc deux mesures. La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu :

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Chacune des deux variables peut être, soit quantitative, soit qualitative. On examine deux cas.

— Les deux variables sont quantitatives.
— Les deux variables sont qualitatives.

3.2 Deux variables quantitatives

3.2.1 Représentation graphique de deux variables

Dans ce cas, chaque couple est composé de deux valeurs numériques. Un couple de nombres (entiers ou réels) peut toujours être représenté comme un point dans un plan

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Exemple 3.1 Le Tableau 3.1 contient le poids Y et la taille X de 20 individus. Le nuage de points est représenté dans la Figure 3.1.

Table 3.1 – Poids Y et taille X de 20 individus

y_i	60	61	64	67	68	69	70	70	72	73	75	76	78	80	85	90	96	96	98	101
x_i	155	162	157	170	164	162	169	170	178	173	180	175	173	175	179	175	180	185	189	187

En langage R

```
# nuage de points
poids=c(60,61,64,67,68,69,70,70,72,73,75,76,78,80,85,90,96,96,98,101)
taille=c(155,162,157,170,164,162,169,170,178,173,
180,175,173,175,179,175,180,185,189,187)
plot(taille,poids)
```

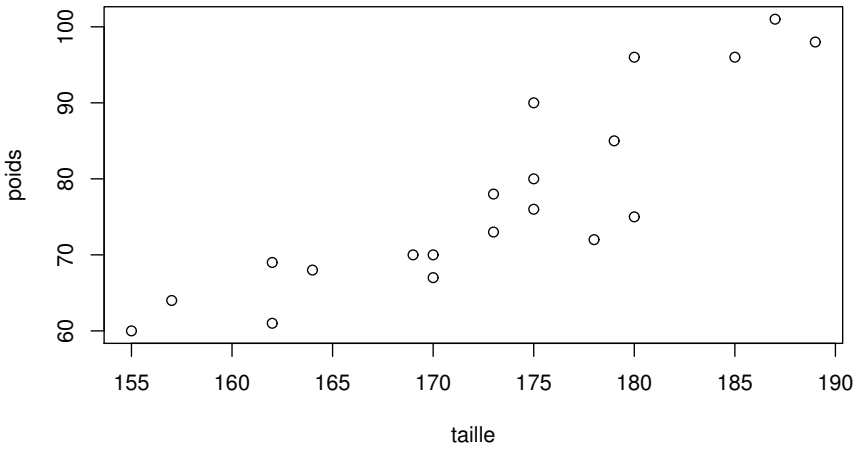


FIGURE 3.1 – Le nuage de points

3.2.2 Analyse des variables

Les variables x et y peuvent être analysées séparément. On peut calculer tous les paramètres dont les moyennes et les variances :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Ces paramètres sont appelés *paramètres marginaux* : *variances marginales*, *moyennes marginales*, *écarts-types marginaux*.

3.2.3 Covariance

La *covariance* est définie

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Remarque 3.1
— La covariance peut prendre des valeurs positives, négatives ou nulles.
— Quand $x_i = y_i$, pour tout $i = 1, \ldots, n$, la covariance est égale à la variance.

Théorème 3.1. *La covariance peut également s'écrire :*

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Démonstration.

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - y_i\bar{x} - \bar{y}x_i + \bar{x}\bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\bar{x} - \frac{1}{n}\sum_{i=1}^{n}\bar{y}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}. \end{aligned}$$

□

3.2.4 Corrélation

Le *coefficient de corrélation* est la covariance divisée par les deux écart-types marginaux :

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Le *coefficient de détermination* est le carré du coefficient de corrélation :

$$r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Remarque 3.2

— Le coefficient de corrélation mesure la dépendance linéaire entre deux variables :
— $-1 \leq r_{xy} \leq 1$,
— $0 \leq r_{xy}^2 \leq 1$.
— Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante.
— Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante.
— Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.

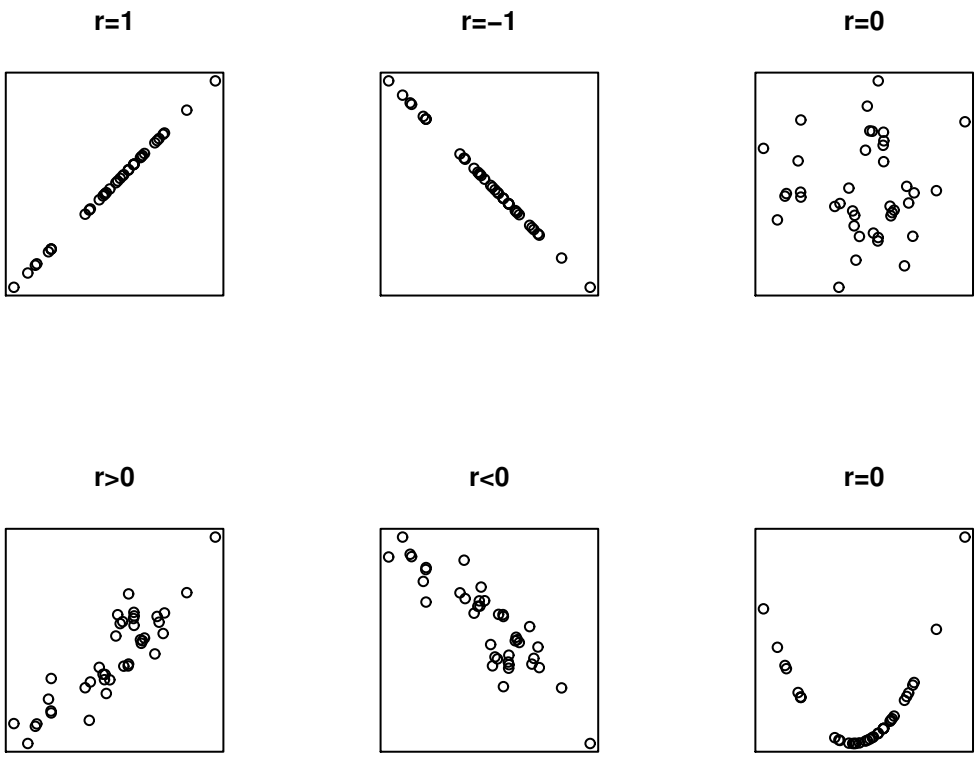


FIGURE 3.2 – Exemples de nuages de points et coefficients de corrélation

3.2.5 Droite de régression

La *droite de régression* est la droite qui ajuste au mieux un nuage de points au sens des moindres carrés.

On considère que la variable X est explicative et que la variable Y est dépendante. L'équation d'une droite est

$$y = a + bx.$$

Le problème consiste à identifier une droite qui ajuste bien le nuage de points. Si les coefficients a et b étaient connus, on pourrait calculer les résidus de la régression définis par :

$$e_i = y_i - a - b\ x_i.$$

Le résidu e_i est l'erreur que l'on commet (voir Figure 3.3) en utilisant la droite de régression pour prédire y_i à partir de x_i. Les résidus peuvent être positifs ou négatifs.

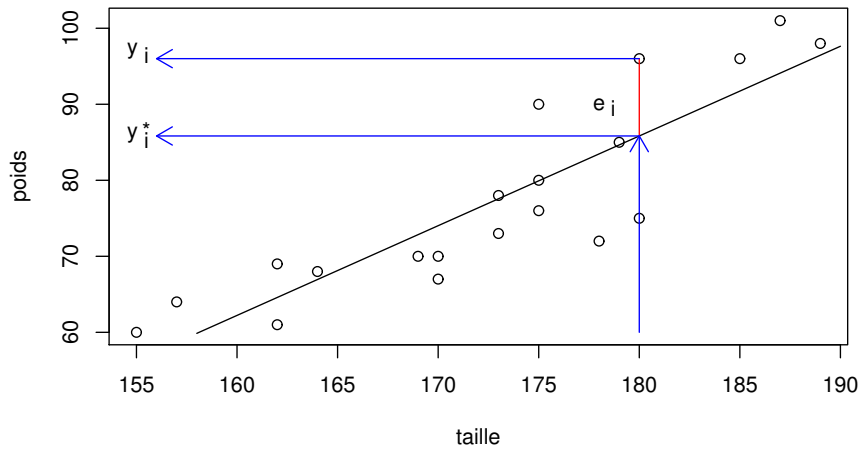


FIGURE 3.3 – Le nuage de points, le résidu

En langage R

```
# Graphique avec les résidus
plot(taille,poids)
segments(158,a+b*158,190,a+b*190)
segments(180,a+b*180,180,96,col="red")
#
text(178,90,expression(e))
text(178.7,89.5,"i")
#
arrows(180,a+b*180,156,a+b*180,col="blue",length=0.14)
arrows(180,60,180,a+b*180,col="blue",length=0.14)
arrows(180,96,156,96,col="blue",length=0.14)
#
text(154.8,86,expression(y))
text(155.5,85.5,"i")
#
text(154.8,97,expression(y))
text(155.5,97.8,"*")
text(155.5,96.5,"i")
```

Pour déterminer la valeur des coefficients a et b on utilise le principe des *moindres carrés* qui consiste à chercher la droite qui minimise la somme des carrés des résidus :

$$M(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2 .$$

Théorème 3.2. *Les coefficients a et b qui minimisent le critère des moindres carrés sont donnés par :*

$$b = \frac{s_{xy}}{s_x^2} \quad et \quad a = \bar{y} - b\,\bar{x}.$$

Démonstration. Le minimum $M(a,b)$ en (a,b) s'obtient en annulant les dérivées partielles par rapport à a et b.

$$\begin{cases} \dfrac{\partial M(a,b)}{\partial a} = -\displaystyle\sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) = 0 \\ \dfrac{\partial M(a,b)}{\partial b} = -\displaystyle\sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) x_i = 0. \end{cases}$$

On obtient un système de deux équations à deux inconnues. En divisant les deux équations par $-2n$, on obtient :

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(y_i - a - bx_i\right) = 0 \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(y_i - a - bx_i\right)x_i = 0, \end{cases}$$

ou encore

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i - \dfrac{1}{n}\sum_{i=1}^{n} a - b\dfrac{1}{n}\sum_{i=1}^{n} x_i = 0 \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i x_i - \dfrac{1}{n}\sum_{i=1}^{n} a x_i - \dfrac{1}{n}\sum_{i=1}^{n} b x_i^2 = 0, \end{cases}$$

ce qui s'écrit aussi

$$\begin{cases} \bar{y} = a + b\ \bar{x} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i x_i - a\bar{x} - \dfrac{1}{n}\sum_{i=1}^{n} b x_i^2 = 0. \end{cases}$$

La première équation montre que la droite passe par le point $(\bar{x}, \bar{y})$. On obtient

$$a = \bar{y} - b\ \bar{x}.$$

En remplaçant a par $\bar{y} - b\ \bar{x}$ dans la seconde équation, on a

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{y} - b\ \bar{x})\bar{x} - b\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - b\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = s_{xy} - b s_x^2 = 0,$$

ce qui donne

$$s_{xy} - b s_x^2 = 0.$$

Donc,

$$b = \frac{s_{xy}}{s_x^2}.$$

On a donc identifié les deux paramètres

$$\begin{cases} b = \dfrac{s_{xy}}{s_x^2} \text{ (la pente)} \\ a = \bar{y} - b\ \bar{x} = \bar{y} - \dfrac{s_{xy}}{s_x^2}\bar{x} \text{ (la constante).} \end{cases}$$

On devrait en outre vérifier qu'il s'agit bien d'un minimum en montrant que la matrice des dérivées secondes est définie positive. □

La droite de régression est donc

$$y = a + b\ x = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2}x,$$

ce qui peut s'écrire aussi

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Remarque 3.3 La droite de régression de y en x n'est pas la même que la droite de régression de x en y.

3.2.6 Résidus et valeurs ajustées

Les *valeurs ajustées* sont obtenues au moyen de la droite de régression :

$$y_i^* = a + bx_i.$$

Les valeurs ajustées sont les 'prédictions' des y_i réalisées au moyen de la variable x et de la droite de régression de y en x.

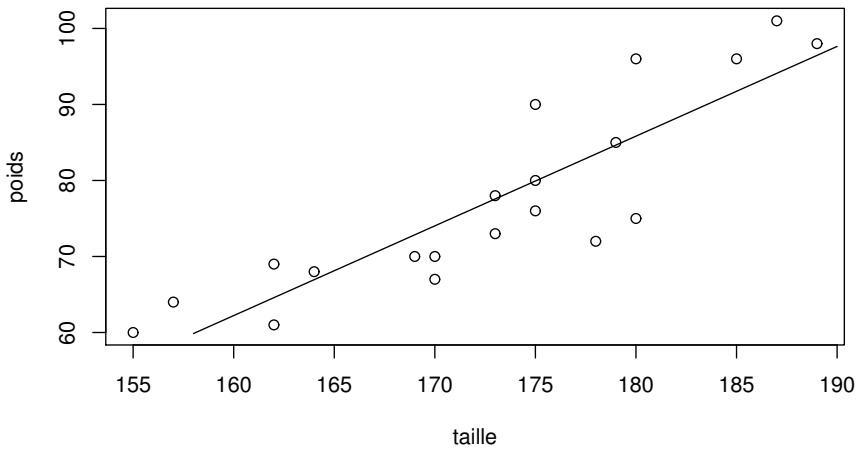


FIGURE 3.4 – La droite de régression

Remarque 3.4 La moyenne des valeurs ajustées est égale à la moyenne des valeurs observées $\bar{y}$. En effet,

$$\frac{1}{n}\sum_{i=1}^{n} y_i^* = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\ \bar{x}.$$

Or, $\bar{y} = a + b\ \bar{x}$, car le point $(\bar{x}, \bar{y})$ appartient à la droite de régression.

Les résidus sont les différences entre les valeurs observées et les valeurs ajustées de la variable dépendante.

$$e_i = y_i - y_i^*.$$

Les résidus représentent la partie inexpliquée des y_i par la droite de régression.
Remarque 3.5

— La moyenne des résidus est nulle. En effet

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*) = \bar{y} - \bar{y} = 0.$$

— De plus,

$$\sum_{i=1}^{n} x_i e_i = 0.$$

La démonstration est un peu plus difficile.

3.2.7 Sommes de carrés et variances

Définition 3.1. *On appelle somme des carrés totale la quantité*

$$SC_{TOT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

La variance marginale peut alors être définie par

$$s_y^2 = \frac{\mathrm{SC}_{TOT}}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Définition 3.2. *On appelle somme des carrés de la régression la quantité*

$$SC_{REGR} = \sum_{i=1}^{n} (y_i^* - \bar{y})^2.$$

Définition 3.3. *La variance de régression est la variance des valeurs ajustées.*

$$s_{y^*}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i^* - \bar{y})^2.$$

Définition 3.4. *On appelle somme des carrés des résidus (ou résiduelle) la quantité*

$$SC_{RES} = \sum_{i=1}^{n} e_i^2.$$

Définition 3.5. *La variance résiduelle est la variance des résidus.*

$$s_e^2 = \frac{SC_{RES}}{n} = \frac{1}{n} \sum_{i=1}^{n} e_i^2.$$

Note : Il n'est pas nécessaire de centrer les résidus sur leurs moyennes pour calculer la variance, car la moyenne des résidus est nulle.

Théorème 3.3.

$$SC_{TOT} = SC_{REGR} + SC_{RES}.$$

Démonstration.

$$\begin{aligned} \mathrm{SC}_{TOT} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - y_i^* + y_i^* - \bar{y})^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 + \sum_{i=1}^{n} (y_i^* - \bar{y})^2 + 2 \sum_{i=1}^{n} (y_i - y_i^*)(y_i^* - \bar{y}) \\ &= \mathrm{SC}_{RES} + \mathrm{SC}_{REGR} + 2 \sum_{i=1}^{n} (y_i - y_i^*)(y_i^* - \bar{y}). \end{aligned}$$

Le troisième terme est nul. En effet,

$$\sum_{i=1}^{n} (y_i - y_i^*)(y_i^* - \bar{y}) = \sum_{i=1}^{n} (y_i - a - b\ x_i)(a + b\ x_i - \bar{y}).$$

En remplaçant a par $\bar{y} - b\ \bar{x}$, on obtient

$$\begin{aligned} \sum_{i=1}^{n} (y_i - y_i^*)(y_i^* - \bar{y}) &= \sum_{i=1}^{n} \left\{ y_i - \bar{y} - b(x_i - \bar{x}) \right\} b(x_i - \bar{x}) \\ &= \sum_{i=1}^{n} \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\} b(x_i - \bar{x}) = b \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - b^2 \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) \\ &= b n s_{xy} - b^2 n s_x^2 = \frac{s_{xy}}{s_x^2} n s_{xy} - \frac{s_{xy}^2}{s_x^4} n s_x^2 = 0. \end{aligned}$$

□

3.2.8 Décomposition de la variance

Théorème 3.4. *La variance de régression peut également s'écrire*

$$s_{y^*}^2 = s_y^2 r^2,$$

où r^2 est le coefficient de détermination.

Démonstration.

$$s_{y^*}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i^* - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} \left\{ \bar{y} + \frac{s_{xy}}{s_x^2}(x_i - \bar{x}) - \bar{y} \right\}^2 = \frac{s_{xy}^2}{s_x^4} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{s_{xy}^2}{s_x^2} = s_y^2 \frac{s_{xy}^2}{s_x^2 s_y^2} = s_y^2 r^2.$$

□

La *variance résiduelle* est la variance des résidus.

$$s_e^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

Théorème 3.5. *La variance résiduelle peut également s'écrire*

$$s_e^2 = s_y^2(1 - r^2),$$

où r^2 est le coefficient de détermination.

Démonstration.

$$\begin{aligned} s_e^2 &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \bar{y} - \frac{s_{xy}}{s_x^2}(x_i - \bar{x})\right\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{s_{xy}^2}{s_x^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\frac{s_{xy}}{s_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= s_y^2 + \frac{s_{xy}^2}{s_x^2} - 2\frac{s_{xy}^2}{s_x^2} = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right). \end{aligned}$$

□

Théorème 3.6. *La variance marginale est la somme de la variance de régression et de la variance résiduelle,*

$$s_y^2 = s_{y^*}^2 + s_e^2.$$

La démonstration découle directement des deux théorèmes précédents.

3.3 Deux variables qualitatives

3.3.1 Données observées

Si les deux variables x et y sont qualitatives, alors les données observées sont une suite de couples de variables

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

chacune des deux variables prend comme valeurs des modalités qualitatives.

Les valeurs distinctes de x et y sont notées respectivement

$$x_1, \ldots, x_j, \ldots, x_J$$

et

$$y_1, \ldots, y_k, \ldots, y_K.$$

3.3.2 Tableau de contingence

Les données observées peuvent être regroupées dans le Tableau 3.2 sous la forme d'un *tableau de contingence* Les $n_{j.}$ et $n_{.k}$ sont appelés les effectifs marginaux. Dans ce tableau,

— $n_{j.}$ représente le nombre de fois que la modalité x_j apparaît,
— $n_{.k}$ représente le nombre de fois que la modalité y_k apparaît,
— n_{jk} représente le nombre de fois que les modalités x_j et y_k apparaissent ensemble.

On a les relations

$$\sum_{j=1}^{J} n_{jk} = n_{.k}, \text{ pour tout } k = 1, \ldots, K,$$

$$\sum_{k=1}^{K} n_{jk} = n_{j.}, \text{ pour tout } j = 1, \ldots, J,$$

et

$$\sum_{j=1}^{J} n_{j.} = \sum_{k=1}^{K} n_{.k} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = n \quad .$$

Exemple 3.2 On s'intéresse à une éventuelle relation entre le sexe de 200 personnes et la couleur des yeux. Le Tableau 3.3 reprend le tableau de contingence.

TABLE 3.2 – Tableau de contingence

	y_1	$\cdots$	y_k	$\cdots$	y_K	total
x_1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jK}	$n_{j.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_J	n_{J1}	$\cdots$	n_{Jk}	$\cdots$	n_{JK}	$n_{J.}$
total	$n_{.1}$	$\cdots$	$n_{.k}$		$n_{.K}$	n

TABLE 3.3 – Tableau des effectifs n_{jk}

	Bleu	Vert	Marron	Total
Homme	10	50	20	80
Femme	20	60	40	120
Total	30	110	60	200

3.3.3 Tableau des fréquences

Le *tableau de fréquences* s'obtient en divisant tous les effectifs par la taille de l'échantillon :

$$f_{jk} = \frac{n_{jk}}{n}, j = 1, \ldots, J, k = 1, \ldots, K$$

$$f_{j.} = \frac{n_{j.}}{n}, j = 1, \ldots, J,$$

$$f_{.k} = \frac{n_{.k}}{n}, k = 1, \ldots, K.$$

Le Tableau 3.4 contient les fréquences.

TABLE 3.4 – Tableau de fréquences

	y_1	$\cdots$	y_k	$\cdots$	y_K	total
x_1	f_{11}	$\cdots$	f_{1k}	$\cdots$	f_{1K}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_j	f_{j1}	$\cdots$	f_{jk}	$\cdots$	f_{jK}	$f_{j.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_J	f_{J1}	$\cdots$	f_{Jk}	$\cdots$	f_{JK}	$f_{J.}$
total	$f_{.1}$	$\cdots$	$f_{.k}$		$f_{.K}$	1

Exemple 3.3 Le Tableau 3.5 reprend le tableau des fréquences.

TABLE 3.5 – Tableau des fréquences

	Bleu	Vert	Marron	Total
Homme	0.05	0.25	0.10	0.40
Femme	0.10	0.30	0.20	0.60
Total	0.15	0.55	0.30	1.00

3.3.4 Profils lignes et profils colonnes

Un tableau de contingence s'interprète toujours en comparant des fréquences en lignes ou des fréquences en colonnes (appelés aussi *profils lignes* et *profils colonnes*).

Les profils lignes sont définis par

$$f_k^{(j)} = \frac{n_{jk}}{n_{j.}} = \frac{f_{jk}}{f_{j.}}, k = 1, \ldots, K, j = 1, \ldots, J,$$

et les profils colonnes par

$$f_j^{(k)} = \frac{n_{jk}}{n_{.k}} = \frac{f_{jk}}{f_{.k}}, j = 1, \ldots, J, k = 1, \ldots, K.$$

Exemple 3.4 Le Tableau 3.6 reprend le tableau des profils lignes et le Tableau 3.7 reprend le tableau des profils colonnes.

Table 3.6 – Tableau des profils lignes

	Bleu	Vert	Marron	Total
Homme	0.13	0.63	0.25	1.00
Femme	0.17	0.50	0.33	1.00
Total	0.15	0.55	0.30	1.00

Table 3.7 – Tableau des profils colonnes

	Bleu	Vert	Marron	Total
Homme	0.33	0.45	0.33	0.40
Femme	0.67	0.55	0.67	0.60
Total	1.00	1.00	1.00	1.00

3.3.5 Effectifs théoriques et khi-carré

On cherche souvent une interaction entre des lignes et des colonnes, un lien entre les variables. Pour mettre en évidence ce lien, on construit un tableau d'effectifs théoriques qui représente la situation où les variables ne sont pas liées (indépendance). Ces *effectifs théoriques* sont construits de la manière suivante :

$$n_{jk}^* = \frac{n_{j.} n_{.k}}{n}.$$

Les effectifs observés n_{jk} ont les mêmes marges que les effectifs théoriques n_{jk}^*.

Enfin, les *écarts à l'indépendance* sont définis par

$$e_{jk} = n_{jk} - n_{jk}^*.$$

— La dépendance du tableau se mesure au moyen du khi-carré défini par

$$\chi^2_{obs} = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*} = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{e_{jk}^2}{n_{jk}^*}. \tag{3.1}$$

— Le khi-carré peut être normalisé pour ne plus dépendre du nombre d'observations. On définit le phi-deux par :

$$\phi^2 = \frac{\chi^2_{obs}}{n}.$$

Le ϕ^2 ne dépend plus du nombre d'observations. Il est possible de montrer que

$$\phi^2 \leq \min(J - 1, K - 1),$$

où J et le nombre de lignes du tableau et K est le nombre de colonnes.

— Le V de Cramer est définit par

$$V = \sqrt{\frac{\phi^2}{\min(J-1, K-1)}} = \sqrt{\frac{\chi^2_{obs}}{n \min(J-1, K-1)}}.$$

Le V de Cramer est compris entre 0 et 1. Il ne dépend ni de la taille de l'échantillon ni de la taille du tableau. Si $V \approx 0$, les deux variables sont indépendantes. Si $V = 1$, il existe une relation fonctionnelle entre les variables, ce qui signifie que chaque ligne et chaque colonne du tableau de contingence ne contiennent qu'un seul effectif différent de 0 (il faut que le tableau ait le même nombre de lignes que de colonnes).

Exemple 3.5 Le Tableau 3.8 reprend le tableau des effectifs théoriques, le Tableau 3.9 reprend le tableau des écarts à l'indépendance. Enfin, les e^2_{jk}/n^*_{jk} sont présentés dans le Tableau 3.10.

Table 3.8 – Tableau des effectifs théoriques n^*_{jk}

	Bleu	Vert	Marron	Total
Homme	12	44	24	80
Femme	18	66	36	120
Total	30	110	60	200

Table 3.9 – Tableau des écarts à l'indépendance e_{jk}

	Bleu	Vert	Marron	Total
Homme	-2	6	-4	0
Femme	2	-6	4	0
Total	0	0	0	0

Table 3.10 – Tableau des e^2_{jk}/n^*_{jk}

	Bleu	Vert	Marron	Total
Homme	0.33	0.82	0.67	1.82
Femme	0.22	0.55	0.44	1.21
Total	0.56	1.36	1.11	3.03

— Le khi-carré observé vaut $\chi^2_{obs} = 3.03$.
— Le phi-deux vaut $\phi^2 = 0.01515$.
— Comme le tableau a deux lignes $\min(J-1, K-1) = \min(2-1, 3-1) = 1$. Le V de Cramer est égal à $\sqrt{\phi^2}$.
— On a $V = 0.123$. La dépendance entre les deux variables est très faible.

En langage R

```
yeux= c(rep("bleu",times=10),rep("vert",times=50),rep("marron",times=20),
rep("bleu",times=20),rep("vert",times=60),rep("marron",times=40))
sexe= c(rep("homme",times=80),rep("femme",times=120))
yeux=factor(yeux,levels=c("bleu","vert","marron"))
sexe=factor(sexe,levels=c("homme","femme"))
T=table(sexe,yeux)
T
plot(T,main="")
summary(T)
```

Exemple 3.6 Le tableau suivant est extrait de Boudon (1979, p. 57). La variable X est le niveau d'instruction du fils par rapport au père (plus élevé, égal, inférieur) et la variable Y est le statut professionnel du fils par rapport au père (plus élevé, égal, inférieur).

Table 3.11 – Tableau de contingence : effectifs n_{jk}

Niveau d'instruction du fils par rapport au père	Statut professionnel du fils par rapport au père			
	Plus élevé	Égal	inférieur	total
plus élevé	134	96	61	291
égal	23	33	24	80
inférieur	7	16	22	45
total	164	145	107	416

Table 3.12 – Tableau des fréquences f_{jk}

$X \backslash Y$	Plus élevé	Égal	inférieur	total
plus élevé	0.322	0.231	0.147	0.700
égal	0.055	0.079	0.058	0.192
inférieur	0.017	0.038	0.053	0.108
total	0.394	0.349	0.257	1.000

Table 3.13 – Tableau des profils lignes

$X \backslash Y$	Plus élevé	Égal	inférieur	total
plus élevé	0.460	0.330	0.210	1
égal	0.288	0.413	0.300	1
inférieur	0.156	0.356	0.489	1
total	0.394	0.349	0.257	1

Table 3.14 – Tableau des profils colonnes

$X \backslash Y$	Plus élevé	Égal	inférieur	total
plus élevé	0.817	0.662	0.570	0.700
égal	0.140	0.228	0.224	0.192
inférieur	0.043	0.110	0.206	0.108
total	1	1	1	1

Table 3.15 – Tableau des effectifs théoriques n^*_{jk}

$X \backslash Y$	Plus élevé	Égal	inférieur	total
plus élevé	114.72	101.43	74.85	291
égal	31.54	27.88	20.58	80
inférieur	17.74	15.69	11.57	45
total	164	145	107	416

Table 3.16 – Tableau des écarts à l'indépendance e_{jk}

$X \backslash Y$	Plus élevé	Égal	inférieur	total
plus élevé	19.28	−5.43	−13.85	0
égal	−8.54	5.12	3.42	0
inférieur	−10.74	0.31	10.43	0
total	0	0	0	0

Table 3.17 – Tableau des e_{jk}^2/n_{jk}^*

$X \backslash Y$	Plus élevé	Égal	inférieur	total
plus élevé	3.24	0.29	2.56	6.09
égal	2.31	0.94	0.57	3.82
inférieur	6.50	0.01	9.39	15.90
total	12.05	1.24	12.52	$\chi^2_{obs} = 25.81$

On a donc

$$\begin{aligned}
\chi^2_{obs} &= 25.81 \\
\phi^2 &= \frac{\chi^2_{obs}}{n} = \frac{25.81}{416} = 0.062 \\
V &= \sqrt{\frac{\phi^2}{\min(J-1, K-1)}} = \sqrt{\frac{0.062}{2}} = 0.176.
\end{aligned}$$

Chapitre 4

Théorie des indices, mesures d'inégalité

4.1 Nombres indices

4.1.1 Définition

Un indice est la valeur d'une grandeur par rapport à une valeur de référence. Prenons l'exemple du Tableau 4.1 contenant le prix (fictif) d'un bien de consommation de 2000 à 2006. Le temps varie de $0, 1, 2, \ldots, 6$ et 0 est considéré comme le temps de référence par rapport auquel l'indice est calculé.

Table 4.1 – Tableau du prix d'un bien de consommation de 2000 à 2006

année	t	prix p_t
2000	0	2.00
2001	1	2.30
2002	2	2.40
2003	3	2.80
2004	4	3.00
2005	5	3.50
2006	6	4.00

L'indice simple est défini par

$$I(t/t') = 100 \times \frac{p_t}{p_{t'}}, t, t' = 0, 1, \ldots, 6.$$

Le Tableau 4.2 contient la matrice des indices de prix du bien. Par exemple de 2000 à 2006, le prix a doublé. Donc, $I(6/0) = 200$.

Table 4.2 – Tableau de l'indice simple du prix du Tableau 4.1

	$t = 0$	1	2	3	4	5	6
$t' = 0$	100.00	115.00	120.00	140.00	150.00	175.00	200.00
1	86.96	100.00	104.35	121.74	130.43	152.17	173.91
2	83.33	95.83	100.00	116.67	125.00	145.83	166.67
3	71.43	82.14	85.71	100.00	107.14	125.00	142.86
4	66.67	76.67	80.00	93.33	100.00	116.67	133.33
5	57.14	65.71	68.57	80.00	85.71	100.00	114.29
6	50.00	57.50	60.00	70.00	75.00	87.50	100.00

4.1.2 Propriétés des indices

Considérons un indice quelconque $I(t/0)$. On dit que cet indice possède les propriétés de

— *réversibilité* si $I(t/0) = 100^2 \times \frac{1}{I(0/t)}$,

— *identité* si $I(t/t) = 100$,

— *circularité (ou transitivité)* si $I(t/u) \times I(u/v) = 100 \times I(t/v)$.

Il est facile de montrer que ces trois propriétés sont satisfaites pour un indice simple.

4.1.3 Indices synthétiques

Quand on veut calculer un indice à partir de plusieurs prix, le problème devient sensiblement plus compliqué. Un indice synthétique est une grandeur d'un ensemble de biens par rapport à une année de référence. On ne peut pas construire un indice synthétique en additionnant simplement des indices simples. Il faut, en effet, tenir compte des quantités achetées.

Pour calculer un indice de prix de n biens de consommation étiquetés de $1, 2, \ldots, n$, on utilise la notation suivante :

— p_{ti} représente le prix du bien de consommation i au temps t,

— q_{ti} représente la quantité de biens i consommée au temps t.

Considérons par exemple le Tableau 4.3 qui contient 3 biens de consommation et pour lesquels ont connaît les prix et les quantités achetées.

Table 4.3 – Exemple : prix et quantités de trois bien pendant 3 ans

Temps	0		1		2	
	Prix (p_{0i})	Quantités (q_{0i})	Prix (p_{1i})	Quantités (q_{1i})	Prix (p_{2i})	Quantités (q_{2i})
Bien 1	100	14	150	10	200	8
Bien 2	60	10	50	12	40	14
Bien 3	160	4	140	5	140	5

Il existe deux méthodes fondamentales pour calculer les indices de prix, l'indice de Paasche et l'indice de Laspeyres.

4.1.4 Indice de Laspeyres

L'indice de Laspeyres, est défini par

$$L(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_{0i} p_{ti}}{\sum_{i=1}^{n} q_{0i} p_{0i}}.$$

On utilise pour le calculer, les quantités q_{0i} du temps de référence.

L'indice de Laspeyres peut aussi être présenté comme une moyenne pondérée des indices simples. Soient l'indice simple du bien i :

$$I_i(t/0) = 100 \times \frac{p_{ti}}{p_{0i}},$$

et le poids w_{0i} correspondant à la recette totale du bien i au temps 0

$$w_{0i} = p_{0i} q_{0i}.$$

L'indice de Laspeyres peut alors être défini comme une moyenne des indices simples pondérés par les recettes au temps 0 :

$$L(t/0) = \frac{\sum_{i=1}^{n} w_{0i} I_i(t/0)}{\sum_{i=1}^{n} w_{0i}} = \frac{\sum_{i=1}^{n} p_{0i} q_{0i} 100 \times \frac{p_{ti}}{p_{0i}}}{\sum_{i=1}^{n} p_{0i} q_{0i}} = 100 \times \frac{\sum_{i=1}^{n} q_{0i} p_{ti}}{\sum_{i=1}^{n} p_{0i} q_{0i}}.$$

L'indice de Laspeyres ne possède ni la propriété de circularité ni de réversibilité. L'indice de Laspeyres est facile à calculer, car seules les quantités q_{0i} du temps de référence sont nécessaires pour le calculer.

Exemple 4.1 Si on utilise les données du Tableau 4.3, les indices de Laspeyres sont les suivants

$$L(1/0) = 100 \times \frac{\sum_{i=1}^{n} q_{0i} p_{1i}}{\sum_{i=1}^{n} q_{0i} p_{0i}} = 100 \times \frac{14 \times 150 + 10 \times 50 + 4 \times 140}{14 \times 100 + 10 \times 60 + 4 \times 160} = 119.6970,$$

$$L(2/0) = 100 \times \frac{\sum_{i=1}^{n} q_{0i} p_{2i}}{\sum_{i=1}^{n} q_{0i} p_{0i}} = 100 \times \frac{14 \times 200 + 10 \times 40 + 4 \times 140}{14 \times 100 + 10 \times 60 + 4 \times 160} = 142.4242,$$

$$L(2/1) = 100 \times \frac{\sum_{i=1}^{n} q_{1i} p_{2i}}{\sum_{i=1}^{n} q_{1i} p_{1i}} = 100 \times \frac{10 \times 200 + 12 \times 40 + 5 \times 140}{10 \times 150 + 12 \times 50 + 5 \times 140} = 113.5714.$$

4.1.5 Indice de Paasche

L'indice de Paasche, est défini par

$$P(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_{ti} p_{ti}}{\sum_{i=1}^{n} q_{ti} p_{0i}}.$$

On utilise, pour le calculer, les quantités q_{ti} du temps par rapport auquel on veut calculer l'indice.

L'indice de Paasche peut aussi être présenté comme une moyenne harmonique pondérée des indices simples. Soient l'indice simple du bien i :

$$I_i(t/0) = 100 \times \frac{p_{ti}}{p_{0i}},$$

et le poids w_{ti} correspondant à la recette totale du bien i au temps t

$$w_{ti} = p_{ti} q_{ti}.$$

L'indice de Paasche peut alors être défini comme une moyenne harmonique des indices simples pondérés par les recettes au temps t :

$$P(t/0) = \frac{\sum_{i=1}^{n} w_{ti}}{\sum_{i=1}^{n} w_{ti}/I_i(t/0)} = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{ti} q_{ti} \frac{p_{0i}}{100 \times p_{ti}}} = 100 \times \frac{\sum_{i=1}^{n} q_{ti} p_{ti}}{\sum_{i=1}^{n} q_{ti} p_{0i}}.$$

L'indice de Paasche ne possède ni la propriété de circularité ni de réversibilité. L'indice de Paasche est plus difficile à calculer que l'indice de Laspeyres, car on doit connaître les quantités pour chaque valeur de t.

Exemple 4.2 Si on utilise les données du Tableau 4.3, les indices de Paasche sont les suivants

$$P(1/0) = 100 \times \frac{\sum_{i=1}^{n} q_{1i} p_{1i}}{\sum_{i=1}^{n} q_{1i} p_{0i}} = 100 \times \frac{10 \times 150 + 12 \times 50 + 5 \times 140}{10 \times 100 + 12 \times 60 + 5 \times 160} = 111.1111,$$

$$P(2/0) = 100 \times \frac{\sum_{i=1}^{n} q_{2i} p_{2i}}{\sum_{i=1}^{n} q_{2i} p_{0i}} = 100 \times \frac{8 \times 200 + 14 \times 40 + 5 \times 140}{8 \times 100 + 14 \times 60 + 5 \times 160} = 117.2131,$$

$$P(2/1) = 100 \times \frac{\sum_{i=1}^{n} q_{2i} p_{2i}}{\sum_{i=1}^{n} q_{2i} p_{1i}} = 100 \times \frac{8 \times 200 + 14 \times 40 + 5 \times 140}{8 \times 150 + 14 \times 50 + 5 \times 140} = 110.$$

4.1.6 L'indice de Fisher

L'indice de Laspeyres est en général plus grand que l'indice de Paasche, ce qui peut s'expliquer par le fait que l'indice de Laspeyres est une moyenne arithmétique d'indices élémentaires tandis que l'indice de Paasche est une moyenne harmonique. Nous avons vu qu'une moyenne harmonique est toujours inférieure ou égale à une moyenne arithmétique (voir la remarque de la page 29). Cependant ici, ce résultat est approximatif, car on n'utilise pas les mêmes poids pour calculer l'indice de Paasche (w_{ti}) et de Laspeyres (w_{0i}).

Fisher a proposé d'utiliser un compromis entre l'indice de Paasche et de Laspeyres en calculant simplement la moyenne géométrique de ces deux indices

$$F(t/0) = \sqrt{L(t/0) \times P(t/0)}.$$

L'avantage de l'indice de Fisher est qu'il jouit de la propriété de réversibilité.

Exemple 4.3 Si on utilise toujours les données du Tableau 4.3, les indices de Fisher sont les suivants :

$$F(1/0) = \sqrt{L(1/0) \times P(1/0)} = 115.3242,$$

$$F(2/0) = \sqrt{L(2/0) \times P(2/0)} = 129.2052,$$

$$F(2/1) = \sqrt{L(2/1) \times P(2/1)} = 111.7715.$$

4.1.7 L'indice de Sidgwick

L'indice de Sidgwick est la moyenne arithmétique des indices de Paasche et de Laspeyres.

$$S(t/0) = \frac{L(t/0) + P(t/0)}{2}.$$

4.1.8 Indices chaînes

Le défaut principal des indices de Laspeyres, de Paasche, de Fisher et de Sidgwick est qu'il ne possèdent pas la propriété de circularité. Un indice qui possède cette propriété est appelé indice chaîne. Pour construire un indice chaîne, avec l'indice de Laspeyres, on peut faire un produit d'indice de Laspeyres annuels.

$$CL(t/0) = 100 \times \frac{L(t/t-1)}{100} \times \frac{L(t-1/t-2)}{100} \times \cdots \times \frac{L(2/1)}{100} \times \frac{L(1/0)}{100}.$$

Pour calculer un tel indice, on doit évidemment connaître les quantités pour chaque valeur de t. L'indice suisse des prix à la consommation est un indice chaîne de Laspeyres.

Exemple 4.4 En utilisant encore les données du Tableau 4.3, les indices chaînes de Laspeyres sont les suivants :

$$CL(1/0) = L(1/0) = 119.6970,$$

$$CL(2/1) = L(2/1) = 113.5714,$$

$$CL(2/0) = \frac{L(2/1) \times L(1/0)}{100} = 135.9416.$$

4.2 Mesures de l'inégalité

4.2.1 Introduction

Des indicateurs particuliers ont été développés pour mesurer les inégalités des revenus ou les inégalités de patrimoine. On considère qu'une société est parfaitement égalitaire si tous les individus reçoivent le même revenu. La situation théorique la plus inégalitaire est la situation où un individu perçoit la totalité des revenus et les autres individus n'ont aucun revenu.

4.2.2 Courbe de Lorenz

Plusieurs indices d'inégalité sont liés à la courbe de Lorenz. On note

$$x_1, \ldots, x_i, \ldots, x_n$$

les revenus des n individus de la population étudiée. On note également

$$x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(n)},$$

la statistique d'ordre, c'est-à-dire la série de revenus triés par ordre croissant.

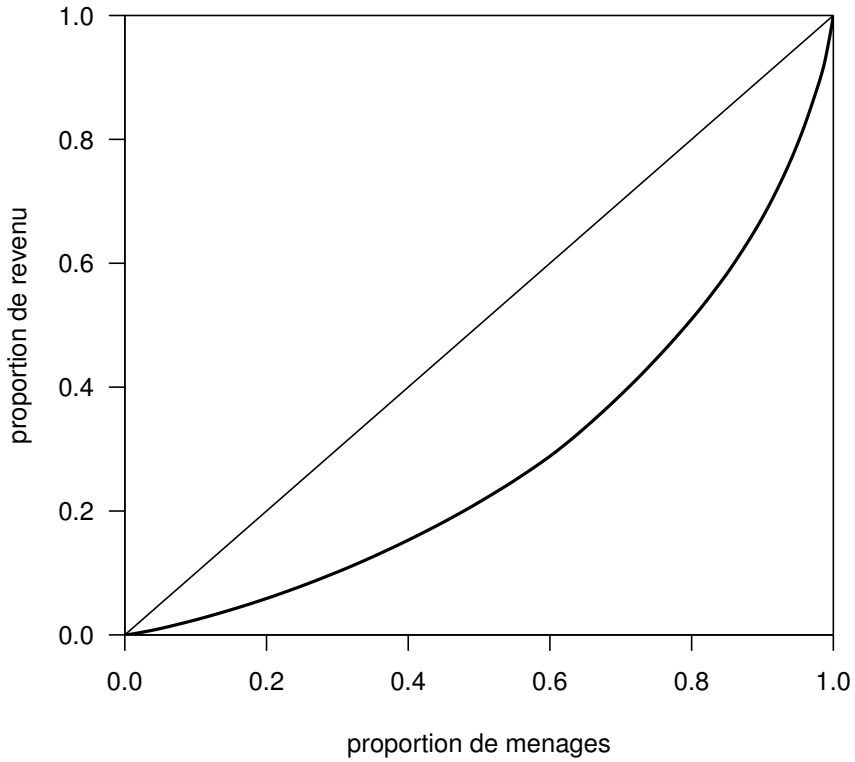
Notons maintenant q_i la proportion de revenus par rapport au revenu total qu'ont gagné les i individus ayant les plus bas revenus, ce qui s'écrit

$$q_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}} \text{ avec } q_0 = 0 \text{ et } q_n = 1.$$

La courbe de Lorenz est la représentation graphique de la fonction qui à la part des individus les moins riches associe la part y du revenu total qu'ils perçoivent. Plus précisément, la courbe de Lorenz relie les points $(i/n, q_i)$ pour $i = 1, \ldots, n$. En abscisse, on a donc une proportion d'individus classés par ordre de revenu et en ordonnée la proportion du revenu total reçu par ces individus.

Exemple 4.5 On utilise une enquête ménage sur le revenu dans une région des Philippines appelée Ilocos. Cette enquête de 1997 sur le revenu des ménages a été produite par l'Office philippin de Statistique. La courbe de Lorenz est présentée en Figure 4.1.

FIGURE 4.1 – Courbe de Lorenz



Remarque 4.1 Sur le graphique, on indique toujours la diagonale. La courbe de Lorenz est égale à la diagonale si tous les individus ont le même revenu. Plus l'écart entre la courbe de Lorenz et la diagonale est important, plus les revenus sont distribués de manière inégalitaire.

En langage R

```
#
# Courbe de Lorenz et indices d'inégalité
#
# Etape 1 : on installe la package ineq
install.packages("ineq")
#
#Etape 2 : on charge le package ineq
#
library(ineq)
#
# Utilisation de la base de données Ilocos
# Enquête sur le revenu de l'Office de Statistique Philippin
data(Ilocos)
attach(Ilocos)
#
plot(Lc(income),xlab="proportion de menages",
ylab="proportion de revenu",main="")
```

4.2.3 Indice de Gini

L'indice de Gini, noté G est égal à deux fois la surface comprise entre la courbe de Lorenz et la diagonale. Il est possible de montrer que :

$$G = \frac{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2\bar{x}}.$$

En utilisant la statistique d'ordre $x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(n)}$, l'indice de Gini peut également s'écrire

$$G = \frac{1}{n-1}\left(\frac{2\sum_{i=1}^{n} i x_{(i)}}{n\bar{x}} - (n+1)\right).$$

L'indice de Gini est compris entre 0 et 1. S'il est proche de 0, tous le revenus sont égaux. S'il est proche de 1, les revenus sont très inégaux.

4.2.4 Robin Hood index

L'indice d'équirépartition de Pietra ou indice de Hoover ou encore indice de Schutz (ou *Robin Hood index*) est défini comme la proportion de revenus qu'il faudrait prendre aux individus gagnant plus que la moyenne et redistribuer aux individus gagnant moins que la moyenne pour que tout le monde ait le même revenu. Il est formellement définit par :

$$H = \frac{\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|}{2\bar{x}}.$$

Cet indice est également compris entre 0 et 1. Il vaut 0 si tous les individus ont le même revenu.

Cet indice est également lié à la courbe de Lorenz, car il est possible de montrer qu'il correspond à la plus grande distance verticale entre la courbe de Lorenz et la diagonale.

4.2.5 Quintile et Decile share ratio

On définit d'abord :

— S_{10} revenu moyen des individus ayant un revenu inférieur au premier décile $x_{1/10}$,
— S_{20} revenu moyen des individus ayant un revenu inférieur au premier quintile ou deuxième décile $x_{1/5}$,
— S_{80} revenu moyen des individus ayant un revenu supérieur au quatrième quintile ou huitième décile $x_{4/5}$,
— S_{90} revenu moyen des individus ayant un revenu supérieur au neuvième décile $x_{9/10}$.

Le quintile share ratio est définit par

$$QSR = \frac{S_{80}}{S_{20}}.$$

Le decile share ratio est définit par

$$DSR = \frac{S_{90}}{S_{10}}.$$

Ces quantités sont toujours plus grandes que 1 et augmentent avec l'inégalité. Ces deux rapports sont facilement interprétables, par exemple si le $QSR = 5$, cela signifie que le revenu moyen de 20% des plus riches est 5 fois plus grand que le revenu moyen de 20% des plus pauvres.

4.2.6 Indice de pauvreté

Un indice simple de pauvreté consiste à calculer le pourcentage de la population gagnant moins que 60% de la médiane.

4.2.7 Indices selon les pays

Le Tableau 4.4 reprend pour tous les pays l'indice de Gini et le rapport des 20% les plus riches sur les 20% les plus pauvres. (référence : United Nations 2005 Development Programme Report, page 270).

Table 4.4 – Mesures de l'inégalité par pays : source Eurostat

	Gini 2010	Gini 2011	QSR 2010	QSR 2011
European Union (27 countries) and Croatia	30.5	30.7	5.0	5.1
European Union (27 countries)	30.5	30.7	5.0	5.1
European Union (15 countries)	30.5	30.8	5.0	5.1
New Member States (12 countries)	30.3	30.5	5.0	5.1
Euro area (17 countries)	30.2	30.5	4.9	5.0
Belgium	26.6	26.3	3.9	3.9
Bulgaria	33.2	35.1	5.9	6.5
Czech Republic	24.9	25.2	3.5	3.5
Denmark	26.9	27.8	4.4	4.4
Germany (until 1990 former territory of the FRG)	29.3	29.0	4.5	4.5
Estonia	31.3	31.9	5.0	5.3
Ireland	33.2	:	5.3	:
Greece	32.9	33.5	5.6	6.0
Spain	33.9	34.0	6.9	6.8
France	29.8	30.8	4.5	4.6
Italy	31.2	31.9	5.2	5.6
Cyprus	29.8	29.1	4.5	4.3
Latvia	36.1	35.4	6.9	6.6
Lithuania	36.9	32.9	7.3	5.8
Luxembourg	27.9	27.2	4.1	4.0
Hungary	24.1	26.8	3.4	3.9
Malta	28.4	27.4	4.3	4.1
Netherlands	25.5	25.8	3.7	3.8
Austria	26.1	26.3	3.7	3.8
Poland	31.1	31.1	5.0	5.0
Portugal	33.7	34.2	5.6	5.7
Romania	33.3	33.2	6.0	6.2
Slovenia	23.8	23.8	3.4	3.5
Slovakia	25.9	25.7	3.8	3.8
Finland	25.4	25.8	3.6	3.7
Sweden	24.1	24.4	3.5	3.6
United Kingdom	33.0	33.0	5.4	5.3
Iceland	25.7	23.6	3.6	3.3
Norway	23.6	22.9	3.4	3.3
Switzerland	29.6	29.7	4.5	4.5
Croatia	31.5	31.0	5.6	5.4

Chapitre 5

Séries temporelles

5.1 Définitions générales et exemples

5.1.1 Définitions

Définition 5.1. *Une série temporelle est une suite d'observations d'une quantité répétée dans le temps.*

On énonce en général l'hypothèse que les intervalles de temps sont équidistants. La série temporelle est notée

$$y_1, \ldots, y_t, \ldots, y_T.$$

On note également $\mathcal{T} = \{1, 2, \ldots, t, \ldots, T\}$ l'ensemble des instants auxquels les observations sont réalisées.

Une série temporelle peut se composer de :
— une tendance T_t,
— une composante cyclique C_t (nous n'étudierons pas cette question),
— une composante saisonnière S_t,
— un résidu E_t (partie inexpliquée).

On étudie deux types de modèles :
— Le modèle additif :

$$y_t = T_t + C_t + S_t + E_t$$

— Le modèle multiplicatif :

$$y_t = T_t \times C_t \times S_t \times E_t.$$

Il peut être intéressant de décomposer la série, ce qui consiste à séparer les composantes T_t, C_t, S_t, E_t.

5.1.2 Traitement des séries temporelles

Le traitement des séries temporelles peut avoir plusieurs objectifs.
— isoler et estimer une tendance,
— isoler et estimer une composante saisonnière et désaisonnaliser la série,
— réaliser une prévision pour des valeurs inconnues manquantes, futures ou passées,
— construire un modèle explicatif en terme de causalité,
— déterminer la durée d'un cycle.

5.1.3 Exemples

Exemple 5.1 Extrait de "The Data and Story Library" Ces données trimestrielles, ont été produites par le service des statistiques d'entreprise du Bureau of Census (États-Unis). Les données concernant les ventes reprennent le nombre de biens expédiés durant 32 trimestres.
— QTR : Quarter, trimestres depuis le 1er trimestre 1978 jusqu'au 4ème trimestre 1985
— DISH : Nombre de lave-vaisselles (dishwashers) expédiés (milliers)
— DISP : Nombre de broyeurs d'ordures (disposers) expédiés (milliers)
— FRIG : Nombre de réfrigérateurs expédiés (milliers)
— WASH : Nombre de machines à laver (washing machine) expédiées (milliers)
— DUR : Dépenses en biens durables USA (milliards de dollars de 1982)
— RES : Investissement résidentiel privé USA (milliards de dollars de 1982)

Table 5.1 – Biens manufacturés aux USA

QTR	DISH	DISP	FRIG	WASH	DUR	RES
1	841	798	1317	1271	252.6	172.9
2	957	837	1615	1295	272.4	179.8
3	999	821	1662	1313	270.9	180.8
4	960	858	1295	1150	273.9	178.6
5	894	837	1271	1289	268.9	174.6
6	851	838	1555	1245	262.9	172.4
7	863	832	1639	1270	270.9	170.6
8	878	818	1238	1103	263.4	165.7
9	792	868	1277	1273	260.6	154.9
10	589	623	1258	1031	231.9	124.1
11	657	662	1417	1143	242.7	126.8
12	699	822	1185	1101	248.6	142.2
13	675	871	1196	1181	258.7	139.3
14	652	791	1410	1116	248.4	134.1
15	628	759	1417	1190	255.5	122.3
16	529	734	919	1125	240.4	110.4
17	480	706	943	1036	247.7	101.2
18	530	582	1175	1019	249.1	103.4
19	557	659	1269	1047	251.8	100.1
20	602	837	973	918	262.0	115.8
21	658	867	1102	1137	263.3	127.8
22	749	860	1344	1167	280.0	147.4
23	827	918	1641	1230	288.5	161.9
24	858	1017	1225	1081	300.5	159.9
25	808	1063	1429	1326	312.6	170.5
26	840	955	1699	1228	322.5	173.1
27	893	973	1749	1297	324.3	170.3
28	950	1096	1117	1198	333.1	169.6
29	838	1086	1242	1292	344.8	170.3
30	884	990	1684	1342	350.3	172.9
31	905	1028	1764	1323	369.1	175.0
32	909	1003	1328	1274	356.4	179.4

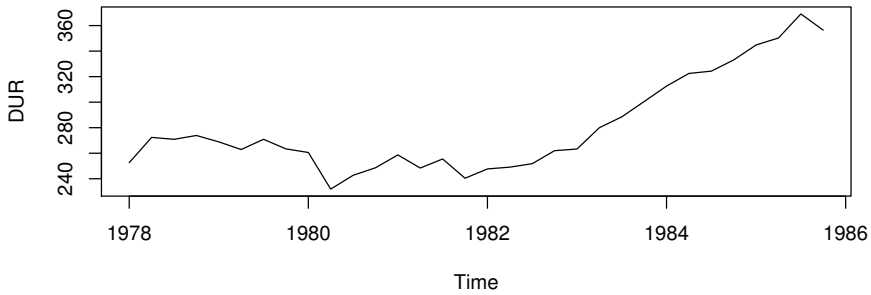


Figure 5.1 – Dépenses en biens durables USA (milliards de dollars de 1982)

Exemple 5.2 La variable "nombre" de réfrigérateurs vendus a manifestement une composante saisonnière et une tendance.

En langage R

```
QTR=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,
26,27,28,29,30,31,32)
```

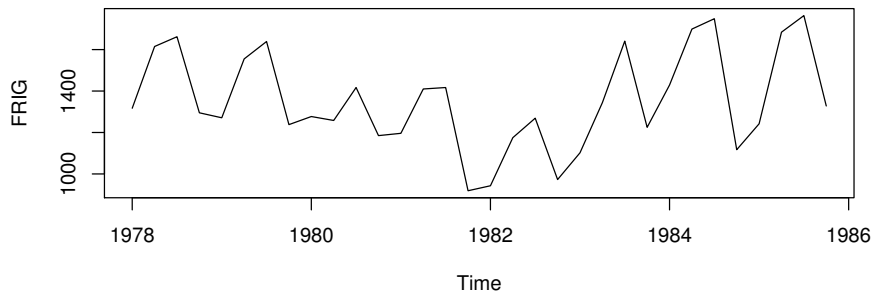


FIGURE 5.2 – Nombre de réfrigérateurs vendus de 1978 à 1985

```
DISH=c(841,957,999,960,894,851,863,878,792,589,657,699,675,652,628,
529,480,530,557,602,658,749,827,858,808,840,893,950,838,884,905,909)
DISP=c(798,837,821,858,837,838,832,818,868,623,662,822,871,791,759,734,706,
582,659,837,867,860,918,1017,1063,955,973,1096,1086,990,1028,1003)
FRIG=c(1317,1615,1662,1295,1271,1555,1639,1238,1277,1258,1417,1185,1196,
1410,1417,919,943,1175,1269,973,1102,1344,1641,1225,1429,1699,1749,
1117,1242,1684,1764,1328)
WASH=c(1271,1295,1313,1150,1289,1245,1270,1103,1273,1031,1143,1101,1181,
1116,1190,1125,1036,1019,1047,918,1137,1167,1230,1081,1326,1228,1297,
1198,1292,1342,1323,1274)
DUR=c(252.6,272.4,270.9,273.9,268.9,262.9,270.9,263.4,260.6,231.9,242.7,248.6,
258.7,248.4,255.5,240.4,247.7,249.1,251.8,262,263.3,280,288.5,300.5,
312.6,322.5,324.3,333.1,344.8,350.3,369.1,356.4)
RES=c(172.9,179.8,180.8,178.6,174.6,172.4,170.6,165.7,154.9,124.1,126.8,
142.2,139.3,134.1,122.3,110.4,101.2,103.4,100.1,115.8,127.8,147.4,161.9,
159.9,170.5,173.1,170.3,169.6,170.3,172.9,175,179.4)
plot(QTR,DUR,type="l") plot(QTR,FRIG,type="l")
```

Exemple 5.3 Le Tableau 5.2 reprend l'indice des prix à la consommation (base 100 en juillet 1970). La Figure 5.3 reprend l'indice brut y_t tel qu'il est présenté dans le Tableau 5.2. La Figure 5.4 présente le rapport mensuel de cet indice y_t/y_{t-1}. Enfin, la Figure 5.5 présente le rapport en glissement annuel y_t/y_{t-12}.

TABLE 5.2 – Indice des prix à la consommation, France (Source : Gouriéroux and Monfort, 1983)

p_t	1970	1971	1972	1973	1974	1975	1976	1977	1978
janvier	97.9	102.5	108.3	115.5	127.4	145.9	159.9	174.3	190.3
février	98.2	103.0	108.9	115.8	129.1	147.0	161.0	175.5	191.7
mars	98.5	103.4	109.4	116.4	130.6	148.2	162.4	177.1	193.4
avril	99.0	104.0	109.8	117.2	132.7	149.5	163.8	179.4	195.5
mai	99.4	104.7	110.4	118.3	134.3	150.6	164.9	181.1	197.4
juin	99.8	105.1	111.0	119.2	135.8	151.7	165.6	182.5	198.9
juillet	100.0	105.6	111.9	120.2	137.5	152.8	167.2	184.1	201.5
août	100.4	106.0	112.5	121.0	138.6	153.8	168.4	185.1	202.5
septembre	100.8	106.5	113.2	122.1	140.1	155.1	170.2	186.7	203.8
octobre	101.2	107.1	114.2	123.4	141.8	156.3	171.8	188.2	205.7
novembre	101.6	107.5	114.9	124.5	143.1	157.3	173.2	188.9	206.8
décembre	101.9	108.0	115.5	125.3	144.3	158.2	173.8	189.4	207.8

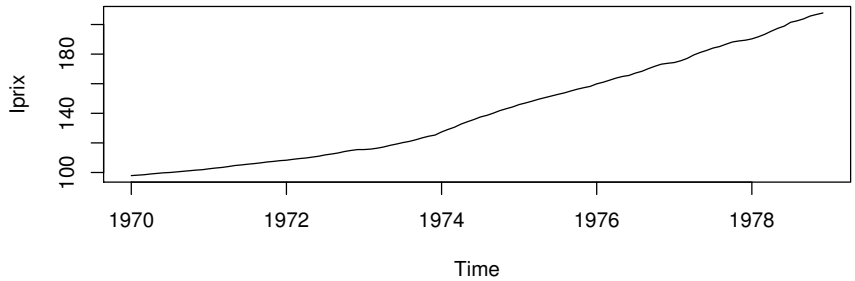


FIGURE 5.3 – Indice des prix à la consommation p_t

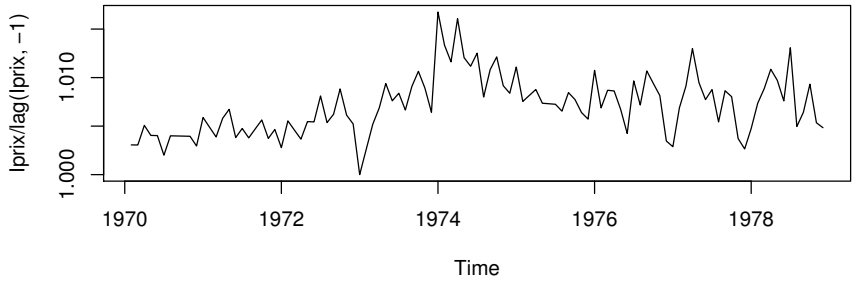


FIGURE 5.4 – Rapport mensuel des indices de prix p_t/p_{t-1}

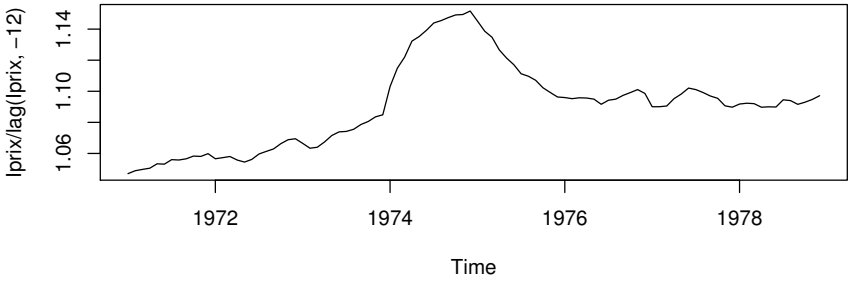


FIGURE 5.5 – Rapport en glissement annuel des indices de prix p_t/p_{t-12}

En langage R

```
# # Indices des prix # Différences d'ordre 1 et 12 #
Iprix=c(97.9,98.2,98.5,99,99.4,99.8,100,100.4,100.8,101.2,101.6,101.9,
102.5,103,103.4,104,104.7,105.1,105.6,106,106.5,107.1,107.5,108,
108.3,108.9,109.4,109.8,110.4,111,111.9,112.5,113.2,114.2,114.9,115.5,
115.5,115.8,116.4,117.2,118.3,119.2,120.2,121,122.1,123.4,124.5,125.3,
127.4,129.1,130.6,132.7,134.3,135.8,137.5,138.6,140.1,141.8,143.1,144.3,
145.9,147,148.2,149.5,150.6,151.7,152.8,153.8,155.1,156.3,157.3,158.2,
159.9,161,162.4,163.8,164.9,165.6,167.2,168.4,170.2,171.8,173.2,173.8,
174.3,175.5,177.1,179.4,181.1,182.5,184.1,185.1,186.7,188.2,188.9,189.4,
190.3,191.7,193.4,195.5,197.4,198.9,201.5,202.5,203.8,205.7,206.8,207.8)
Iprix <- ts(Iprix,start = c(1970, 1), frequency = 12) plot(Iprix)
plot(Iprix/lag(Iprix,-1)) plot(Iprix/lag(Iprix,-12))
```

Exemple 5.4 Données du nombre de voyageurs-kilomètres en deuxième classe exprimées en millions de kilomètres (Source : Gouriéroux and Monfort, 1983).

Exemple 5.5 Hauteur du lac de Neuchâtel (moyenne mensuelle).

Table 5.3 – Trafic du nombre de voyageurs SNCF

mois/année	janv.	fév.	mars	avril	mai	juin	juil.	août	sept.	oct.	nov.	déc.
1963	1750	1560	1820	2090	1910	2410	3140	2850	2090	1850	1630	2420
1964	1710	1600	1800	2120	2100	2460	3200	2960	2190	1870	1770	2270
1965	1670	1640	1770	2190	2020	2610	3190	2860	2140	1870	1760	2360
1966	1810	1640	1860	1990	2110	2500	3030	2900	2160	1940	1750	2330
1967	1850	1590	1880	2210	2110	2480	2880	2670	2100	1920	1670	2520
1968	1834	1792	1860	2138	2115	2485	2581	2639	2038	1936	1784	2391
1969	1798	1850	1981	2085	2120	2491	2834	2725	1932	2085	1856	2553
1970	1854	1823	2005	2418	2219	2722	2912	2771	2153	2136	1910	2537
1971	2008	1835	2120	2304	2264	2175	2928	2738	2178	2137	2009	2546
1972	2084	2034	2152	2522	2318	2684	2971	2759	2267	2152	1978	2723
1973	2081	2112	2279	2661	2281	2929	3089	2803	2296	2210	2135	2862
1974	2223	2248	2421	2710	2505	3021	3327	3044	2607	2525	2160	2876
1975	2481	2428	2596	2923	2795	3287	3598	3118	2875	2754	2588	3266
1976	2667	2668	2804	2806	2976	3430	3705	3053	2764	2802	2707	3307
1977	2706	2586	2796	2978	3053	3463	3649	3095	2839	2966	2863	3375
1978	2820	2857	3306	3333	3141	3512	3744	3179	2984	2950	2896	3611
1979	3313	2644	2872	3267	3391	3682	3937	3284	2849	3085	3043	3541
1980	2848	2913	3248	3250	3375	3640	3771	3259	3206	3269	3181	4008

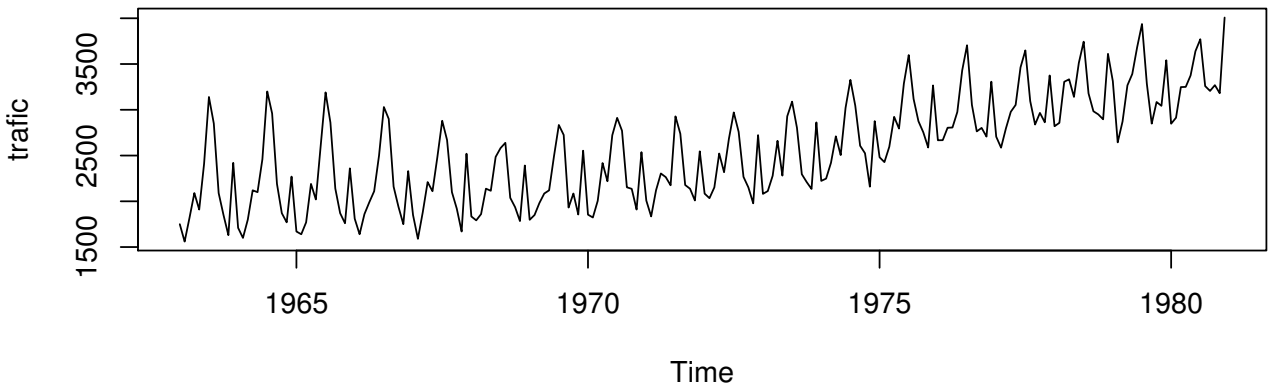


Figure 5.6 – Trafic du nombre de voyageurs SNCF

En langage R

```
S=rbind(a1998=c(429.12,429.06,429.2,429.4,429.39,429.36,429.34,429.33,429.33,429.23,429.27,429.09),
a1999=c(429.05,429.2,429.3,429.36,429.7,429.58,429.45,429.44,429.31,429.25,429.12,429.2),
a2000=c(429.09,429.2,429.3,429.39,429.46,429.38,429.4,429.38,429.27,429.24,429.15,429.09),
a2001=c(429.18,429.12,429.78,429.66,429.54,429.63,429.48,429.38,429.37,429.21,429.03,429.03),
a2002=c(429.03,429.1,429.33,429.36,429.5,429.44,429.45,429.47,429.33,429.32,429.61,429.24),
a2003=c(429.13,429.06,429.21,429.36,429.42,429.38,429.36,429.33,429.23,429.2,429.06,429.02),
a2004=c(429.23,429.09,429.24,429.4,429.42,429.42,429.35,429.44,429.3,429.24,429.08,429.03),
a2005=c(429.03,429.06,429.13,429.43,429.46,429.38,429.33,429.45,429.34,429.21,429.01,428.97),
a2006=c(428.98,428.99,429.25,429.89,429.57,429.42,429.39,429.42,429.35,429.25,429.02,429.05),
a2007=c(429.08,429.1,429.37,429.38,429.42,429.53,429.62,429.73,429.29,429.15,429.03,429.06),
a2008=c(429.05,429.05,429.21,429.42,429.41,429.4,429.39,429.36,429.34,429.17,429.07,429.01),
a2009=c(429.04,429.03,429.13,429.22,429.37,429.34,429.4,429.33,429.18,429.05,429.03,429.09),
a2010=c(429.09,429.04,429.17,429.31,429.33,429.39,429.33,429.39,429.28,429.17,429.13,429.1),
a2011=c(429.06,428.99,428.98,429.13,429.24,429.38,429.38,429.35,429.28,429.22,428.98,429.09),
a2012=c(429.11,428.99,429.13,429.3,429.35,429.43,429.4,429.34,429.33,429.28,429.21,429.27),
a2013=c(429.09,429.14,429.14,429.43,0,0,0,0,0,0,0,0))
SS=ts(c(t(S))[1~: (16*12-8)],frequency=12,start=c(1998,1))
```

Table 5.4 – Hauteur du lac de Neuchâtel : moyennes mensuelles (Source : Office fédéral de l'environnement)

	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
1998	429.12	429.06	429.2	429.4	429.39	429.36	429.34	429.33	429.33	429.23	429.27	429.09
1999	429.05	429.2	429.3	429.36	429.7	429.58	429.45	429.44	429.31	429.25	429.12	429.2
2000	429.09	429.2	429.3	429.39	429.46	429.38	429.4	429.38	429.27	429.24	429.15	429.09
2001	429.18	429.12	429.78	429.66	429.54	429.63	429.48	429.38	429.37	429.21	429.03	429.03
2002	429.03	429.1	429.33	429.36	429.5	429.44	429.45	429.47	429.33	429.32	429.61	429.24
2003	429.13	429.06	429.21	429.36	429.42	429.38	429.36	429.33	429.23	429.2	429.06	429.02
2004	429.23	429.09	429.24	429.4	429.42	429.42	429.35	429.44	429.3	429.24	429.08	429.03
2005	429.03	429.06	429.13	429.43	429.46	429.38	429.33	429.45	429.34	429.21	429.01	428.97
2006	428.98	428.99	429.25	429.89	429.57	429.42	429.39	429.42	429.35	429.25	429.02	429.05
2007	429.08	429.1	429.37	429.38	429.42	429.53	429.62	429.73	429.29	429.15	429.03	429.06
2008	429.05	429.05	429.21	429.42	429.41	429.4	429.39	429.36	429.34	429.17	429.07	429.01
2009	429.04	429.03	429.13	429.22	429.37	429.34	429.4	429.33	429.18	429.05	429.03	429.09
2010	429.09	429.04	429.17	429.31	429.33	429.39	429.33	429.39	429.28	429.17	429.13	429.1
2011	429.06	428.99	428.98	429.13	429.24	429.38	429.38	429.35	429.28	429.22	428.98	429.09
2012	429.11	428.99	429.13	429.3	429.35	429.43	429.4	429.34	429.33	429.28	429.21	429.27
2013	429.09	429.14	429.14	429.43								

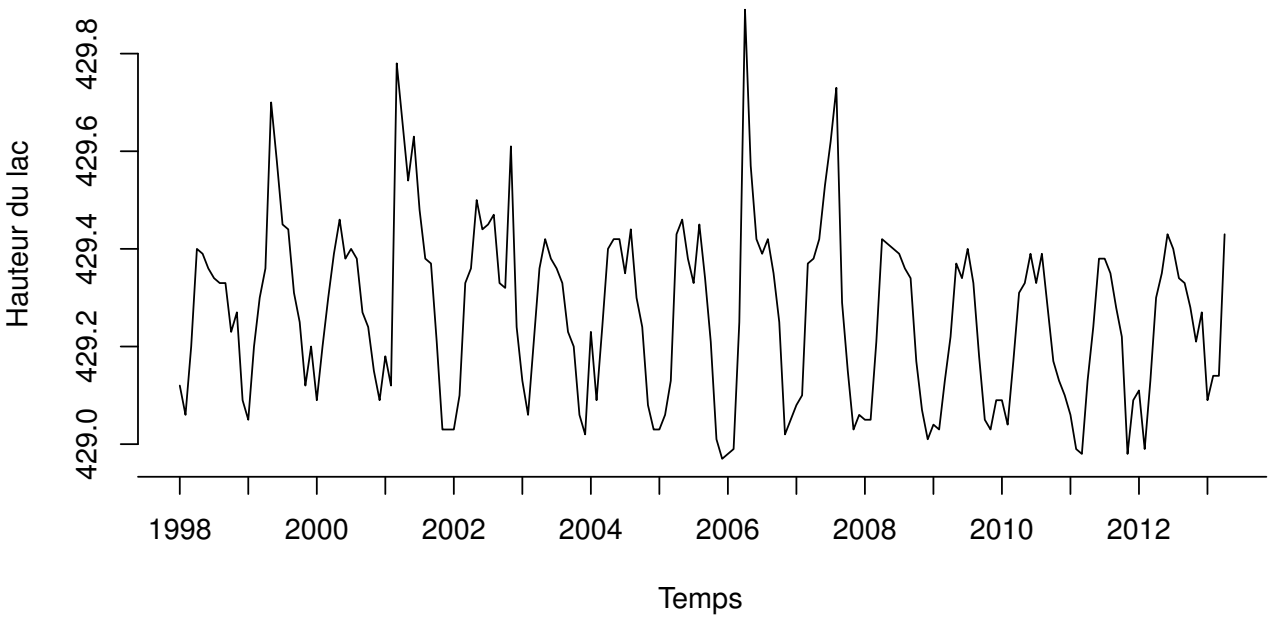


Figure 5.7 – Hauteur du lac de Neuchâtel

```
plot.ts(SS,ylab="Hauteur du lac",xlab="Temps",axes=FALSE)
axis(1,1997:2014)
axis(2)
```

5.2 Description de la tendance

5.2.1 Les principaux modèles

Plusieurs types de modèles peuvent être utilisés pour décrire la tendance.

— Modèles dépendant du temps. La série dépend directement du temps. Le modèle peut être additif

$$y_t = f(t) + E_t,$$

ou multiplicatif

$$y_t = f(t) \times E_t.$$

— Modèles explicatifs statiques : la série chronologique dépend des valeurs prises par une ou plusieurs autres séries chronologiques.

$$y_t = f(x_t) + E_t$$

Le cas linéaire est le plus facile à traiter

$$y_t = b_0 + b_1 x_t + E_t.$$

— Modèles auto-projectifs. La série chronologique au temps t dépend de ses propres valeurs passées

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \ldots, y_{t-p}) + E_t$$

— Modèles explicatifs dynamiques : la série chronologique dépend des valeurs présentes et passées d'une ou de plusieurs autres séries chronologiques, par exemple :

$$y_t = \mu + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_q x_{t-q} + E_t.$$

5.2.2 Tendance linéaire

La tendance la plus simple est linéaire. On peut estimer les paramètres au moyen de la méthode des moindres carrés. C'est une régression simple.

$$T_t = a + bt.$$

5.2.3 Tendance quadratique

On peut utiliser une tendance parabolique. Les paramètres peuvent être estimés au moyen de la méthode des moindres carrés. C'est une régression avec deux variables explicatives.

$$T_t = a + bt + ct^2$$

5.2.4 Tendance polynomiale d'ordre q

On peut ajuster la série par un polynôme d'ordre q. Les paramètres peuvent être estimés au moyen de la méthode des moindres carrés. C'est une régression avec q variables explicatives.

$$T_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_q t^q$$

5.2.5 Tendance logistique

La fonction logistique permet de modéliser des processus ne pouvant dépasser une certaine valeur c (par exemple des taux).

$$T_t = \frac{c}{1 + be^{-at}} \text{ où } a, b, c \in \mathbb{R}^+$$

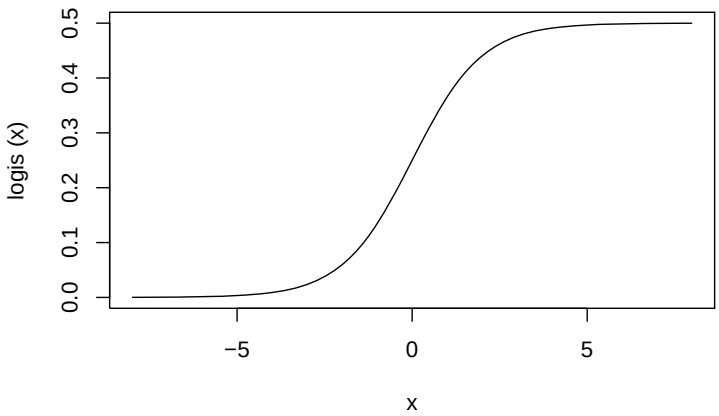


Figure 5.8 – Exemple de fonction logistique avec $c = 0.5$

5.3 Opérateurs de décalage et de différence

5.3.1 Opérateurs de décalage

Afin de simplifier la notation, on utilise des opérateurs de décalage. On définit l'opérateur de décalage "retard" (en anglais *lag operator*) L par

$$Ly_t = y_{t-1},$$

et l'opérateur "avance" (en anglais *forward operator*) F

$$Fy_t = y_{t+1}.$$

L'opérateur identité est donné par

$$Iy_t = y_t.$$

L'opérateur avance est l'inverse de l'opérateur retard

$$FL = LF = I.$$

On peut donc écrire

$$F^{-1} = L \text{ et } L^{-1} = F.$$

On a également
- $L^2 y_t = LLy_t = y_{t-2}$,
- $L^q y_t = y_{t-q}$,
- $F^q y_t = y_{t+q}$,
- $L^0 = F^0 = I$,
- $L^{-q} y_t = F^q y_t = y_{t+q}$.

5.3.2 Opérateur différence

L'opérateur différence d'ordre un est un filtre linéaire

$$\nabla = I - L.$$

L'opérateur différence permet d'enlever une tendance linéaire. En effet, si la série s'écrit

$$y_t = a + b \times t + E_t,$$

alors

$$\nabla y_t = a + b \times t + E_t - a - b \times (t-1) - E_{t-1} = b + E_t - E_{t-1}.$$

Exemple 5.6 On génère une série selon un modèle linéaire dépendant du temps

$$y_t = 10 + 0.3 \times t + E_t, \text{ avec } t = 1, \ldots, 50.$$

La série brute y_t est représentée dans le graphique 5.9 et la différence d'ordre 1 de la série ∇y_t est représentée dans le graphique 5.10.

En langage R

```
# # Tendance linéaire et différence #
lin=10+0.3*(0:50)+rnorm(50,0,1) plot(lin,main="",xlab="",ylab="")
Dlin=diff(lin) plot(Dlin,main="",xlab="",ylab="")
```

On peut construire l'opérateur différence d'ordre deux en élevant ∇ au carré :

$$\nabla^2 = \nabla \times \nabla = I - 2L + L^2$$

L'opérateur différence d'ordre deux permet d'enlever une tendance quadratique. En effet, si la série s'écrit

$$y_t = a + b \times t + c \times t^2 + E_t,$$

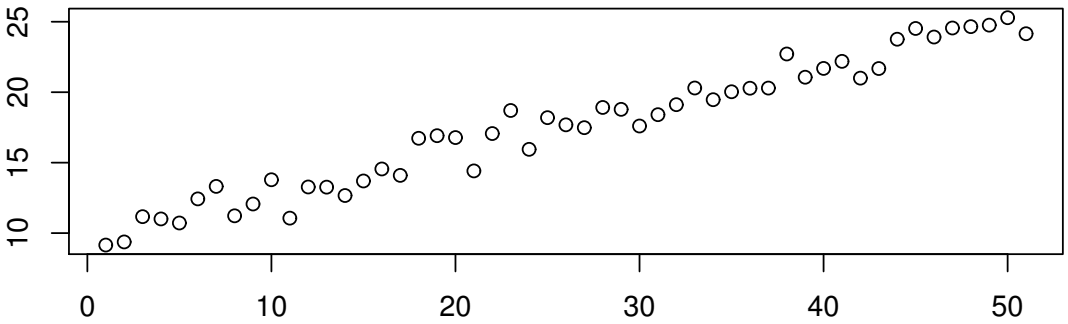


FIGURE 5.9 – Série avec une tendance linéaire dépendant du temps

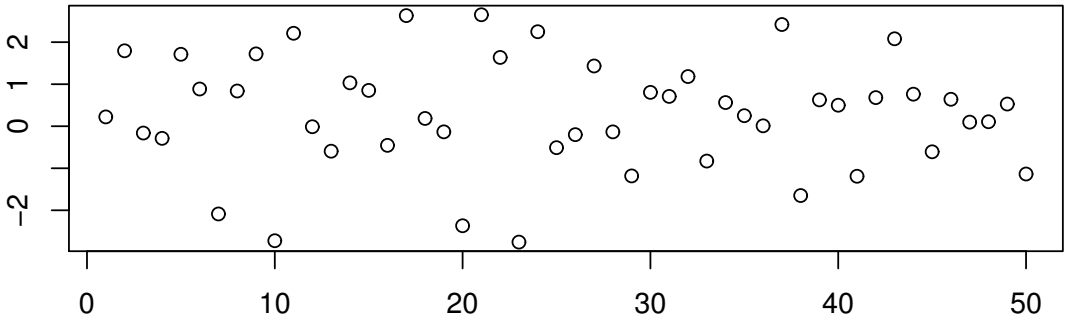


FIGURE 5.10 – Différence d'ordre un de la série avec une tendance linéaire

alors

$$\begin{aligned}
\nabla^2 y_t &= (I - 2L + L^2)y_t \\
&= a + b \times t + c \times t^2 + E_t \\
&\quad -2a - 2b \times (t-1) - 2c \times (t-1)^2 - 2E_{t-1} \\
&\quad +a + b \times (t-2) + c \times (t-2)^2 + E_{t-2} \\
&= 2c + E_t - 2E_{t-1} + E_{t-2}.
\end{aligned}$$

Une tendance polynomiale d'ordre q peut également être supprimée grâce à ∇^q, la différence d'ordre q.

5.3.3 Différence saisonnière

L'opérateur de différence saisonnière s'écrit :

$$\nabla_s = I - L^s,$$

où s vaut 4 pour des données trimestrielles, 7 pour des données journalières et 12 pour des données mensuelles :

Exemple 5.7 Si on applique une différence saisonnière d'ordre 4 sur les données de ventes de réfrigérateurs, la composante saisonnière disparaît.

En langage R

```
# # Vente de réfrigérateurs différence d'ordre 4 #
FRIGm4=FRIG-lag(FRIG,-4) plot(FRIGm4)
```

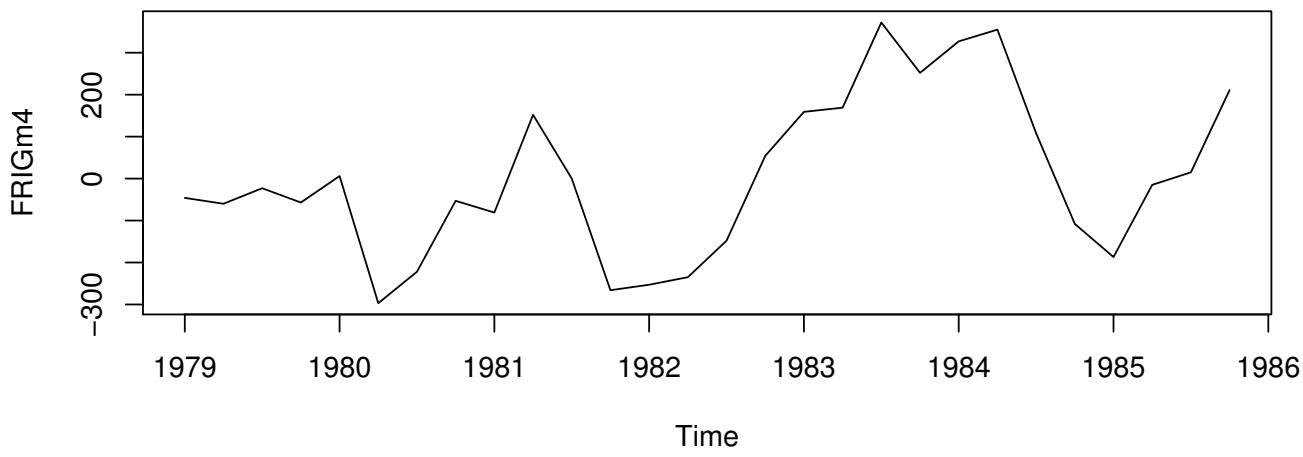


FIGURE 5.11 – Différence d'ordre 4 de la variable vente de 'réfrigérateurs'

Exemple 5.8 Si on applique une différence saisonnière d'ordre 12 sur les données du nombre de voyageurs-kilomètres y_t en deuxième classe exprimées en millions de kilomètres de la SNCF, la tendance saisonnière disparaît (voir Figure 5.13). On a ainsi la nouvelle variable

$$z_t = \nabla_{12} y_t = (I - L^{12}) y_t = y_t - y_{t-12}.$$

Une autre manière de faire consiste à prendre le logarithme de la variable et ensuite à calculer la différence, ce qui revient à prendre le logarithme du rapport de la variable (voir Figure 5.14). On définit ainsi une nouvelle variable v_t :

$$v_t = \nabla_{12} \log y_t = (I - L^{12}) \log y_t = \log y_t - \log y_{t-12} = \log \frac{y_t}{y_{t-12}}.$$

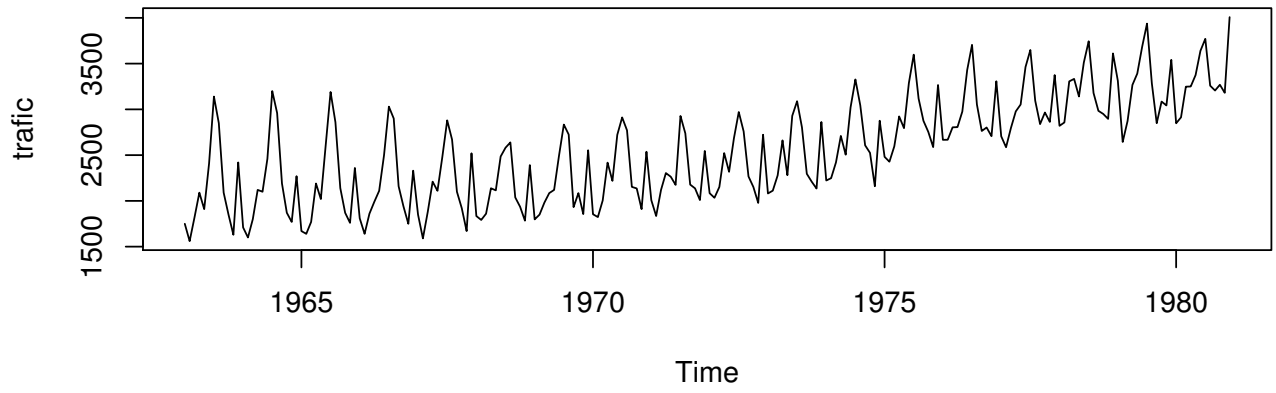


FIGURE 5.12 – Trafic du nombre de voyageurs SNCF

En langage R

```
trafic=c(1750,1560,1820,2090,1910,2410,3140,2850,2090,1850,1630,2420,
1710,1600,1800,2120,2100,2460,3200,2960,2190,1870,1770,2270,
1670,1640,1770,2190,2020,2610,3190,2860,2140,1870,1760,2360,
1810,1640,1860,1990,2110,2500,3030,2900,2160,1940,1750,2330,
1850,1590,1880,2210,2110,2480,2880,2670,2100,1920,1670,2520,
1834,1792,1860,2138,2115,2485,2581,2639,2038,1936,1784,2391,
```

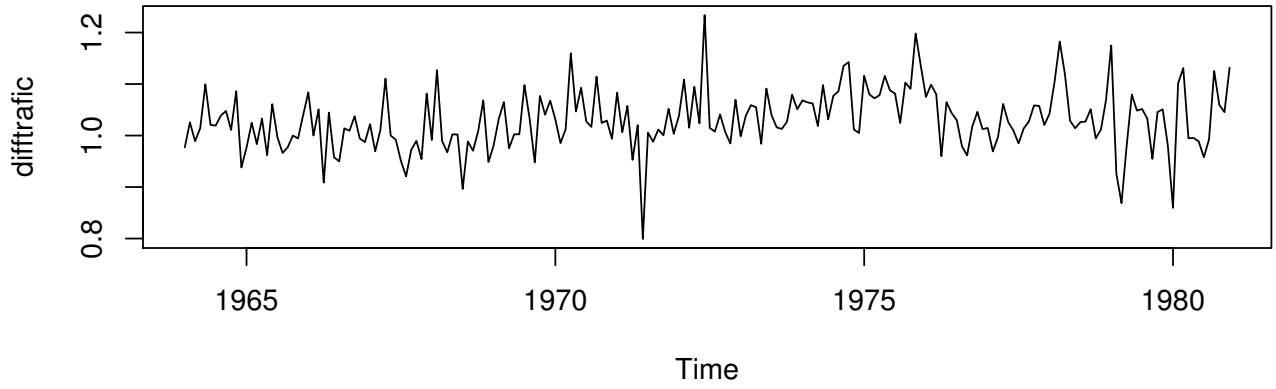


FIGURE 5.13 – Différence d'ordre 12 sur la série trafic du nombre de voyageurs SNCF

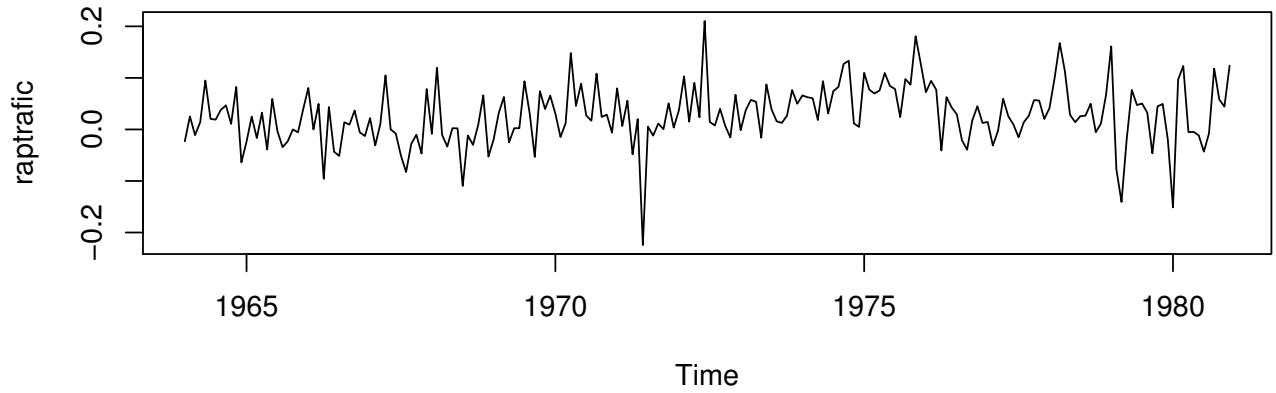


FIGURE 5.14 – Logarithme du rapport d'ordre 12 sur la série trafic du nombre de voyageurs SNCF

```
1798,1850,1981,2085,2120,2491,2834,2725,1932,2085,1856,2553,
1854,1823,2005,2418,2219,2722,2912,2771,2153,2136,1910,2537,
2008,1835,2120,2304,2264,2175,2928,2738,2178,2137,2009,2546,
2084,2034,2152,2522,2318,2684,2971,2759,2267,2152,1978,2723,
2081,2112,2279,2661,2281,2929,3089,2803,2296,2210,2135,2862,
2223,2248,2421,2710,2505,3021,3327,3044,2607,2525,2160,2876,
2481,2428,2596,2923,2795,3287,3598,3118,2875,2754,2588,3266,
2667,2668,2804,2806,2976,3430,3705,3053,2764,2802,2707,3307,
2706,2586,2796,2978,3053,3463,3649,3095,2839,2966,2863,3375,
2820,2857,3306,3333,3141,3512,3744,3179,2984,2950,2896,3611,
3313,2644,2872,3267,3391,3682,3937,3284,2849,3085,3043,3541,
2848,2913,3248,3250,3375,3640,3771,3259,3206,3269,3181,4008)
trafic <- ts(trafic,start = c(1963, 1), frequency = 12)
plot(trafic) difftrafic=trafic-lag(trafic,-12) plot(difftrafic)
raptrafic=log(trafic/lag(trafic,-12)) plot(raptrafic)
```

5.4 Filtres linéaires et moyennes mobiles

5.4.1 Filtres linéaires

Un filtre linéaire d'ordre $m = p_1 + p_2 + 1$ est défini par

$$\begin{aligned} FL &= \sum_{j=-p_1}^{p_2} w_j L^{-j} \\ &= w_{-p_1} L^{p_1} + w_{-p_1+1} L^{p_1-1} + \cdots + w_{-1} L + w_0 I + w_1 F + \cdots + w_{p_2-1} F^{p_2-1} + w_{p_2} F^{p_2}, \end{aligned}$$

où $p_1, p_2 \in \mathbb{N}$ et $w_j \in \mathbb{R}$.

5.4.2 Moyennes mobiles : définition

Une moyenne mobile d'ordre $m = p_1 + p_2 + 1$ est un filtre linéaire tel que

$$\sum_{j=-p_1}^{p_2} w_j = 1, \text{ pour tout } j = -p_1, \ldots, p_2.$$

Beaucoup de moyennes mobiles ont des poids w_j positifs, mais pas toutes.

Une moyenne mobile est symétrique si $p_1 = p_2 = p$, et

$$w_j = w_{-j}, \text{ pour tout } j = 1, \ldots, p.$$

Une moyenne mobile symétrique est dite non-pondérée si

$$w_j = cst \text{ pour tout } j = -p_1, \ldots, p_2.$$

5.4.3 Moyenne mobile et composante saisonnière

Une moyenne mobile est un outil intéressant pour lisser une série temporelle et donc pour enlever une composante saisonnière. On utilise de préférence des moyennes mobiles non-pondérées d'ordre égal à la période, par exemple d'ordre 7 pour des données journalières, d'ordre 12 pour des données mensuelles. Par exemple, pour enlever la composante saisonnière due au jour de la semaine, on peut appliquer une moyenne mobile non-pondérée d'ordre 7.

$$\text{MM}(7) = \frac{1}{7}\left(L^3 + L^2 + L + I + F + F^2 + F^3\right).$$

Cette moyenne mobile accorde le même poids à chaque jour de la semaine. En effet,

$$\text{MM}(7) y_t = \frac{1}{7}\left(y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3}\right).$$

Pour les composantes saisonnières d'une période paire, il n'existe pas de moyennes mobiles centrées non-pondérées. Il existe deux types de moyenne mobile centrée pondérée :

— Si la période est paire et égale à m ($m = 4$ pour des données trimestrielles), on utilise une moyenne mobile d'ordre impair accordant un demi-poids aux deux extrémités. Par exemple, pour des données trimestrielles, la moyenne mobile est définie par

$$\text{MM}(4) = \frac{1}{8}\left(L^2 + 2L + 2I + 2F + F^2\right).$$

Ainsi, chaque trimestre conserve le même poids. En effet,

$$\text{MM}(4) y_t = \frac{1}{8}\left(y_{t-2} + 2y_{t-1} + 2y_t + 2y_{t+1} + y_{t+2}\right).$$

— Si la période est paire et égale à m, on peut aussi utiliser la composée de deux moyennes mobiles non-pondérées et non-centrées afin d'obtenir une moyenne mobile centrée :

$$\text{MMC} = \frac{1}{4}\left(L^2 + L + I + F\right)\frac{1}{4}\left(L + I + F + F^2\right) = \frac{1}{16}\left(L^3 + 2L^2 + 3L + 4I + 3F + 2F^2 + F^3\right).$$

À nouveau, chaque trimestre est affecté du même poids, mais cette méthode est moins avantageuse car la moyenne mobile est plus étendue. Donc, plus des données seront "perdues" aux extrémités de la séries.

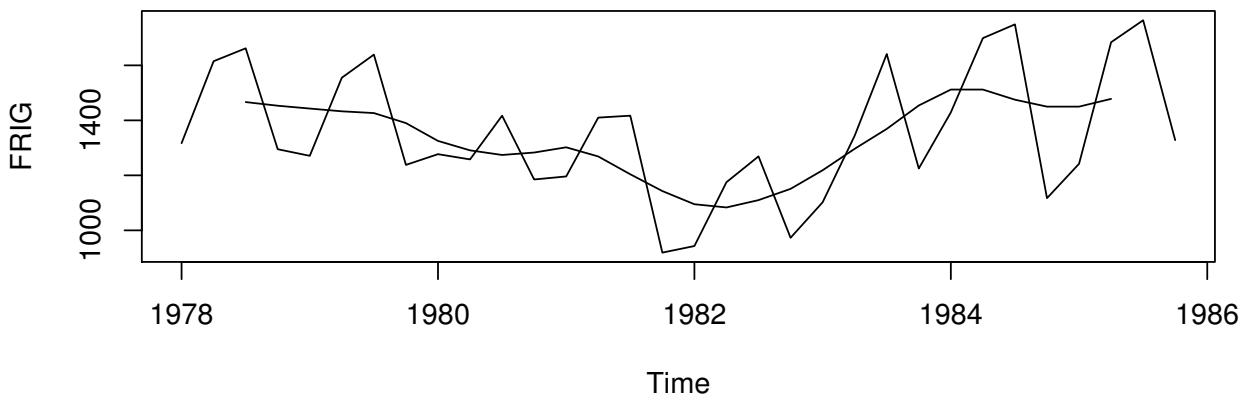


FIGURE 5.15 – Nombre de réfrigérateurs et moyenne mobile d'ordre 4

Exemple 5.9 La variable "réfrigérateur" est lissée grâce à une moyenne mobile qui accorde le même coefficient de pondération à chaque trimestre.

En langage R

```
dec=decompose(FRIG)
moving_average= dec$trend plot(FRIG)
lines(moving_average)
```

Une moyenne mobile qui accorde le même poids à chaque saison permet d'enlever une tendance saisonnière.

5.5 Moyennes mobiles particulières

5.5.1 Moyenne mobile de Van Hann

$$\mathrm{MM}_{VH} = \frac{1}{2}(I+F) \times \frac{1}{2}(L+I) = \frac{1}{4}(L+2I+F)$$

5.5.2 Moyenne mobile de Spencer

$$\begin{aligned}\mathrm{MM}_S &= \frac{1}{4}(L+I+F+F^2) \times \frac{1}{4}(L^2+L+I+F) \times \frac{1}{5}(L^2+L+I+F+F^2) \times \frac{1}{4}(-3L^2+3L+4I+3F-3F^2) \\ &= \frac{1}{320}(-3L^7-6L^6-5L^5+3L^4+21L^3+46L^2+67L+74I+67F+46F^2+21F^3+3F^4-5F^5-6F^6-3F^7)\end{aligned}$$

La moyenne mobile de Spencer supprime les composantes saisonnières de période 4 et 5 et conserve les tendances polynomiales jusqu'à l'ordre 3.

5.5.3 Moyenne mobile de Henderson

Les moyennes mobiles d'Henderson conservent les tendances polynomiales de degré 2 tout en conservant une "souplesse" aux coefficients de la moyenne mobile. La souplesse est obtenue en minimisant la quantité

$$\sum_j (I-L)^3 \theta_j.$$

Moyenne mobile de Henderson d'ordre $2m-3$, **où** $m \geq 4$

$$\mathrm{MM}_H = \sum_{j=-m-1}^{m+1} \theta_j L^j,$$

où

$$\theta_i = \frac{315((m-1)^2-i^2)(m^2-i^2)((m+1)^2-i^2)(3m^2-16-11i^2)}{8m(m^2-1)(4m^2-1)(4m^2-9)(4m^2-25)}.$$

Moyenne mobile de Henderson d'ordre $2m-3=5$ $(m=4)$

$$\frac{1}{286}(-21L^2+84L+160I+84F-21F^2).$$

Moyenne mobile de Henderson d'ordre $2m-3=9$ $(m=6)$

$$\frac{1}{2431}(-99L^4-24L^3-288L^2+648L+805I+648F+288F^2-24F^3-99F^4).$$

Moyenne mobile de Henderson d'ordre $2m-3=11$ $(m=7)$

$$\begin{aligned}\frac{1}{92378}(-2574L^5 &- 2475L^4+3300L^3+13050L^2+22050L+25676I\\ &+ 22050F+13050F^2+3300F^3-2475F^4-2574F^5)\end{aligned}$$

Moyenne mobile de Henderson d'ordre $2m-3=15$ $(m=9)$

$$\begin{aligned}\frac{1}{193154}(-2652L^7 &- 4732L^6-2730L^5+4641L^4+16016L^3+28182L^2+37422L+40860I\\ &+ 37422F+28182F^2+16016F^3+4641F^4-2730F^5-4732F^6-2652F^7)\end{aligned}$$

5.5.4 Médianes mobiles

Si les données contiennent des valeurs aberrantes ou extrêmes, on peut remplacer la moyenne mobile par une médiane mobile. Par exemple la médiane mobile d'ordre 5 est définie par :

$$Med(5)_t = \text{Médiane}(y_{t-2}, y_{t-1}, y_t, y_{t+1}, y_{t+2}).$$

5.6 Désaisonnalisation

5.6.1 Méthode additive

Soit une série temporelle régie par un modèle additif du type

$$Y_{am} = T_{am} + S_m + E_{am},$$

où $a = 1, \ldots, A$, représente par exemple l'année et $m = 1, \ldots, M$ représente par exemple le mois. La tendance est supposée connue soit par un ajustement, soit par une moyenne mobile. On isole la composante saisonnière en faisant, pour chaque mois, la moyenne des différences entre les valeurs observées et la tendance

$$S_m = \frac{1}{A-1}\sum_a (Y_{am} - T_{am}).$$

En général, on ne dispose pas du même nombre d'observations, pour chaque mois. On procède à un ajustement afin que la somme des composantes saisonnières soit égale à zéro :

$$S'_m = S_m - \frac{1}{M}\sum_m S_m.$$

On peut ensuite procéder à la désaisonnalisation de la série par

$$\widetilde{Y}_{am} = Y_{am} - S'_m.$$

5.6.2 Méthode multiplicative

Soit une série temporelle régie par un modèle multiplicatif du type

$$Y_{am} = T_{am} \times S_m \times E_{am},$$

où $a = 1, \ldots, A$ représente par exemple l'année et $m = 1, .., M$ représente par exemple le mois. La tendance est supposée connue soit par un ajustement, soit par une moyenne mobile.

On isole la composante saisonnière en faisant, pour chaque mois, la moyenne des rapports entre les valeurs observées et la tendance :

$$S_m = \frac{1}{A-1} \sum_a \frac{Y_{am}}{T_{am}}.$$

À nouveau, on réalise un ajustement afin que la moyenne des composantes saisonnières soit égale à 1. On corrige donc les coefficients S_m par

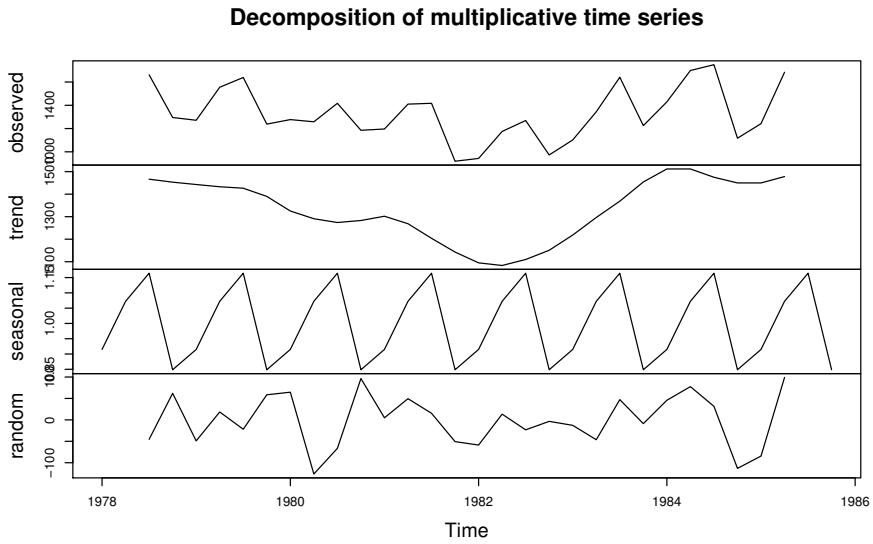
$$S'_m = S_m \frac{1}{\frac{1}{M} \sum_m S_m}.$$

La désaisonnalisation se réalise alors par une division

$$\widetilde{Y}_{am} = \frac{Y_{am}}{S'_m} = T_{am} \times E_{am}$$

Exemple 5.10 L'objectif est de désaisonnaliser la série trimestrielle des ventes de réfrigérateurs. Le Tableau 5.5 contient la variable 'vente de réfrigérateurs', la moyenne mobile d'ordre 4, la composante saisonnière et série désaisonnalisée au moyen de la méthode additive. Le Tableau 5.7 présente la désaisonnalisation au moyen de la méthode multiplicative.

FIGURE 5.16 – Décomposition de la série de ventes de réfrigérateurs 5.1



En langage R

```
deco=decompose(FRIG,type="multiplicative")
plot(deco)
```

Table 5.5 – Décomposition de la variable FRIG, méthode additive

QTR	FRIG	MM	FRIG-MM	Desaison
1	1317			1442.58
2	1615			1505.13
3	1662	1466.50	195.50	1451.20
4	1295	1453.25	-158.25	1490.09
5	1271	1442.88	-171.88	1396.58
6	1555	1432.88	122.13	1445.13
7	1639	1426.50	212.50	1428.20
8	1238	1390.13	-152.13	1433.09
9	1277	1325.25	-48.25	1402.58
10	1258	1290.88	-32.88	1148.13
11	1417	1274.13	142.88	1206.20
12	1185	1283.00	-98.00	1380.09
13	1196	1302.00	-106.00	1321.58
14	1410	1268.75	141.25	1300.13
15	1417	1203.88	213.13	1206.20
16	919	1142.88	-223.88	1114.09
17	943	1095.00	-152.00	1068.58
18	1175	1083.25	91.75	1065.13
19	1269	1109.88	159.13	1058.20
20	973	1150.88	-177.88	1168.09
21	1102	1218.50	-116.50	1227.58
22	1344	1296.50	47.50	1234.13
23	1641	1368.88	272.13	1430.20
24	1225	1454.13	-229.13	1420.09
25	1429	1512.00	-83.00	1554.58
26	1699	1512.00	187.00	1589.13
27	1749	1475.13	273.88	1538.20
28	1117	1449.88	-332.88	1312.09
29	1242	1449.88	-207.88	1367.58
30	1684	1478.13	205.88	1574.13
31	1764			1553.20
32	1328			1523.09

Table 5.6 – Moyenne des composantes saisonnières

S_1	-126.50	S'_1	-125.58
S_2	108.95	S'_2	109.87
S_3	209.88	S'_3	210.80
S_4	-196.02	S'_4	-195.09
Total	-3.70	Total	0.00

5.7 Lissage exponentiel

5.7.1 Lissage exponentiel simple

Une manière simple de réaliser une prédiction est de réaliser un lissage exponentiel simple. On suppose que l'on dispose de T observations $X_1, \ldots, X_T$ indicées par les dates $1, \ldots, T$. On veut réaliser une prédiction pour les dates suivantes $T + k$, $k \geq 1$. La prédiction faite à la date T pour la date $T + k$ est notée $\widehat{X}_T(k)$ (prédiction au temps T et à l'horizon k). Le lissage exponentiel simple donne une prédiction à l'horizon 1 et consiste à réaliser une moyenne des valeurs passées en affectant des poids moins importants aux valeurs qui sont éloignées de la

Table 5.7 – Décomposition de la variable FRIG, méthode multiplicative

QTR	FRIG	MM	FRIG/MM	Desaison
1	1317			1453.85
2	1615			1493.76
3	1662	1466.50	1.13	1434.00
4	1295	1453.25	0.89	1516.45
5	1271	1442.88	0.88	1403.07
6	1555	1432.88	1.09	1438.26
7	1639	1426.50	1.15	1414.15
8	1238	1390.13	0.89	1449.70
9	1277	1325.25	0.96	1409.70
10	1258	1290.88	0.97	1163.56
11	1417	1274.13	1.11	1222.61
12	1185	1283.00	0.92	1387.64
13	1196	1302.00	0.92	1320.28
14	1410	1268.75	1.11	1304.15
15	1417	1203.88	1.18	1222.61
16	919	1142.88	0.80	1076.15
17	943	1095.00	0.86	1040.99
18	1175	1083.25	1.08	1086.79
19	1269	1109.88	1.14	1094.91
20	973	1150.88	0.85	1139.39
21	1102	1218.50	0.90	1216.51
22	1344	1296.50	1.04	1243.10
23	1641	1368.88	1.20	1415.88
24	1225	1454.13	0.84	1434.48
25	1429	1512.00	0.95	1577.49
26	1699	1512.00	1.12	1571.45
27	1749	1475.13	1.19	1509.06
28	1117	1449.88	0.77	1308.01
29	1242	1449.88	0.86	1371.06
30	1684	1478.13	1.14	1557.58
31	1764			1522.01
32	1328			1555.09

Table 5.8 – Moyenne des composantes saisonnières

S_1	0.90	S'_1	0.91
S_2	1.08	S'_2	1.08
S_3	1.16	S'_3	1.16
S_4	0.85	S'_4	0.85
Total	3.99	Total	4.00

prédiction :

$$\widehat{X}_T(1) = (1-\beta)\sum_{j=0}^{T-1}\beta^j X_{T-j} = (1-\beta)\sum_{j=0}^{T-1}\beta^j L^j X_T,$$

où β est un coefficient appartenant à $]0,1[$. Comme

$$\widehat{X}_{T-1}(1) = (1-\beta)\sum_{j=0}^{T-2}\beta^j X_{T-1-j} = \frac{(1-\beta)}{\beta}\sum_{j=1}^{T-1}\beta^j X_{T-j},$$

on a

$$\widehat{X}_T(1) = (1-\beta)\sum_{j=0}^{T-1}\beta^j X_{T-j} = (1-\beta)X_T + \beta\widehat{X}_{T-1}(1).$$

Cette formule peut être utilisée pour mettre à jour le lissage exponentiel simple.

Afin d'initialiser le lissage exponentiel on peut prendre

$$\widehat{X}_0(1) = X_1.$$

Le lissage exponentiel simple est adapté au cas ou la série peut être ajustée par une droite horizontale. Autrement dit, on suppose que

$$X_T \approx a.$$

Le lissage exponentiel peut être obtenu au moyen de la méthode des moindres carrés en minimisant en a le critère

$$Q = \sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - a\right)^2.$$

En annulant la dérivée par rapport à a, on obtient

$$2\sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - a\right) = 0,$$

ce qui donne

$$\widehat{X}_T(1) = a = \frac{\sum_{j=0}^{T-1} \beta^j X_{T-j}}{\sum_{j=0}^{T-1} \beta^j} \approx (1-\beta)\sum_{j=0}^{T-1} \beta^j X_{T-j}.$$

On peut choisir β sur base de critères subjectifs, cependant on peut également déterminer une valeur optimale au moyen de la méthode des moindres carrés. On minimise alors en β :

$$\sum_{j=0}^{T-1} \left(X_{T-j} - \widehat{X}_{T-j-1}(1)\right)^2,$$

ce qui aboutit à un système non-linéaire qu'il est cependant possible de résoudre numériquement.

5.7.2 Lissage exponentiel double

Si la série peut être ajustée par une droite quelconque de type $a + b(t - T)$. On applique alors un lissage exponentiel double pour obtenir la prédiction

$$\widehat{X}_T(k) = a + bk.$$

Comme

$$\widehat{X}_T(-j) = a - bj,$$

on obtient les valeurs de a et b au moyen de la méthode des moindres carrés en minimisant en a et b le critère

$$Q = \sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - \widehat{X}_T(-j)\right)^2 = \sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - a + bj\right)^2.$$

En annulant les dérivées partielles par rapport à a et b, on obtient

$$\begin{cases} 2\displaystyle\sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - a + bj\right) = 0 \\ 2\displaystyle\sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - a + bj\right) j = 0. \end{cases}$$

ce qui donne

$$\begin{cases} \displaystyle\sum_{j=0}^{T-1} \beta^j X_{T-j} - a\sum_{j=0}^{T-1} \beta^j + b\sum_{j=0}^{T-1} j\beta^j = 0 \\ \displaystyle\sum_{j=0}^{T-1} j\beta^j X_{T-j} - a\sum_{j=0}^{T-1} j\beta^j + b\sum_{j=0}^{T-1} j^2\beta^j = 0. \end{cases}$$

Comme on a

$$\sum_{j=0}^{\infty} \beta^j = \frac{1}{1-\beta} \qquad \sum_{j=0}^{\infty} j\beta^j = \frac{\beta}{(1-\beta)^2} \qquad \sum_{j=0}^{\infty} j^2\beta^j = \frac{\beta(1+\beta)}{(1-\beta)^3},$$

on obtient

$$\begin{cases} \displaystyle\sum_{j=0}^{T-1} \beta^j X_{T-j} - \frac{a}{1-\beta} + \frac{b\beta}{(1-\beta)^2} = 0 \\ \displaystyle\sum_{j=0}^{T-1} j\beta^j X_{T-j} - \frac{a\beta}{(1-\beta)^2} + \frac{b\beta(1+\beta)}{(1-\beta)^3} = 0. \end{cases} \tag{5.1}$$

En notant maintenant S_T^1 la série lissée

$$S_T^1 = (1-\beta) \sum_{j=0}^{T-1} \beta^j X_{T-j},$$

et S_T^2 la série doublement lissée,

$$\begin{aligned} S_T^2 &= (1-\beta) \sum_{j=0}^{T-1} \beta^j S_{T-j}^1 = (1-\beta) \sum_{j=0}^{T-1} \beta^j (1-\beta) \sum_{i=0}^{T-1-j} \beta^i X_{T-j-i} = (1-\beta)^2 \sum_{j=0}^{T-1} \sum_{i=0}^{T-1-j} \beta^{i+j} X_{T-j-i} \\ &= (1-\beta)^2 \sum_{k=0}^{T-1} (k+1)\beta^k X_{T-k} = (1-\beta)^2 \sum_{k=0}^{T-1} k\beta^k X_{T-k} + (1-\beta) S_T^1. \end{aligned}$$

On obtient finalement

$$\sum_{k=0}^{T-1} k\beta^k X_{T-k} = \frac{S_T^2}{(1-\beta)^2} - \frac{S_T^1}{(1-\beta)^1}.$$

Le système (5.1) peut alors s'écrire

$$\begin{cases} \dfrac{S_T^1}{1-\beta} - \dfrac{a}{1-\beta} + \dfrac{b\beta}{(1-\beta)^2} = 0 \\ \dfrac{S_T^2}{(1-\beta)^2} - \dfrac{S_T^1}{1-\beta} - \dfrac{a\beta}{(1-\beta)^2} + \dfrac{b\beta(1+\beta)}{(1-\beta)^3} = 0. \end{cases}$$

En résolvant ce système en a et b, on obtient finalement

$$\begin{cases} a = 2S_T^1 - S_T^2 \\ b = \dfrac{1-\beta}{\beta}(S_T^1 - S_T^2). \end{cases}$$

Exemple 5.11 Le Tableau 5.9 rend compte du prix moyen du mazout pour 100 ℓ (achat entre 800 et 1500 ℓ) en CHF pour chaque mois de 2004 à 2007 (Source : Office fédéral de la statistique, 2008).

Nous allons effectuer un lissage exponentiel double avec $\beta = 0.7$.

— On réalise d'abord un premier lissage en utilisant la formule récursive

$$\widehat{X}_t(1) = (1-\beta) X_t + \beta \widehat{X}_{t-1}(1),$$

$$\widehat{X}_0(1) = X_1,$$

où $S_t^1 = \widehat{X}_t(1)$.

On obtient :

$$S_1^1 = \widehat{X}_1(1) = (1-\beta)X_1 + \beta\widehat{X}_0(1) = (1-0.7)X_1 + 0.7X_1 = X_1 = 54.23,$$

$$S_2^1 = \widehat{X}_2(1) = (1-\beta)X_2 + \beta\widehat{X}_1(1) = 0.3 \times 51.51 + 0.7 \times 54.23 = 53.414,$$

$$S_3^1 = \widehat{X}_3(1) = (1-\beta)X_3 + \beta\widehat{X}_2(1) = 0.3 \times 55.60 + 0.7 \times 53.41 = 54.070,$$

et ainsi de suite.

TABLE 5.9 – Prix moyen du Mazout pour 100 ℓ (achat entre 800 et 1500 ℓ)

mois/année	2004	2005	2006	2007
janvier	54.23	63.00	86.16	79.39
février	51.51	67.32	88.70	81.32
mars	55.60	75.52	88.92	82.06
avril	55.72	79.83	92.58	88.05
mai	58.71	73.22	93.65	88.24
juin	58.82	75.38	91.88	88.95
juillet	58.41	83.97	95.35	92.10
août	64.92	84.23	95.83	91.65
septembre	63.95	97.29	91.16	95.35
octobre	72.98	99.31	87.63	97.54
novembre	70.25	89.88	84.57	106.94
décembre	68.24	87.18	84.10	108.94

— On réalise ensuite un second lissage que l'on applique à la série lissée.

$$S_t^2 = (1-\beta)S_t^1 + \beta S_{t-1}^2,$$

$$S_0^2 = S_1^1.$$

On obtient :

$$S_1^2 = (1-\beta)S_1^1 + \beta S_0^2 = (1-\beta)S_1^1 + \beta S_1^1 = S_1^1 = 54.23,$$

$$S_2^2 = (1-\beta)S_2^1 + \beta S_1^2 = 0.3 \times 53.414 + 0.7 \times 54.23 = 53.99,$$

$$S_3^2 = (1-\beta)S_3^1 + \beta S_2^2 = 0.3 \times 54.070 + 0.7 \times 53.99 = 54.01,$$

et ainsi de suite.

— On cherche alors $\widehat{X}_t(k) = a + bk$ pour chaque t. On prend ici $k = 1$, $\widehat{X}_t(1) = a + b$ avec :

$$\begin{aligned} a &= 2S_t^1 - S_t^2 \\ b &= \frac{1-\beta}{\beta}\left(S_t^1 - S_t^2\right) = \frac{0.3}{0.7}\left(S_t^1 - S_t^2\right). \end{aligned}$$

Le Tableau 5.10 rend compte des résultats pour les années 2004 à 2007. La Figure 5.17 représente la série initiale, le lissage exponentiel simple et le lissage exponentiel double.

TABLE 5.10 – Lissage exponentiel simple et double de la série temporelle Prix moyen du Mazout pour 100 litres (achat entre 800 et 1500 litres) en CHF

Année	mois	X_t	$S_t^1 = \widehat{X}_{LES}(1)$	S_t^2	a	b	$\widehat{X}_{LED}(1) = a + b$
2004	1	54.23	54.23	54.23	54.23	0	54.23
	2	51.51	54.23	54.23	54.23	0	54.23
	3	55.60	53.41	54.23	52.60	-0.350	52.25
	4	55.72	54.07	53.99	54.15	0.036	54.19
	5	58.71	54.56	54.01	55.12	0.238	55.36
	6	58.82	55.81	54.18	57.44	0.699	58.14
	7	58.41	56.71	54.67	58.76	0.877	59.63
	8	64.92	57.22	55.28	59.16	0.832	59.99
	9	63.95	59.53	55.86	63.20	1.572	64.77
	10	72.98	60.86	56.96	64.75	1.669	66.42
	11	70.25	64.49	58.13	70.86	2.727	73.58
	12	68.24	66.22	60.04	72.40	2.649	75.05
2005	1	63.00	66.83	61.89	71.76	2.114	73.87
	2	67.32	65.68	63.37	67.98	0.988	68.97
	3	75.52	66.17	64.07	68.28	0.902	69.18
	4	79.83	68.98	64.70	73.25	1.834	75.09
	5	73.22	72.23	65.98	78.48	2.679	81.16
	6	75.38	72.53	67.86	77.20	2.002	79.20
	7	83.97	73.38	69.26	77.51	1.768	79.28
	8	84.23	76.56	70.50	82.62	2.599	85.22
	9	97.29	78.86	72.31	85.41	2.805	88.21
	10	99.31	84.39	74.28	94.50	4.333	98.83
	11	89.88	88.87	77.31	100.42	4.952	105.37
	12	87.18	89.17	80.78	97.56	3.597	101.16
2006	1	86.16	88.57	83.30	93.85	2.262	96.11
	2	88.70	87.85	84.88	90.82	1.273	92.09
	3	88.92	88.10	85.77	90.44	1.000	91.44
	4	92.58	88.35	86.47	90.23	0.805	91.03
	5	93.65	89.62	87.03	92.20	1.108	93.31
	6	91.88	90.83	87.81	93.85	1.294	95.14
	7	95.35	91.14	88.71	93.57	1.041	94.61
	8	95.83	92.41	89.44	95.37	1.269	96.64
	9	91.16	93.43	90.33	96.53	1.329	97.86
	10	87.63	92.75	91.26	94.24	0.638	94.88
	11	84.57	91.21	91.71	90.72	-0.212	90.51
	12	84.10	89.22	91.56	86.88	-1.003	85.88
2007	1	79.39	87.68	90.86	84.51	-1.360	83.15
	2	81.32	85.20	89.91	80.49	-2.019	78.47
	3	82.06	84.03	88.49	79.57	-1.911	77.66
	4	88.05	83.44	87.16	79.73	-1.592	78.14
	5	88.24	84.82	86.04	83.61	-0.522	83.09
	6	88.95	85.85	85.68	86.02	0.074	86.10
	7	92.10	86.78	85.73	87.83	0.451	88.28
	8	91.65	88.38	86.04	90.71	0.999	91.71
	9	95.35	89.36	86.74	91.97	1.121	93.09
	10	97.54	91.16	87.53	94.78	1.555	96.34
	11	106.94	93.07	88.62	97.53	1.909	99.44
	12	108.94	97.23	89.95	104.51	3.120	107.63
2008	1		100.74	92.14	109.35	3.689	113.04

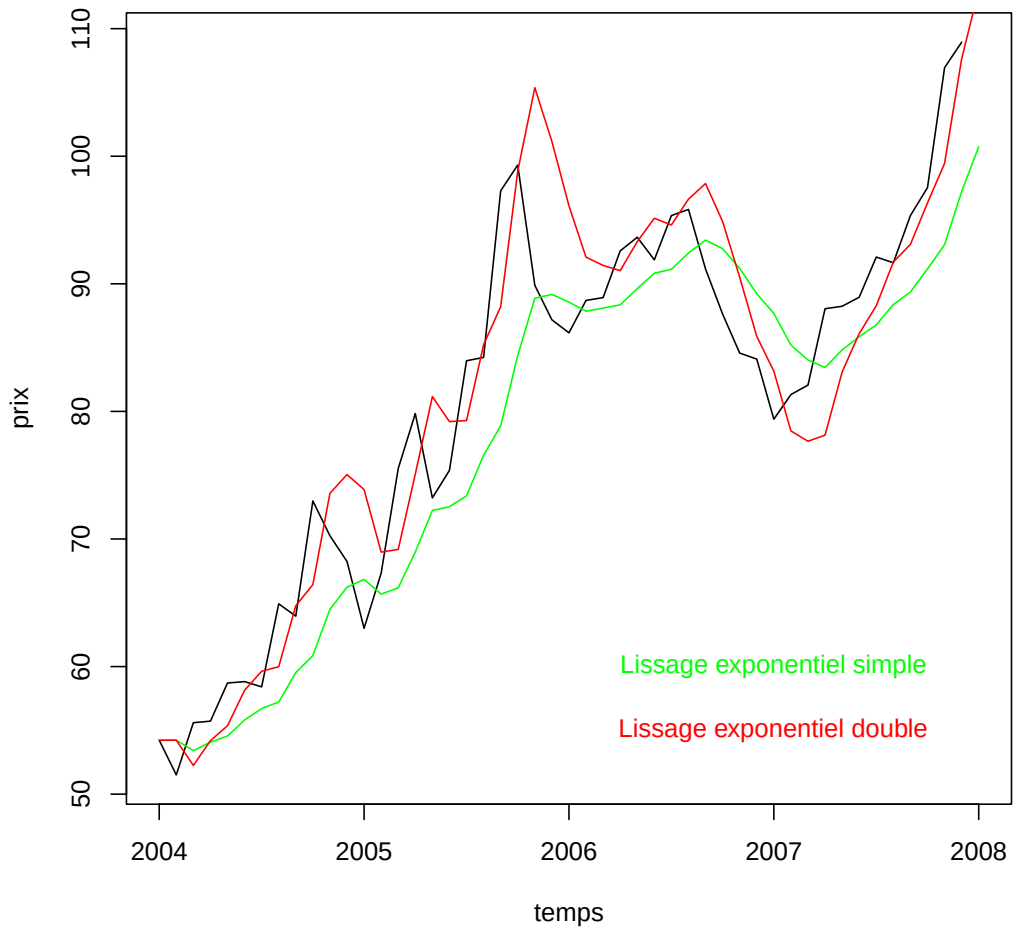


FIGURE 5.17 – Evolution du prix du mazout en CHF (achat entre 800 et1500 ℓ), lissage exponentiel double et lissage exponentiel simple

Chapitre 6

Calcul des probabilités et variables aléatoires

6.1 Probabilités

6.1.1 Événement

Une expérience est dite aléatoire si on ne peut pas prédire *a priori* son résultat. On note ω un résultat possible de cette expérience aléatoire. L'ensemble de tous les résultats possibles est noté Ω. Par exemple, si on jette deux pièces de monnaie, on peut obtenir les résultats

$$\Omega = \{(P,P),(F,P),(P,F),(F,F)\} ,$$

avec F pour "face" et P pour "pile". Un événement est une assertion logique sur une expérience aléatoire comme "avoir deux fois pile" ou "avoir au moins une fois pile". Formellement, un événement est un sous-ensemble de Ω.

— L'événement "avoir deux fois pile" est le sous ensemble $\{(P,P)\}$.
— L'événement "avoir au moins une fois pile" est le sous ensemble $\{(P,P),(F,P),(P,F)\}$.

L'ensemble Ω est appelé événement certain et l'ensemble vide $\emptyset$ est appelé événement impossible.

6.1.2 Opérations sur les événements

Sur les événements, on peut appliquer les opérations habituelles de la théorie des ensembles.

L'union

L'événement $A \cup B$ est réalisé dès que A ou B est réalisé. Dans un lancer de dé, si l'événement A est "obtenir un nombre pair" et l'événement B "obtenir un multiple de 3", l'événement $A \cup B$ est l'événement "obtenir un nombre pair OU un multiple de 3", c'est-à-dire $\{2,3,4,6\}$.

L'intersection

L'événement $A \cap B$ est réalisé dès que A et B sont réalisés conjointement dans la même expérience. Dans un lancer de dé, si l'événement A est "obtenir un nombre pair" et l'événement B "obtenir un multiple de 3", l'événement $A \cap B$ est l'événement "obtenir un nombre pair ET multiple de 3", c'est-à-dire $\{6\}$.

La différence

L'événement $A \backslash B$ est réalisé quand A est réalisé et que B ne l'est pas.

Le complémentaire

Le complémentaire de l'événement A est l'événement $\Omega \backslash A$. Le complémentaire est noté $\overline{A}$.

Exemple 6.1 L'expérience peut consister à jeter un dé, alors

$$\Omega = \{1,2,3,4,5,6\},$$

et un événement, noté A, est "obtenir un nombre pair". On a

$$A = \{2, 4, 6\} \text{ et } \overline{A} = \{1, 3, 5\}.$$

6.1.3 Relations entre les événements

Événements mutuellement exclusifs

Si $A \cap B = \emptyset$ on dit que A et B sont mutuellement exclusifs, ce qui signifie que A et B ne peuvent pas se produire ensemble.

Exemple 6.2 Si on jette un dé, l'événement "obtenir un nombre pair" et l'événement "obtenir un nombre impair" ne peuvent pas être obtenus en même temps. Ils sont mutuellement exclusifs. D'autre part, si l'on jette un dé, les événements A : "obtenir un nombre pair" n'est pas mutuellement exclusif avec l'événement B : "obtenir un nombre inférieur ou égal à 3". En effet, l'intersection de A et B est non-vide et consiste en l'événement "obtenir 2".

Inclusion

Si A est inclus dans B, on écrit $A \subset B$. On dit que A implique B.

Exemple 6.3 Si on jette un dé, on considère les événements A "obtenir 2" et B "obtenir un nombre pair".

$$A = \{2\} \text{ et } B = \{2, 4, 6\}.$$

On dit que A implique B.

6.1.4 Ensemble des parties d'un ensemble et système complet

On va associer à Ω l'ensemble $\mathcal{A}$ de toutes les parties (ou sous-ensembles) de Ω.

Exemple 6.4 Si on jette une pièce de monnaie alors $\Omega = \{P, F\}$, et

$$\mathcal{A} = \{\emptyset, \{F\}, \{P\}, \{F, P\}\}.$$

Définition 6.1. *Les événements $A_1, \ldots, A_n$ forment un système complet d'événements, si ils constituent une partition de Ω, c'est-à-dire si*

— *tous les couples A_i, A_j sont mutuellement exclusifs quand $i \neq j$,*
— $\bigcup_{i=1}^{n} A_i = \Omega$.

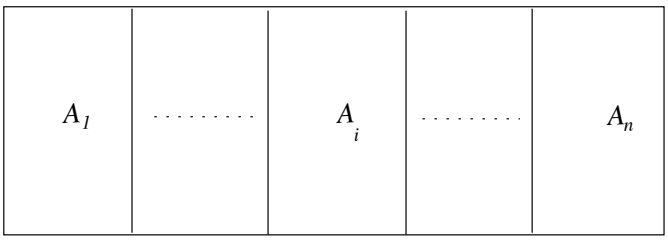


FIGURE 6.1 – Système complet d'événements

6.1.5 Axiomatique des Probabilités

Définition 6.2. *Une probabilité $P(.)$ est une application de $\mathcal{A}$ dans $[0, 1]$, telle que :*

— $\Pr(\Omega) = 1$,

— *Pour tout ensemble dénombrable d'événements $A_1, .., A_n$ mutuellement exclusifs (tels que $A_i \cap A_j = \emptyset$, pour tout $i \neq j$),*

$$\Pr(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n) = \Pr(A_1) + \Pr(A_2) + \Pr(A_3) + \cdots + \Pr(A_n).$$

À partir des axiomes, on peut déduire les propriétés suivantes :

Propriété 6.1. $\Pr(\emptyset) = 0$.

Démonstration. Comme $\emptyset$ est d'intersection vide avec $\emptyset$, on a que

$$\Pr(\emptyset \cup \emptyset) = \Pr(\emptyset) + \Pr(\emptyset).$$

Donc,

$$\Pr(\emptyset) = 2\Pr(\emptyset),$$

ce qui implique que $\Pr(\emptyset) = 0$. □

Propriété 6.2. $\Pr(\overline{A}) = 1 - \Pr(A)$.

Démonstration. On sait que

$$A \cup \overline{A} = \Omega \text{ et } A \cap \overline{A} = \emptyset.$$

Ainsi, on a que

$$\Pr(\Omega) = \Pr(A \cup \overline{A}) = \Pr(A) + \Pr(\overline{A}).$$

Mais, par la définition d'une probabilité, $\Pr(\Omega) = 1$. Donc,

$$\Pr(A) + \Pr(\overline{A}) = 1.$$

On en déduit que $\Pr(\overline{A}) = 1 - \Pr(A)$. □

Propriété 6.3. $\Pr(A) \leq \Pr(B)$ *si* $A \subset B$.

Démonstration. Comme $A \subset B$, on a

$$B = (B \cap \overline{A}) \cup A.$$

Mais on a que

$$(B \cap \overline{A}) \cap A = \emptyset.$$

Ainsi, on a

$$\Pr(B) = \Pr(B \cap \overline{A}) + \Pr(A).$$

Or une probabilité est à valeur dans [0,1]. Donc, $\Pr(B \cap \overline{A}) \geq 0$. On a alors

$$\Pr(B) \geq \Pr(A).$$

□

Propriété 6.4. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Démonstration. On a

$$A \cup B = A \cup (B \cap \overline{A}),$$

et

$$A \cap (B \cap \overline{A}) = \emptyset.$$

Donc,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B \cap \overline{A}).$$

Il reste à montrer que

$$\Pr(B \cap \overline{A}) = \Pr(B) - \Pr(A \cap B)$$

En effet,

$$B = (B \cap \overline{A}) \cup (B \cap A),$$

avec

$$(B \cap \overline{A}) \cap (B \cap A) = \emptyset.$$

Donc,

$$\Pr(B) = \Pr(B \cap \overline{A}) + \Pr(B \cap A),$$

ce qui donne

$$\Pr(B \cap \overline{A}) = \Pr(B) - \Pr(A \cap B).$$

□

Propriété 6.5. $\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i)$

Démonstration. Notons respectivement

$$B_1 = A_1, \quad B_2 = (A_2 \backslash A_1), \quad B_3 = (A_3 \backslash (A_1 \cup A_2)),$$

$$B_4 = (A_4 \backslash (A_1 \cup A_2 \cup A_3)), \quad \ldots, B_n = (A_n \backslash (A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_{n-1})).$$

Comme

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i,$$

et que $B_i \cap B_j = \emptyset$ pour tout $j \neq i$, alors

$$\Pr\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \Pr(B_i).$$

De plus, comme, pour tout i, $B_i \subset A_i$, on a que $\Pr(B_i) \leq \Pr(A_i)$, ce qui donne finalenent

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \Pr\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \Pr(B_i) \leq \sum_{i=1}^{n} \Pr(A_i).$$

□

Propriété 6.6. *Si $A_1, \ldots, A_n$ forment un système complet d'événements, alors*

$$\sum_{i=1}^{n} \Pr(B \cap A_i) = \Pr(B).$$

Démonstration. Si $A_1, \ldots, A_n$ forment un système complet d'événements, alors

$$B = \bigcup_{i=1}^{n} (B \cap A_i).$$

Mais on a, pour tout i, j tels que $i \neq j$

$$(B \cap A_i) \cap (B \cap A_j) = \emptyset.$$

Finalement, on a que

$$\Pr(B) = \Pr\left(\bigcup_{i=1}^{n} (B \cap A_i)\right) = \sum_{i=1}^{n} \Pr(B \cap A_i).$$

□

6.1.6 Probabilités conditionnelles et indépendance

Définition 6.3. *Soient deux événements A et B, si $\Pr(B) > 0$, alors*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Exemple 6.5 Si on jette un dé et que l'on considère les deux événements suivants :
- A l'évènement 'avoir un nombre pair' et
- B l'évènement 'avoir un nombre supérieur ou égal à 4'.

On a donc
- $\Pr(A) = \Pr(\{2, 4, 6\}) = \frac{1}{2}$,
- $\Pr(B) = \Pr(\{4, 5, 6\}) = \frac{3}{6} = \frac{1}{2}$,
- $\Pr(A \cap B) = \Pr(\{4, 6\}) = \frac{2}{6} = \frac{1}{3}$,
- $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/3}{1/2} = \frac{2}{3}$.

Définition 6.4. *Deux événements A et B sont dits indépendants si*

$$\Pr(A|B) = \Pr(A).$$

On peut montrer facilement que si A et B sont indépendants, alors

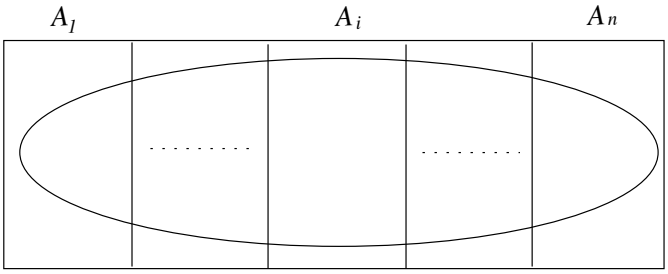
$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

6.1.7 Théorème des probabilités totales et théorème de Bayes

Théorème 6.1. *(des probabilités totales) Soient $A_1, \ldots, A_n$ un système complet d'événements, alors*

$$\Pr(B) = \sum_{i=1}^{n} \Pr(A_i)\Pr(B|A_i).$$

Table 6.1 – Illustration du théorème des probabilités totales



En effet,

$$\sum_{i=1}^{n} \Pr(A_i)\Pr(B|A_i) = \sum_{i=1}^{n} \Pr(B \cap A_i).$$

Comme les événements $A_i \cap B$ sont mutuellement exclusifs,

$$\sum_{i=1}^{n} \Pr(B \cap A_i) = \Pr \bigcup_{i=1}^{n} (B \cap A_i) = \Pr(B).$$

Théorème 6.2. *(de Bayes) Soit $A_1, \ldots, A_n$ un système complet d'événements, alors*

$$\Pr(A_i|B) = \frac{\Pr(A_i)\Pr(B|A_i)}{\sum_{j=1}^{n} \Pr(A_j)\Pr(B|A_j)}.$$

En effet, par le théorème des probabilités totales,

$$\frac{\Pr(A_i)\Pr(B|A_i)}{\sum_{j=1}^{n} \Pr(A_j)\Pr(B|A_j)} = \frac{\Pr(B \cap A_i)}{\Pr(B)} = \Pr(A_i|B).$$

Exemple 6.6 Supposons qu'une population d'adultes soit composée de 30% de fumeurs (A_1) et de 70% de non-fumeurs (A_2). Notons B l'événement "mourir d'un cancer du poumon". Supposons en outre que la probabilité de mourir d'un cancer du poumon est égale à $\Pr(B|A_1) = 20\%$ si l'on est fumeur et de $\Pr(B|A_2) = 1\%$ si l'on est non-fumeur. Le théorème de Bayes permet de calculer les probabilités a priori, c'est-à-dire la probabilité d'avoir été fumeur si on est mort d'un cancer du poumon. En effet, cette probabilité est notée $\Pr(A_1|B)$ et peut être calculée par

$$\Pr(A_1|B) = \frac{\Pr(A_1)\Pr(B|A_1)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.3 \times 0.2}{0.3 \times 0.2 + 0.7 \times 0.01} = \frac{0.06}{0.06 + 0.007} \approx 0.896.$$

La probabilité d'avoir été non-fumeur si on est mort d'un cancer du poumon vaut quant à elle :

$$\Pr(A_2|B) = \frac{\Pr(A_2)\Pr(B|A_2)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.7 \times 0.01}{0.3 \times 0.2 + 0.7 \times 0.01} = \frac{0.007}{0.06 + 0.007} \approx 0.104.$$

6.2 Analyse combinatoire

6.2.1 Introduction

L'analyse combinatoire est l'étude mathématique de la manière de ranger des objets. L'analyse combinatoire est un outil utilisé dans le calcul des probabilités.

6.2.2 Permutations (sans répétition)

Une permutation sans répétition est un classement ordonné de n objets distincts. Considérons par exemple l'ensemble $\{1,2,3\}$. Il existe 6 manières d'ordonner ces trois chiffres :

$$\{1,2,3\},\{1,3,2\},\{2,1,3\},\{2,3,1\},\{3,1,2\},\{3,2,1\}.$$

Si on dispose de n objets, chacun des n objets peut être placé à la première place. Il reste ensuite $n-1$ objets qui peuvent être placés à la deuxième place, puis $n-2$ objets pour la troisième place, et ainsi de suite. Le nombre de permutations possibles de n objets distincts vaut donc

$$n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = n!.$$

La notation $n!$ se lit factorielle de n (voir Tableau 6.2).

Table 6.2 – Factorielle des nombres de 1 à 10

n	0	1	2	3	4	5	6	7	8	9	10
$n!$	1	1	2	6	24	120	720	5040	40320	362880	3628800

6.2.3 Permutations avec répétition

On peut également se poser la question du nombre de manières de ranger des objets qui ne sont pas tous distincts. Supposons que nous ayons 2 boules rouges (notées R) et 3 boules blanches (notées B). Il existe 10 permutations possibles qui sont :

$$\{R,R,B,B,B\},\{R,B,R,B,B\},\{R,B,B,R,B\},\{R,B,B,B,R\},\{B,R,R,B,B\},$$

$$\{B,R,B,R,B\},\{B,R,B,B,R\},\{B,B,R,R,B\},\{B,B,R,B,R\},\{B,B,B,R,R\}.$$

Si l'on dispose de n objets appartenant à deux groupes de tailles n_1 et n_2, le nombre de permutations avec répétition est

$$\frac{n!}{n_1!n_2!}.$$

Par exemple si l'on a 3 boules blanches et 2 boules rouges, on obtient

$$\frac{n!}{n_1!n_2!} = \frac{5!}{2!3!} = \frac{120}{2\times 6} = 10.$$

Si l'on dispose de n objets appartenant à p groupes de tailles $n_1, n_2, \ldots, n_p$, le nombre de permutations avec répétition est

$$\frac{n!}{n_1!n_2! \times \cdots \times n_p!}.$$

6.2.4 Arrangements (sans répétition)

Soit n objets distincts. On appelle un arrangement une manière de sélectionner k objets parmi les n et de les ranger dans des boîtes numérotées de 1 à k.

Dans la première boîte, on peut mettre chacun des n objets. Dans la seconde boîte, on peut mettre chacun des $n-1$ objets restants, dans la troisième boîte, on peut mettre chacun des $n-2$ objets restants et ainsi de suite. Le nombre d'arrangements possibles est donc égal à :

$$A_n^k = n \times (n-1) \times (n-2) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

6.2.5 Combinaisons

Soit n objets distincts. On appelle une combinaison une manière de sélectionner k objets parmi les n sans tenir compte de leur ordre. Le nombre de combinaisons est le nombre de sous-ensembles de taille k dans un ensemble de taille n. Soit l'ensemble $\{1,2,3,4,5\}$. Il existe 10 sous-ensembles de taille 3 qui sont :

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,3,5\},\{1,4,5\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{3,4,5\}.$$

De manière générale, quel est le nombre de combinaisons de k objets parmi n ? Commençons par calculer le nombre de manières différentes de sélectionner k objets parmi n en tenant compte de l'ordre : c'est le nombre d'arrangements sans répétition A_n^k. Comme il existe $k!$ manières d'ordonner ces k éléments, si l'on ne veut pas tenir compte de l'ordre on divise A_n^k par $k!$. Le nombre de combinaisons de k objets parmi n vaut donc

$$\frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}.$$

Le nombre de combinaisons de k objets parmi n s'écrit parfois $\binom{n}{k}$ et parfois C_n^k :

$$\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}.$$

Par exemple, si on cherche à déterminer le nombre de combinaisons de 3 objets parmi 5, on a

$$\binom{5}{3} = C_5^3 = \frac{5!}{3!(5-3)!} = \frac{120}{6\times 2} = 10.$$

6.3 Variables aléatoires

La notion de variable aléatoire formalise l'association d'une valeur au résultat d'une expérience aléatoire.

Définition 6.5. *Une variable aléatoire X est une application de l'ensemble fondamental Ω dans $\mathbb{R}$.*

Exemple 6.7 On considère une expérience aléatoire consistant à lancer deux pièces de monnaie. L'ensemble des résultats possibles est

$$\Omega = \{(F,F),(F,P),(P,F),(P,P)\}.$$

Chacun des éléments de Ω a une probabilité $1/4$. Une variable aléatoire va associer une valeur à chacun des éléments de Ω. Considérons la variable aléatoire représentant le nombre de "Faces" obtenus :

$$X = \begin{cases} 0 & \text{avec une probabilité } 1/4 \\ 1 & \text{avec une probabilité } 1/2 \\ 2 & \text{avec une probabilité } 1/4. \end{cases}$$

C'est une variable aléatoire discrète dont la distribution de probabilité est présentée en Figure 6.2.

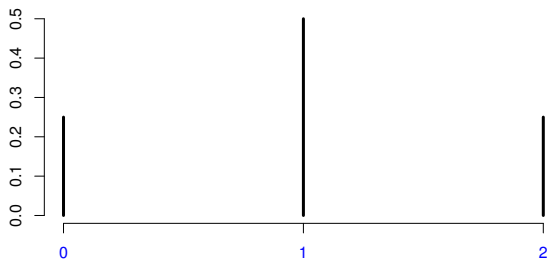


FIGURE 6.2 – Distribution de "faces" obtenus.

6.4 Variables aléatoires discrètes

6.4.1 Définition, espérance et variance

Une variable aléatoire discrète prend uniquement des valeurs entières (de $\mathbb{Z}$).

Une distribution de probabilité $p_X(x)$ est une fonction qui associe à chaque valeur entière une probabilité.

$$p_X(x) = \Pr(X = x), x \in \mathbb{Z}.$$

La fonction de répartition est définie par

$$F_X(x) = \Pr(X \leq x) = \sum_{z \leq x} p_X(z).$$

L'espérance mathématique d'une variable aléatoire discrète est définie de la manière suivante :

$$\mu = \mathrm{E}(X) = \sum_{x \in \mathbb{Z}} x p_X(x),$$

et sa variance

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}\left\{(X - \mathrm{E}(X))^2\right\} = \sum_{x \in \mathbb{Z}} p_X(x)(x - \mu)^2 = \sum_{x \in \mathbb{Z}} p_X(x)x^2 - \mu^2.$$

On peut aussi calculer les moments et tous les autres paramètres.

6.4.2 Variable indicatrice ou bernoullienne

La variable indicatrice X de paramètre $p \in [0, 1]$ a la distribution de probabilité suivante :

$$X = \begin{cases} 1 \text{ avec une probabilité } p \\ 0 \text{ avec une probabilité } 1 - p. \end{cases}$$

L'espérance vaut

$$\mu = \mathrm{E}(X) = 0 \times (1 - p) + 1 \times p = p,$$

et la variance vaut

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}(X - p)^2 = (1 - p)(0 - p)^2 + p(1 - p)^2 = p(1 - p).$$

Exemple 6.8 On tire au hasard une boule dans une urne contenant 18 boules rouges et 12 boules blanches. Si X vaut 1 si la boule est rouge et 0 sinon, alors X a une loi bernoullienne de paramètre $p = 18/(18 + 12) = 0.6$.

6.4.3 Variable binomiale

La variable aléatoire binomiale de paramètres n et p correspond à l'expérience suivante. On renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre p, où p est la probabilité de succès pour une expérience élémentaire. Ensuite, on note X le nombre de succès obtenus. Le nombre de succès est une variable aléatoire prenant des valeurs entières de 0 à n et ayant une distribution binomiale.

Une variable X suit une loi binomiale de paramètre $0 < p < 1$ et d'exposant n, si

$$\Pr(X = x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, \ldots, n - 1, n,$$

où $q = 1 - p$, et

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

De manière synthétique, si X a une distribution binomiale, on note :

$$X \sim \mathcal{B}(n, p).$$

Rappel Cette variable est appelée binomiale car sa distribution de probabilité est un terme du développement du binôme de Newton $(p + q)^n$.

$$\begin{aligned}
(p+q)^0 &= 1 \\
(p+q)^1 &= p + q = 1 \\
(p+q)^2 &= p^2 + 2pq + q^2 = 1 \\
(p+q)^3 &= p^3 + 3p^2q + 3pq^2 + q^3 = 1 \\
(p+q)^4 &= p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4 = 1 \\
&\vdots \\
(p+q)^n &= \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = 1.
\end{aligned}$$

La somme de ces probabilités vaut 1. En effet

$$\sum_{x=0}^{n} \Pr(X = x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1.$$

L'espérance se calcule de la manière suivante :

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{n} x\Pr(X = x) = \sum_{x=0}^{n} x\binom{n}{x} p^x q^{n-x} = \sum_{x=1}^{n} x\binom{n}{x} p^x q^{n-x} \text{ (on peut enlever le terme } x = 0) \\
&= \sum_{x=1}^{n} n\binom{n-1}{x-1} p^x q^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{(n-1)-(x-1)} \\
&= np \sum_{z=0}^{n-1} \binom{n-1}{z} p^z q^{(n-1)-z} \quad \text{(en posant } z = x-1) \\
&= np(p+q)^{n-1} = np.
\end{aligned}$$

Théorème 6.3. *La variance est donnée par*

$$\mathrm{var}(X) = npq.$$

Démonstration. Pour calculer cette variance, nous allons d'abord calculer $\mathrm{E}\{X(X-1)\}$. Ce résultat préliminaire nous permettra de déterminer ensuite la variance.

$$\begin{aligned}
\mathrm{E}\{X(X-1)\} &= \sum_{x=0}^{n} x(x-1)\Pr(X = x) \\
&= \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x q^{n-x} \\
&= \sum_{x=2}^{n} x(x-1)\binom{n}{x} p^x q^{n-x} \text{ (on peut enlever les termes } x = 0 \text{ et } x = 1) \\
&= \sum_{x=2}^{n} n(n-1)\binom{n-2}{x-2} p^x q^{n-x} \\
&= n(n-1)p^2 \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{(n-2)-(x-2)} \\
&= n(n-1)p^2 \sum_{z=0}^{n-2} \binom{n-2}{z} p^z q^{(n-2)-z} \quad \text{(en posant } z = x-2) \\
&= n(n-1)p^2(p+q)^{n-2} \\
&= n(n-1)p^2.
\end{aligned}$$

Comme

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$$

et que

$$\mathrm{E}\{X(X-1)\} = \mathrm{E}(X^2) - \mathrm{E}(X),$$

on obtient

$$\mathrm{var}(X) = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}^2(X) = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq.$$

□

Exemple 6.9 On tire au hasard avec remise et de manière indépendante 5 boules dans une urne contenant 18 boules rouges et 12 boules blanches. Si X est le nombre de boules rouges obtenues, alors X a une loi binomiale de paramètre $p = 18/(18+12) = 0.6$, et d'exposant $n = 5$. Donc,

$$\Pr(X = x) = \binom{5}{x} 0.6^x 0.4^{5-x}, x = 0, 1, \ldots, 4, 5,$$

ce qui donne

$$\begin{aligned}
\Pr(X=0) &= \frac{5!}{0!(5-0)!}0.6^0 \times 0.4^{5-0} = 1 \times 0.4^5 = 0.01024 \\
\Pr(X=1) &= \frac{5!}{1!(5-1)!}0.6^1 \times 0.4^{5-1} = 5 \times 0.6^1 \times 0.4^4 = 0.0768 \\
\Pr(X=2) &= \frac{5!}{2!(5-2)!}0.6^2 \times 0.4^{5-2} = 10 \times 0.6^2 \times 0.4^3 = 0.2304 \\
\Pr(X=3) &= \frac{5!}{3!(5-3)!}0.6^3 \times 0.4^{5-3} = 10 \times 0.6^3 \times 0.4^2 = 0.3456 \\
\Pr(X=4) &= \frac{5!}{4!(5-4)!}0.6^4 \times 0.4^{5-4} = 5 \times 0.6^4 \times 0.4^1 = 0.2592 \\
\Pr(X=5) &= \frac{5!}{5!(5-5)!}0.6^5 \times 0.4^{5-5} = 1 \times 0.6^5 = 0.07776.
\end{aligned}$$

La distribution de probabilité de la variable X est présentée dans la Figure 6.3.

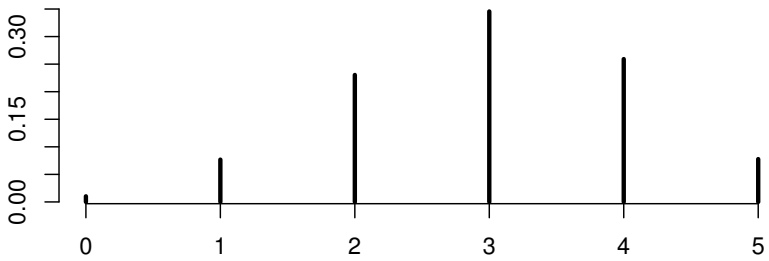


FIGURE 6.3 – Distribution d'une variable aléatoire binomiale avec $n = 5$ et $p = 0.6$.

Exemple 6.10 Supposons que, dans une population d'électeurs, 60% des électeurs s'apprêtent à voter pour le candidat A et 40% pour le candidat B et que l'on sélectionne un échantillon aléatoire de 10 électeurs avec remise dans cette population. Soit X le nombre de personnes s'apprêtant à voter pour le candidat A dans l'échantillon. La variable X a une distribution binomiale de paramètres $n = 10$ et $p = 0.6$ et donc

$$\Pr(X = x) = \binom{10}{x} 0.6^x (0.4)^{10-x}, x = 0, 1, \ldots, n-1, n.$$

6.4.4 Variable de Poisson

La variable X suit une loi de Poisson, ou loi des événements rares, de paramètre $\lambda \in \mathbb{R}^+$ si

$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, 3, \ldots.$$

On note alors $X \sim \mathcal{P}(\lambda)$. La somme des probabilités est bien égale à 1, en effet

$$\sum_{x=0}^{\infty} \Pr(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Cette loi exprime la probabilité de l'occurence d'un nombre d'événements dans un laps de temps fixe si ces événements se produisent avec un taux moyen connu (λ) et indépendamment du temps d'occurrence du dernier événement.

L'espérance et la variance d'une loi de Poisson sont égales au paramètre λ. En effet,

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{\infty} x\Pr(X=x) = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda\sum_{z=0}^{\infty}\frac{\lambda^z}{z!} \text{ (en posant } z = x-1) \\
&= e^{-\lambda}\lambda e^{\lambda} = \lambda.
\end{aligned}$$

En outre, il est possible de montrer que

$$\mathrm{var}(X) = \lambda.$$

La distribution de probabilité d'une variable de Poisson $\mathcal{P}(\lambda = 1)$ est présentée dans la Figure 6.4.

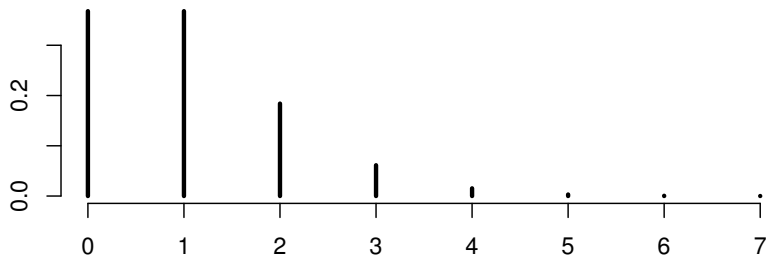


FIGURE 6.4 – Distribution d'une variable de Poisson avec $\lambda = 1$.

En langage R

```
#
# distributions de probabilités discrètes
#
# nombre de faces obtenues en lançant deux pièces
plot(0:2,dbinom(0:2, 2,0.5),type = "h", lwd=3,
ylim=c(0,0.5),xlab="",ylab="",xaxt = "n",frame = FALSE)
axis(1, 0:2, 0:2, col.axis = "blue")
# binomiale B(5,0.6)
plot(dbinom(0:5, 5,0.6),type = "h",
lwd=3,xlab="",ylab="",main="",frame=FALSE)
# Poisson P(1)
plot(dpois(0:7, 1),type = "h",
lwd=3,xlab="",ylab="",main="",frame=FALSE)
```

6.5 Variable aléatoire continue

6.5.1 Définition, espérance et variance

Une variable aléatoire continue prend des valeurs dans $\mathbb{R}$ ou dans un intervalle de $\mathbb{R}$.

La probabilité qu'une variable aléatoire continue soit inférieure à une valeur particulière est donnée par sa fonction de répartition.

$$\Pr(X \leq x) = F(x).$$

La fonction de répartition d'une variable aléatoire continue est toujours :

— dérivable,
— positive : $F(x) \geq 0$, pour tout x,
— croissante,
— $\lim_{x\to\infty} F(x) = 1$,
— $\lim_{x\to-\infty} F(x) = 0$.

On a

$$\Pr(a \leq X \leq b) = F(b) - F(a).$$

La fonction de densité d'une variable aléatoire continue est la dérivée de la fonction de répartition en un point

$$f(x) = \frac{dF(x)}{dx}.$$

Une fonction de densité est toujours :
- positive : $f(x) \geq 0$, pour tout x,
- d'aire égale à un : $\int_{-\infty}^{\infty} f(x)dx = 1$.

On a évidemment la relation :

$$F(b) = \int_{-\infty}^{b} f(x)dx.$$

La probabilité que la variable aléatoire soit inférieure à une valeur quelconque vaut :

$$\Pr(X \leq a) = \int_{-\infty}^{a} f(x)dx = F(a).$$

Dans la Figure 6.5, la probabilité $\Pr(X \leq a)$ est l'aire sous la densité de $-\infty$ à a.

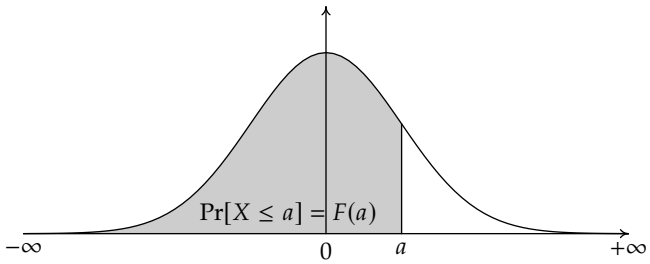


FIGURE 6.5 – Probabilité que la variable aléatoire soit inférieure à a

La probabilité que la variable aléatoire prenne une valeur comprise entre a et b vaut

$$\Pr(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a).$$

Si la variable aléatoire est continue, la probabilité qu'elle prenne exactement une valeur quelconque est nulle :

$$\Pr(X = a) = 0.$$

L'espérance d'une variable aléatoire continue est définie par :

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)dx,$$

et la variance

$$\sigma^2 = \mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

6.5.2 Variable uniforme

Une variable aléatoire X est dite uniforme dans un intervalle [a,b] (avec $a < b$), si sa répartition est :

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

Sa densité est alors

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

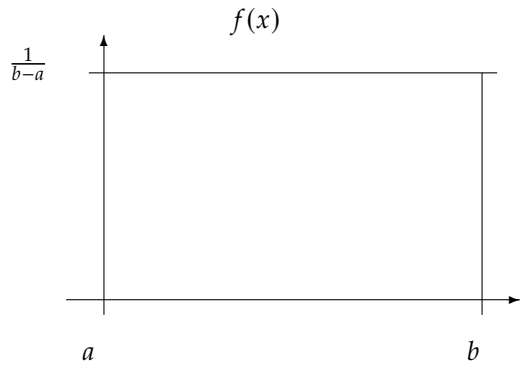


FIGURE 6.6 – Fonction de densité d'une variable uniforme

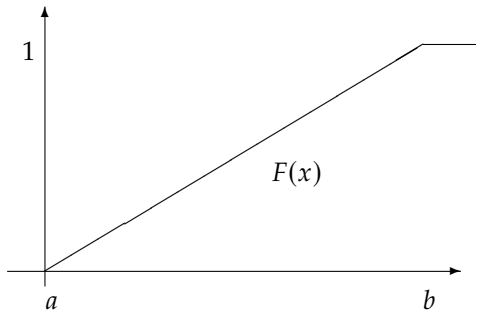


FIGURE 6.7 – Fonction de répartition d'une variable uniforme

De manière synthétique, on écrit :

$$X \sim \mathcal{U}(a, b).$$

Les logiciels peuvent générer des variables aléatoires uniformes dans [0,1] (En R : runif(1)). Les Figures 6.6 et 6.7 représentent respectivement les fonctions de densité et de répartition d'une variable uniforme.

On peut calculer l'espérance et la variance :

Résultat 6.1.

$$\mu = \mathrm{E}(X) = \frac{b+a}{2}$$

Démonstration.

$$\begin{aligned}
\mu &= \mathrm{E}(X) = \int_a^b x f(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b \\
&= \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) = \frac{1}{b-a} \frac{1}{2} (b+a)(b-a) = \frac{a+b}{2}.
\end{aligned}$$

□

Résultat 6.2.

$$\sigma^2 = \mathrm{var}(X) = \frac{(b-a)^2}{12}.$$

Démonstration. De manière générale, une variance peut toujours s'écrire comme un moment à l'origine d'ordre 2 moins le carré de la moyenne. En effet,

$$\begin{aligned}
\sigma^2 &= \mathrm{var}(X) = \int_a^b (x-\mu)^2 f(x) dx = \int_a^b (x^2 + \mu^2 - 2x\mu) f(x) dx \\
&= \int_a^b x^2 f(x) dx + \int_a^b \mu^2 f(x) dx - 2\mu \int_a^b x f(x) dx = \int_a^b x^2 f(x) dx + \mu^2 - 2\mu^2 = \int_a^b x^2 f(x) dx - \mu^2.
\end{aligned}$$

On calcule ensuite un moment à l'origine d'ordre 2 :

$$\begin{aligned}
\int_a^b x^2 f(x) dx &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{1}{b-a} \left(\frac{x^3}{3} \right)_a^b \\
&= \frac{1}{b-a} \left(\frac{b^3}{3} - \frac{a^3}{3} \right) = \frac{1}{b-a} \frac{1}{3} (b^2 + ab + a^2)(b-a) = \frac{b^2 + ab + a^2}{3}.
\end{aligned}$$

On obtient enfin la variance par différence :

$$\begin{aligned}\sigma^2 &= \int_a^b x^2 f(x)dx - \mu^2 = \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4b^2+4ab+4a^2}{12} - \frac{3a^2+6ab+3b^2}{12} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.\end{aligned}$$

□

6.5.3 Variable normale

Une variable aléatoire X est dite normale si sa densité vaut

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \tag{6.1}$$

où $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$ sont les paramètres de la distribution. Le paramètre μ est appelé la moyenne et le paramètre σ l'écart-type de la distribution.

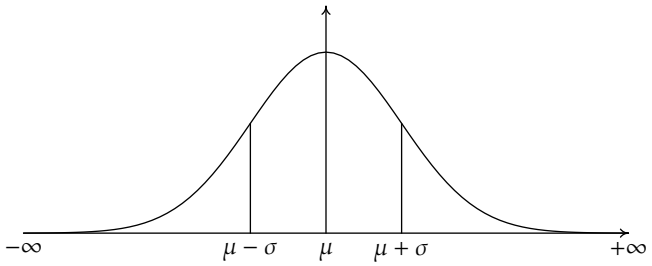


Figure 6.8 – Fonction de densité d'une variable normale

De manière synthétique, pour noter que X suit une loi normale (ou gaussienne, d'après Carl Friedrich Gauss) de moyenne μ et de variance σ^2 on écrit :

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

La loi normale est une des principales distributions de probabilité. Elle a de nombreuses applications en statistique. Sa fonction de densité dessine une courbe dite courbe de Gauss. On peut montrer (sans démonstration) que

$$\mathrm{E}(X) = \mu,$$

et

$$\mathrm{var}(X) = \sigma^2.$$

La fonction de répartition vaut

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2 du.$$

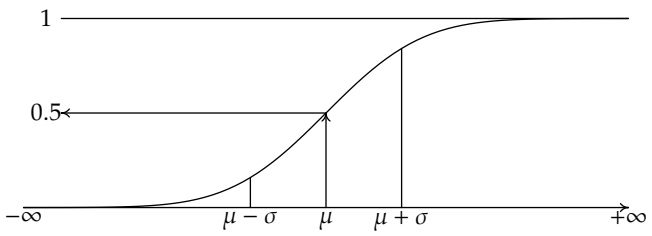


Figure 6.9 – Fonction de répartition d'une variable normale

6.5.4 Variable normale centrée réduite

La variable aléatoire normale centrée réduite est une variable normale, d'espérance nulle, $\mu = 0$, et de variance $\sigma^2 = 1$. Sa fonction de densité vaut

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}.$$

et sa répartition vaut

$$\Phi(x) = F_{0,1}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp -\left(\frac{u^2}{2}\right) du.$$

Du fait de la symétrie de la densité, on a la relation

$$\Phi(-x) = 1 - \Phi(x),$$

qui se comprend facilement en examinant la Figure 6.10.

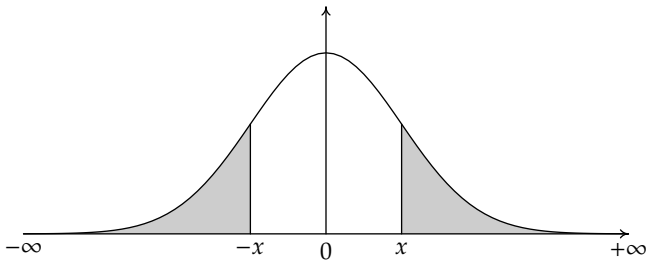


FIGURE 6.10 – Densité d'une normale centrée réduite, symétrie

De plus, le calcul de la répartition d'une variable normale de moyenne μ et de variance σ^2 peut toujours être ramené à une normale centrée réduite.

Résultat 6.3.

$$F_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Démonstration. On a

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp -\left\{\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\} du.$$

En posant

$$z = \frac{u-\mu}{\sigma},$$

on obtient $u = z\sigma + \mu$, et donc $du = \sigma dz$. Donc,

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sigma\sqrt{2\pi}} \exp -\left(\frac{z^2}{2}\right) \sigma dz = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

□

Les tables de la variable normale ne sont données que pour la normale centrée réduite. Les tables ne donnent $\Phi(x)$ que pour les valeurs positives de x, car les valeurs négatives peuvent être trouvées par la relation de symétrie.

6.5.5 Distribution exponentielle

Soit une variable aléatoire X qui définit la durée de vie d'un phénomène ou d'un objet. Si la durée de vie est *sans vieillissement*, c'est-à-dire la durée de vie au delà d'un instant T est indépendante de l'instant T, alors sa fonction de densité est donnée par :

$$f(x) = \begin{cases} \lambda \exp -(\lambda x), & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

On dit que X suit une loi exponentielle de paramètre λ positif. De manière synthétique, on écrit :

$$X \sim \varepsilon(\lambda).$$

Quand $x > 0$, sa fonction de répartition vaut :

$$F(x) = \int_0^x f(u)du = \int_0^x \lambda e^{-\lambda u} du = \left[-e^{-\lambda u}\right]_0^x = 1 - e^{-\lambda x}.$$

On peut alors calculer la moyenne :

Résultat 6.4. $\mathrm{E}(X) = \dfrac{1}{\lambda}$

Démonstration.

$$\mathrm{E}(X) = \int_0^\infty x f(x)dx = \int_0^\infty x\lambda e^{-\lambda x} dx = \left[-\frac{1 + x\lambda}{\lambda} e^{-\lambda x}\right]_0^\infty = \left(0 + \frac{1}{\lambda}\right) = \frac{1}{\lambda}.$$

□

Il est également possible de montrer que la variance vaut :

$$\mathrm{var}(X) = \frac{1}{\lambda^2}.$$

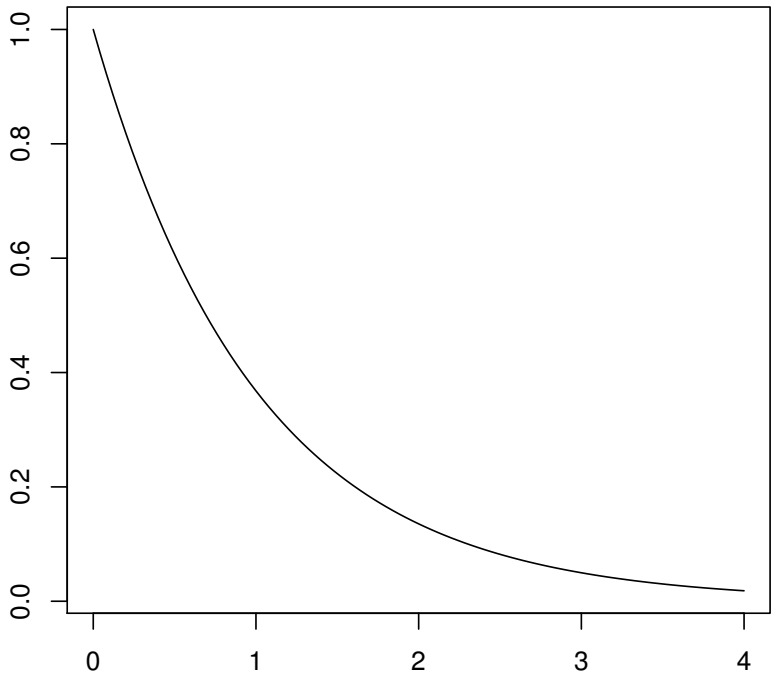


FIGURE 6.11 – Fonction de densité d'une variable exponentielle avec $\lambda = 1$.

6.6 Distribution bivariée

Deux variables aléatoires peuvent avoir une distribution jointe.

6.6.1 Cas continu

Soit deux variables aléatoires X et Y continues, leur distribution de densité $f(x, y)$ est une fonction continue, positive et telle que

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1.$$

La fonction de répartition jointe est définie par

$$F(x,y) = \Pr(X \leq x \text{ et } Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dvdu.$$

On appelle densités marginales les fonctions

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \text{ et } f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Avec les distributions marginales, on peut définir les moyennes marginales et les variances marginales :

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x)dx, \text{ et } \mu_Y = \int_{-\infty}^{\infty} y f_Y(y)dy,$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx, \text{ et } \sigma_Y^2 = \int_{-\infty}^{\infty} (y-\mu_Y)^2 f_Y(y)dy.$$

On appelle densités conditionnelles, les fonctions

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \text{ et } f(y|x) = \frac{f(x,y)}{f_X(x)}.$$

Avec les distributions conditionnelles, on peut définir les moyennes conditionnelles et les variances conditionnelles :

$$\mu_X(y) = \mathrm{E}(X|Y=y) = \int_{-\infty}^{\infty} x f(x|y)dx, \text{ et } \mu_Y(x) = \mathrm{E}(Y|X=x) = \int_{-\infty}^{\infty} y f(y|x)dy,$$

$$\sigma_X^2(y) = \mathrm{var}(X|Y=y) = \int_{-\infty}^{\infty} \left\{x-\mu_X(y)\right\}^2 f(x|y)dx,$$

et

$$\sigma_Y^2(x) = \mathrm{var}(Y|X=x) = \int_{-\infty}^{\infty} \left\{y-\mu_Y(x)\right\}^2 f(y|x)dy.$$

Enfin, la covariance entre X et Y est définie par

$$\sigma_{xy} = \mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)dxdy.$$

6.6.2 Cas discret

Soit deux variables aléatoires X et Y discrètes, leur distribution de probabilité jointe $p(x,y)$ est telle que

$$\sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}} p(x,y) = 1.$$

La fonction de répartition jointe est définie par

$$F(x,y) = \Pr(X \leq x \text{ et } Y \leq y) = \sum_{u\leq x}\sum_{v\leq y} p(u,v).$$

On appelle distributions de probabilté marginales les fonctions

$$p_X(x) = \sum_{y\in\mathbb{Z}} p(x,y), \text{ et } p_Y(y) = \sum_{x\in\mathbb{Z}} p(x,y).$$

Avec les distributions marginales, on peut définir les moyennes marginales et les variances marginales :

$$\mu_X = \sum_{x\in\mathbb{Z}} x p_X(x), \text{ et } \mu_Y = \sum_{y\in\mathbb{Z}} y p_Y(y),$$

$$\sigma_X^2 = \sum_{x\in\mathbb{Z}} (x-\mu_X)^2 p_X(x), \text{ et } \sigma_Y^2 = \sum_{y\in\mathbb{Z}} (y-\mu_Y)^2 p_Y(y).$$

On appelle distributions conditionnelles, les fonctions

$$p(x|y) = \frac{p(x,y)}{p_Y(y)} \text{ et } p(y|x) = \frac{p(x,y)}{p_X(x)}.$$

Avec les distributions conditionnelles, on peut définir les moyennes conditionnelles et les variances conditionnelles :

$$\mu_X(y) = \sum_{x\in\mathbb{Z}} xp(x|y), \text{ et } \mu_Y(x) = \sum_{y\in\mathbb{Z}} yp(y|x),$$

$$\sigma_X^2(y) = \sum_{x\in\mathbb{Z}} \left\{x - \mu_X(y)\right\}^2 p(x|y), \text{ et } \sigma_Y^2(x) = \sum_{x\in\mathbb{Z}} \left\{y - \mu_Y(x)\right\}^2 p(y|x).$$

Enfin, la covariance entre X et Y est définie par

$$\sigma_{xy} = \text{cov}(X,Y) = \sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}}(x-\mu_X)(y-\mu_Y)p(x,y).$$

6.6.3 Remarques

Dans les deux cas discrets et continus, on peut toujours écrire

$$\begin{aligned}\text{var}(X) &= \text{E}\{X - \text{E}(X)\}^2 = \text{E}\left\{X^2 - 2X\text{E}(X) + \text{E}^2(X)\right\}\\ &= \text{E}(X^2) - 2\text{E}(X)\text{E}(X) + \text{E}^2(X) = \text{E}(X^2) - \text{E}^2(X).\end{aligned}$$

De même,

$$\text{var}(X|Y=y) = \text{E}\left[\{X - \text{E}(X|Y=y)\}^2|Y=y\right] = \text{E}(X^2|Y=y) - \text{E}^2(X|Y=y).$$

On a également

$$\begin{aligned}\text{cov}(X,Y) &= \text{E}\{X-\text{E}(X)\}\{Y-\text{E}(Y)\} = \text{E}\{XY - Y\text{E}(X) - X\text{E}(Y) + \text{E}(X)\text{E}(Y)\}\\ &= \text{E}(XY) - \text{E}(X)\text{E}(Y) - \text{E}(X)\text{E}(Y) + \text{E}(X)\text{E}(Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y).\end{aligned}$$

L'opérateur espérance permet donc de définir la variance et la covariance.

6.6.4 Indépendance de deux variables aléatoires

Deux variables aléatoires X et Y sont dites indépendantes, si

$$\Pr(X \le x \text{ et } Y \le y) = \Pr(X \le x)\Pr(Y \le y), \text{ pour tout } x, y \in \mathbb{R}.$$

— Si X et Y sont discrètes, cela implique que

$$\Pr(X = x \text{ et } Y = y) = \Pr(X = x)\Pr(Y = y), \text{ pour tout } x, y \in \mathbb{Z}.$$

— Si X et Y sont continues, en notant $f_X(.)$ et $f_Y(.)$ les fonctions de densité respectives de X et Y et en notant $f_{XY}(x,y)$ la densité jointe des deux variables, alors X et Y sont indépendants si

$$f_{XY}(x,y) = f_X(x)f_Y(y), x, y \in \mathbb{R}.$$

6.7 Propriétés des espérances et des variances

De manière générale, pour des variables aléatoires X et Y et avec a et b constants, on a les résultats suivants qui sont démontrées pour le cas continu. Ces résultats sont également valables pour le cas discret pour lequel les démonstrations sont similaires.

Résultat 6.5.

$$\text{E}(a + bX) = a + b\text{E}(X)$$

Démonstration.

$$\text{E}(a+bX) = \int_{\mathbb{R}}(a + b\ x)f(x)dx = a\int_{\mathbb{R}} f(x)dx + b\int_{\mathbb{R}} xf(x)dx = a + b\text{E}(X).$$

□

Résultat 6.6.

$$\mathrm{E}(aY + bX) = a\mathrm{E}(Y) + b\mathrm{E}(X).$$

Démonstration.

$$\begin{aligned}\mathrm{E}(aY + bX) &= \int_{\mathbb{R}}\int_{\mathbb{R}}(ay + b\ x)f(x,y)dxdy = a\int_{\mathbb{R}}\int_{\mathbb{R}} yf(x,y)dxdy + b\int_{\mathbb{R}}\int_{\mathbb{R}} xf(x,y)dxdy \\ &= a\int_{\mathbb{R}} y\int_{\mathbb{R}} f(x,y)dxdy + b\int_{\mathbb{R}} x\int_{\mathbb{R}} f(x,y)dydx = a\int_{\mathbb{R}} yf_Y(y)dy + b\int_{\mathbb{R}} xf_X(x)dx = a\mathrm{E}(Y) + b\mathrm{E}(X)\end{aligned}$$

□

Quand a et b valent 1, on obtient que l'espérance de la somme de deux variables aléatoires est égale à la somme de leur espérances :

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

Résultat 6.7.

$$\mathrm{var}(a + bX) = b^2\mathrm{var}(X).$$

Démonstration.

$$\begin{aligned}\mathrm{var}(a + bX) &= \int_{\mathbb{R}}\{a + b\ x - \mathrm{E}(a + bX)\}^2 f(x)dx = \int_{\mathbb{R}}\{a + b\ x - (a + b\mathrm{E}(X))\}^2 f(x)dx \\ &= \int_{\mathbb{R}}\{b\ x - b\mathrm{E}(X)\}^2 f(x)dx = b^2\int_{\mathbb{R}}\{x - \mathrm{E}(X)\}^2 f(x)dx = b^2\mathrm{var}(X).\end{aligned}$$

□

La variance n'est donc pas sensible à un changement d'origine, mais est affectée par le carré d'un changement d'unité.

Résultat 6.8.

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y).$$

Démonstration.

$$\begin{aligned}\mathrm{var}(X + Y) &= \int_{\mathbb{R}}\int_{\mathbb{R}}\{x + y - \mathrm{E}(X + Y)\}^2 f(x,y)dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}\{x - \mathrm{E}(X) + y - \mathrm{E}(Y)\}^2 f(x,y)dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}\left\{[x - \mathrm{E}(X)]^2 + [y - \mathrm{E}(Y)]^2 + 2[x - \mathrm{E}(X)][y - \mathrm{E}(Y)]\right\} f(x,y)dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}[x - \mathrm{E}(X)]^2 f(x,y)dxdy + \int_{\mathbb{R}}\int_{\mathbb{R}}[y - \mathrm{E}(Y)]^2]f(x,y)dxdy \\ &\quad +2\int_{\mathbb{R}}\int_{\mathbb{R}}[x - \mathrm{E}(X)][y - \mathrm{E}(Y)]f(x,y)dxdy \\ &= \int_{\mathbb{R}}[x - \mathrm{E}(X)]^2\int_{\mathbb{R}} f(x,y)dydx + \int_{\mathbb{R}}[y - \mathrm{E}(Y)]^2]\int_{\mathbb{R}} f(x,y)dxdy + 2\mathrm{cov}(X,Y) \\ &= \int_{\mathbb{R}}[x - \mathrm{E}(X)]^2 f_X(x)dx + \int_{\mathbb{R}}[y - \mathrm{E}(Y)]^2]f_Y(y)dy + 2\mathrm{cov}(X,Y) \\ &= \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y)\end{aligned}$$

□

Résultat 6.9. *De plus, si X et Y sont indépendantes, on a $f(x,y) = f_X(x)_Y f(y)$ pour tout x, y*

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y).$$

Démonstration.

$$\mathrm{E}(XY) = \int_{\mathbb{R}}\int_{\mathbb{R}} xyf_X(x)f_Y(y)dxdy = \int_{\mathbb{R}} xf_X(x)dx\int_{\mathbb{R}} yf_Y(y)dy = \mathrm{E}(X)\mathrm{E}(Y).$$

□

Comme, de manière générale $\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$, on déduit directement du Résultat 6.9 que, si X et Y sont indépendantes, on a $\mathrm{cov}(X,Y) = 0$, et donc

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

Attention, la réciproque n'est pas vraie. Une covariance nulle n'implique pas que les deux variables sont indépendantes.

Enfin, il est possible de calculer l'espérance et la variance d'une somme de variables aléatoires indépendantes et identiquement distribuées.

Théorème 6.4. *Soit $X_1, \ldots, X_n$ une suite de variables aléatoires, indépendantes et identiquement distribuées et dont la moyenne μ et la variance σ^2 existent et sont finies, alors si*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

on a

$$\mathrm{E}(\overline{X}) = \mu, \text{ et } \mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Démonstration.

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

et

$$\mathrm{var}\left(\overline{X}\right) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

□

6.8 Autres variables aléatoires

6.8.1 Variable khi-carrée

Soit une suite de variables aléatoires indépendantes, normales centrées réduites, $X_1, \ldots, X_p$, (c'est-à-dire de moyenne nulle et de variance égale à 1), alors la variable aléatoire

$$\chi_p^2 = \sum_{i=1}^{p} X_i^2,$$

est appelée variable aléatoire khi-carré à p degrés de liberté.

Il est possible de montrer que

$$\mathrm{E}(\chi_p^2) = p,$$

et que

$$\mathrm{var}(\chi_p^2) = 2p.$$

6.8.2 Variable de Student

Soit une variable aléatoire X normale centrée réduite et une variable aléatoire khi-carré χ_p^2 à p degrés de liberté, indépendante de X, alors la variable aléatoire

$$t_p = \frac{X}{\sqrt{\chi_p^2/p}}$$

est appelée variable aléatoire de Student à p degrés de liberté.

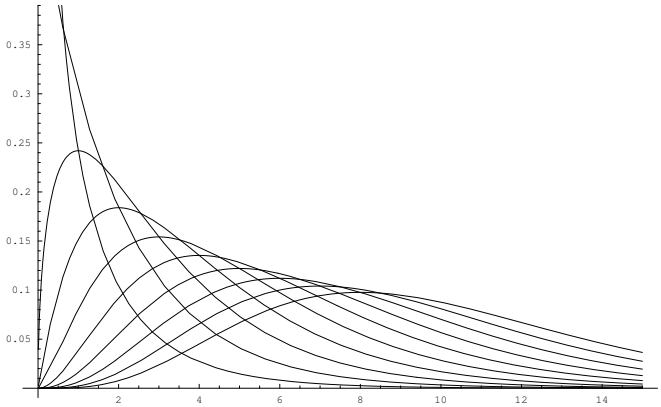


FIGURE 6.12 – Densité d'une variable de chi-carré avec $p = 1, 2, \ldots, 10$

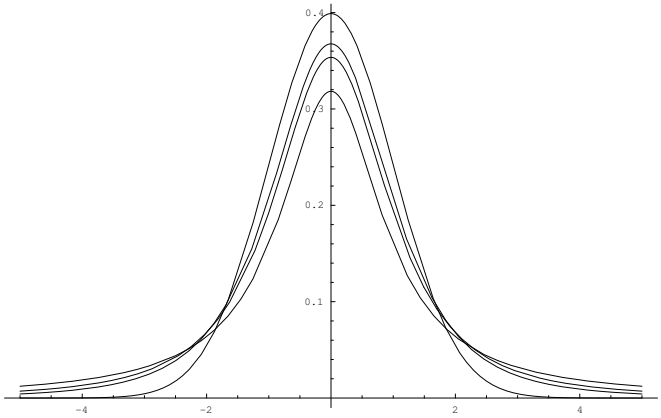


FIGURE 6.13 – Densités de variables de Student avec $p = 1, 2$ et 3 et d'une variable normale

6.8.3 Variable de Fisher

Soient deux variables aléatoires khi-carrés indépendantes χ^2_p, χ^2_q, respectivement à p et q degrés de liberté, alors la variable aléatoire

$$F_{p,q} = \frac{\chi^2_p/p}{\chi^2_q/q}$$

est appelée variable aléatoire de Fisher à p et q degrés de liberté.
Remarque 6.1 Il est facile de montrer que le carré d'une variable de Student à q degrés de liberté est une variable de Fisher à 1 et q degrés de liberté.

6.8.4 Distribution normale bivariée

Les variables X et Y suivent une loi normale bivariée si leur densité jointe est donnée par

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_x\sigma_y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}. \tag{6.2}$$

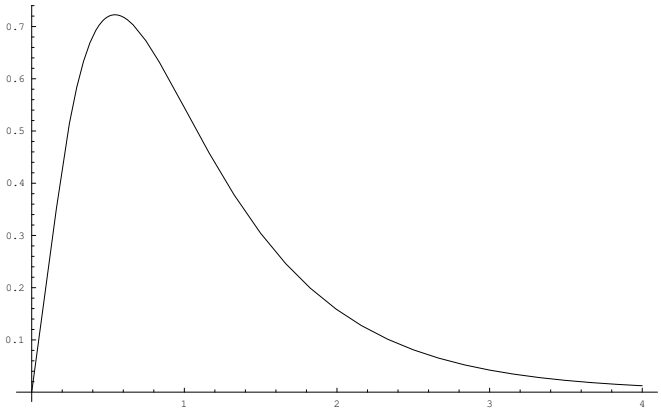


FIGURE 6.14 – Densité d'une variable de Fisher

La fonction de densité dépend de 5 paramètres
- les deux moyennes marginales $\mu_X \in \mathbb{R}$ et $\mu_Y \in \mathbb{R}$,
- les deux variances marginales $\sigma_X^2 > 0$ et $\sigma_Y^2 > 0$,
- le coefficient de corrélation $-1 < \rho < 1$.

Un exemple de normale bivariée est présentée dans la Figure 6.15.

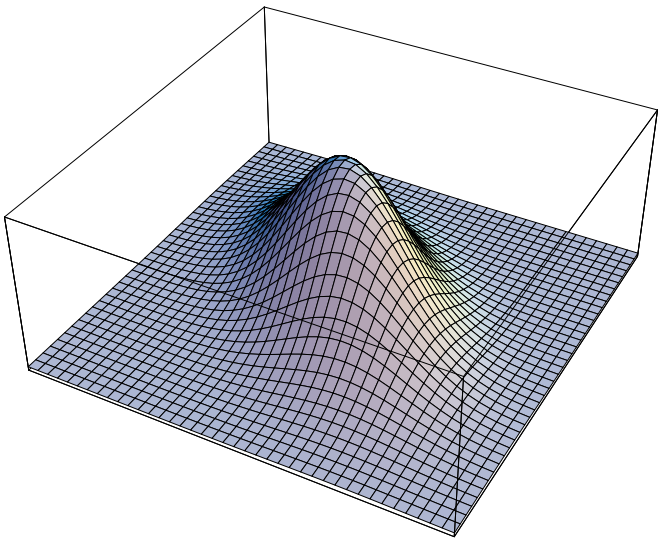


FIGURE 6.15 – Densité d'une normale bivariée

La Figure 6.16 montre le nuage de points de 1000 réalisations d'une normale bivariée avec les paramètres suivants : $\mu_X = 8$, $\mu_Y = 20$, $\sigma_X^2 = 9$, $\sigma_Y^2 = 25$, $\rho = 0.6$.

En langage R

```
a=8; b=3; c=12; d=4
X=a+ b*rnorm(2000)
Y=c+X+d*rnorm(2000)
plot(X,Y,type="p")
```

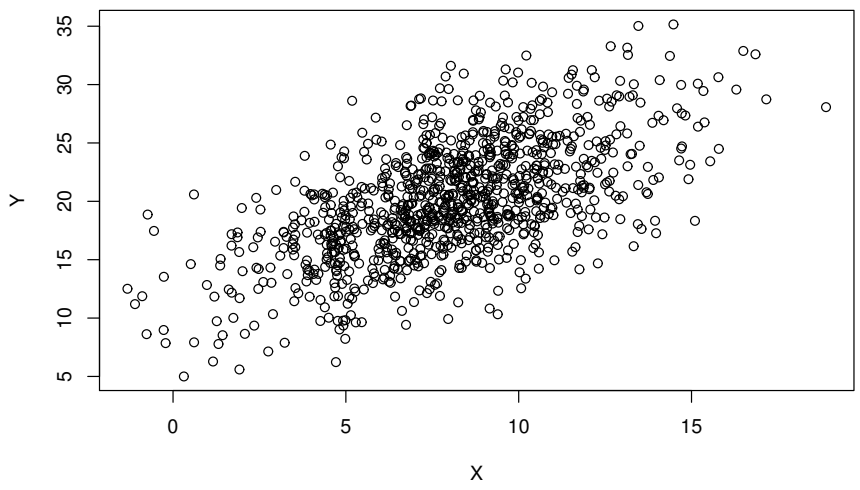


FIGURE 6.16 – Nuage de points de réalisations d'une normale bivariée

Théorème 6.5. *Les deux distributions marginales d'une distribution normale bivariée ont une distribution normale donnée par :*

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\sigma_X\sqrt{2\pi}} \exp -\frac{(x-\mu_X)^2}{2\sigma_X^2}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp -\frac{(y-\mu_Y)^2}{2\sigma_Y^2}$$

Démonstration. (pour $f_X(x)$)
On peut vérifier que la densité jointe peut également s'écrire :

$$f(x,y) = \left(\frac{1}{\sigma_X\sqrt{2\pi}} \exp -\frac{(x-\mu_X)^2}{2\sigma_X^2}\right) \frac{1}{\sigma_Y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_Y(x)}{\sigma_Y(x)}\right)^2\right\},$$

où

$$\mu_Y(x) = \mu_Y + \frac{\sigma_Y\rho}{\sigma_X}(x-\mu_X) \text{ et } \sigma_Y^2(x) = \sigma_Y^2(1-\rho^2).$$

On a

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \left(\frac{1}{\sigma_X\sqrt{2\pi}} \exp -\frac{(x-\mu_X)^2}{2\sigma_X^2}\right) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sigma_Y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_Y(x)}{\sigma_Y(x)}\right)^2\right\} dy}_{=1}. \end{aligned}$$

□

Le Théorème 6.5 montre que les deux distributions marginales sont normales, que μ_X et μ_Y sont les moyennes marginales et que σ_X^2 et σ_X^2 sont les deux variance marginales de la distribution jointes. On peut également montrer à partir du Théorème 6.5 que le volume sous la courbe vaut bien 1. En effet

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty} f_Y(y)dy = 1.$$

Attention, la réciproque du Théorème 6.5 n'est pas nécessairement vraie. Une distribution bivariée dont les deux distributions marginales sont normales, n'est pas nécessairement normale.

Théorème 6.6. *Toutes les distributions conditionnelles d'une distribution normale bivariée ont une distribution normale donnée par :*

$$f(y|x) = \frac{1}{\sigma_Y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_Y(x)}{\sigma_Y(x)}\right)^2\right\},$$

où

$$\mu_Y(x) = \mu_Y + \frac{\sigma_Y \rho}{\sigma_X}(x-\mu_X) \text{ et } \sigma_Y^2(x) = \sigma_Y^2(1-\rho^2).$$

et

$$f(x|y) = \frac{1}{\sigma_X(y)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{x-\mu_X(y)}{\sigma_X(y)}\right)^2\right\},$$

où

$$\mu_X(y) = \mu_X + \frac{\sigma_X \rho}{\sigma_Y}(y-\mu_Y) \text{ et } \sigma_X^2(y) = \sigma_X^2(1-\rho^2).$$

Démonstration. (pour $f(y|x)$)

$$\begin{aligned}
f(y|x) &= \frac{f(x,y)}{f_X(x)} \\
&= \frac{\dfrac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{\dfrac{-1}{2(1-\rho^2)}\left[\dfrac{(x-\mu_X)^2}{\sigma_X^2} - \dfrac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \dfrac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}}{\dfrac{1}{\sigma_X\sqrt{2\pi}} \exp -\dfrac{(x-\mu_X)^2}{2\sigma_X^2}} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right. \\
&\quad \left. + \frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{\rho^2(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \exp\left\{\frac{-1}{2\sqrt{1-\rho^2}}\left(\frac{y-\mu_Y}{\sigma_Y} - \frac{\rho(x-\mu_X)}{\sigma_X}\right)^2\right\} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \exp\left\{\frac{-1}{2\sqrt{1-\rho^2}}\left(\frac{y-\mu_Y-\frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y}\right)^2\right\} \\
&= \frac{1}{\sigma_Y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_Y(x)}{\sigma_Y(x)}\right)^2\right\}.
\end{aligned}$$

□

Le Theorème 6.6 montre que toutes les distributions conditionnelles sont également normales. La variance conditionnelle de Y pour une valeur fixée de x de la variable X vaut :

$$\mathrm{E}(Y|X=x) = \mu_Y(x) = \mu_Y + \frac{\sigma_Y\rho}{\sigma_X}(x-\mu_X).$$

De même, l'espérance conditionnelle de X pour une valeur fixée de y de la variable Y vaut :

$$\mathrm{E}(X|Y=y) = \mu_X(y) = \mu_X + \frac{\sigma_X\rho}{\sigma_Y}(y-\mu_Y).$$

La variance conditionnelle de Y pour une valeur fixée de x de la variable X vaut :

$$\mathrm{var}(Y|X=x) = \sigma_Y^2(x) = \sigma_Y^2(1-\rho^2).$$

Cette variance conditionnelle ne dépend pas de x. La variance conditionnelle de X pour une valeur fixée de y de la variable Y vaut :

$$\mathrm{var}(X|Y=y) = \sigma_X^2(y) = \sigma_X^2(1-\rho^2),$$

et ne dépend pas de y. Cette variance conditionnelle ne dépend pas de y. Les variances conditionnelles sont donc homoscédastiques (même variance).

Théorème 6.7.

$$\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dydx = \sigma_X\sigma_Y\rho.$$

Démonstration. La covariance peut également s'écrire

$$\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dydx - \mu_X\mu_Y.$$

On a :

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f(y|x)dydx = \int_{-\infty}^{\infty} xf_X(x)\int_{-\infty}^{\infty} yf(y|x)dydx \\
&= \int_{-\infty}^{\infty} xf_X(x)\left[\mu_Y + \frac{\sigma_Y\rho}{\sigma_X}(x-\mu_X)\right]dx = \mu_Y\int_{-\infty}^{\infty} xf_X(x)dx + \frac{\sigma_Y\rho}{\sigma_X}\int_{-\infty}^{\infty} xf_X(x)(x-\mu_X)dx \\
&= \mu_Y\mu_X + \frac{\sigma_Y\rho}{\sigma_X}\sigma_X^2 = \mu_Y\mu_X + \sigma_X\sigma_Y\rho.
\end{aligned}$$

Donc,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy = \sigma_X\sigma_Y\rho.$$

□

Le paramètre ρ est bien un coefficient de corrélation entre les variables X et X car il peut s'écrire :

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{\sigma_X\sigma_Y\rho}{\sigma_X\sigma_Y} = \rho.$$

Théorème 6.8. *Si les deux variables X et Y ont une distribution normale bivariée et que leur coefficient de corrélation est nul, alors X et Y sont indépendantes.*

Démonstration. Si $\rho = 0$, alors de l'Expression 6.2, la distribution jointe vaut :

$$\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y}\exp\left\{\frac{-1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left\{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right\}\right) \\
&= f_X(x)f_Y(y).
\end{aligned}$$

Dans ce cas, la densité jointe peut s'écrire comme le produit des deux densités marginales. Les deux variables sont donc indépendantes. □

Attention, si les deux variables n'ont pas une distribution normale bivariée, une covariance nulle n'implique plus que les variables sont indépendantes.

Deuxième partie

Exercices

Chapitre 7

Exercices : Variables, données statistiques, tableaux, effectifs

Exercice 7.1. Types de variables
Indiquer de quel type (qualitative nominale/ordinale, quantitative discrète/continue) sont les variables présentées ci-dessous :

(a) L'état civil des habitants de la Suisse.
(b) La taille des étudiants de l'Université de Neuchâtel.
(c) Le nombre de pages de cent supports de cours.
(d) Les professions reconnues en Suisse.
(e) Le nombre de ventes par jour d'un appareil électro-ménager au cours d'un mois.
(f) Le nombre d'accidents non-professionnels au cours d'une année.
(g) Le nombre d'enfants dans une famille.
(h) Le sexe des élèves d'une classe secondaire.
(i) La nationalité des élèves d'une classe secondaire.
(j) Le poids des nouveau-nés d'une ville.
(k) Le nombre de téléviseurs par famille.
(l) Le degré de qualification du personnel d'une entreprise.
(m) La couleur des yeux des étudiants de l'Université de Neuchâtel.
(n) Le nombre de jours de pluie pendant le mois d'août.

Solution

1. Qualitatives nominales : (a), (d), (h), (i), (m).
2. Qualitatives ordinales : (l).
3. Quantitative continues : (b), (e), (f), (j).
4. Quantitatives discrètes : (c), (e), (f), (g), (k), (n).

Exercice 7.2. Séries statistiques et graphiques
Les Tableaux 7.1, 7.2 et 7.3 présentent des séries statistiques. Pour chacune des séries :

1. Définir les unités statistiques.
2. Définir la variable.
3. Définir le type de variable.
4. Définir le domaine de la variable.
5. Construire le tableau de statistique.
6. Donner, si elle existe la fonction de répartition.
7. Construire les représentations graphiques adéquates de la série statistique et de sa fonction de répartition, si elle existe.

Table 7.1 – Nombre de noisettes mangées par 27 écureuils

11	11	11	11	13	13	13	13	13
16	16	16	16	16	16	17	17	17
17	19	19	19	19	19	19	19	19

Table 7.2 – Branche choisie par 24 étudiants (B=Biologie, C=chimie, M=mathématique, F=français)

B	B	B	B	C	C	C	C	C	C	C	C
M	M	M	M	M	M	F	F	F	F	F	F

Table 7.3 – Note moyenne de 22 étudiants

3	4	4	4	4	4.5	4.5	4.5	4.5	4.5	4.5
5	5	5	5.5	5.5	5.5	5.5	5.5	6	6	6

Note : En supposant qu'une note est multiple de 0.5 et entre 1 et 6.

Solution
Nombre de noisettes mangées par 27 écureuils :

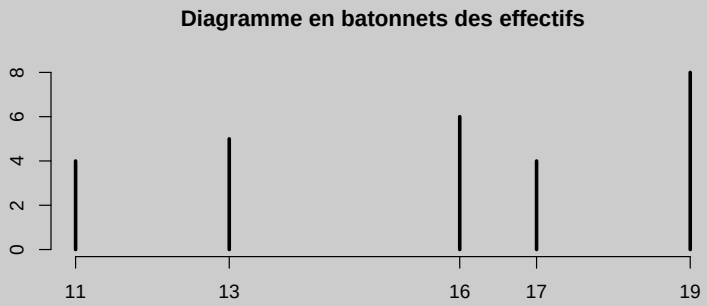
1. Écureuils
2. Nombre de noisettes mangées
3. Quantitative discrète
4. Domaine : $\{11, 13, 16, 17, 19\}$
5. Tableau statistique

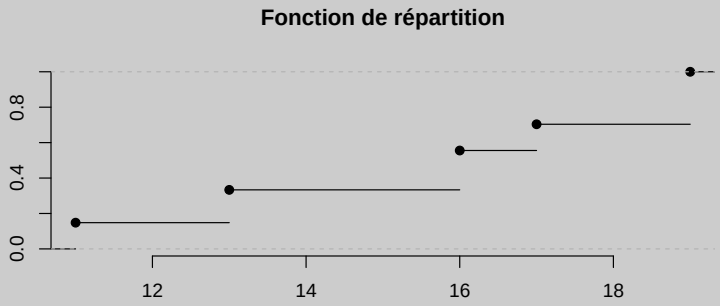
Nombre de noisettes mangées par écureuils

x_j	n_j	N_j	f_j	F_j
11	4	4	0.148	0.148
13	5	9	0.185	0.333
16	6	15	0.222	0.555
17	4	19	0.148	0.703
19	8	27	0.296	1.000
	$n = 27$		1	

6. Fonction de répartition

$$F(x) = \begin{cases} 0 & x < 11 \\ 0.148 & 11 \leq x < 13 \\ 0.333 & 13 \leq x < 15 \\ 0.555 & 15 \leq x < 17 \\ 0.703 & 17 \leq x < 19 \\ 1 & 19 \leq x \end{cases}$$





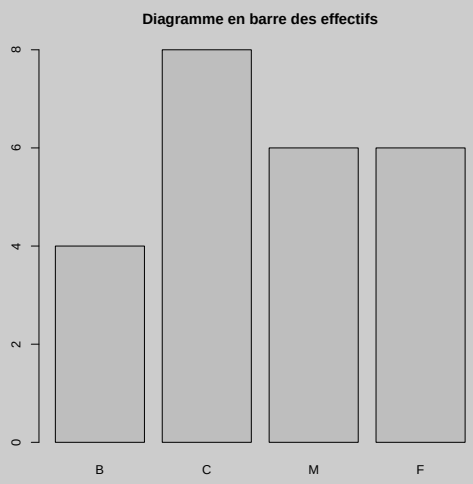
Branche choisie par 24 étudiants :

1. Étudiants
2. Branche choisie
3. Qualitative nominal
4. Domaine : $\{B, C, M, F\}$
5. Tableau statistique

Branche choisie par étudiant

x_j	n_j	f_j
B	4	0.167
C	8	0.333
M	6	0.250
F	6	0.250
	n=24	1.000

6. Pas de fonction de répartition car c'est une variable qualitative.
7. Graphique :



Note moyenne de 22 étudiants :

1. Etudiants
2. Note moyenne
3. Quantitative discrète
4. Domaine : 3, 4, 4.4, 5, 5.5, 6
5. Tableau statistique :

x_j	n_j	N_j	f_j	F_j
3	1	1	0.04	0.04
4	4	5	0.18	0.22
4.5	6	11	0.27	0.49
5	3	14	0.14	0.63
5.5	5	19	0.23	0.86
6	3	22	0.14	1.00
	$n = 2$		1	

6. La fonction de répartition

$$F(x) = \begin{cases} 0 & x < 3 \\ 0.04, & 3 \leq x < 4.0 \\ 0.22, & 4 \leq x < 4.5 \\ 0.49, & 4.5 \leq x < 5.0 \\ 0.63, & 5 \leq x < 5.5 \\ 0.86, & 5.5 \leq x < 6.0 \\ 1, & 6 \leq x \end{cases}$$

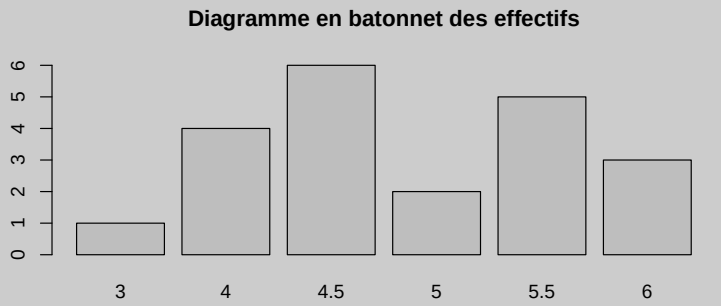
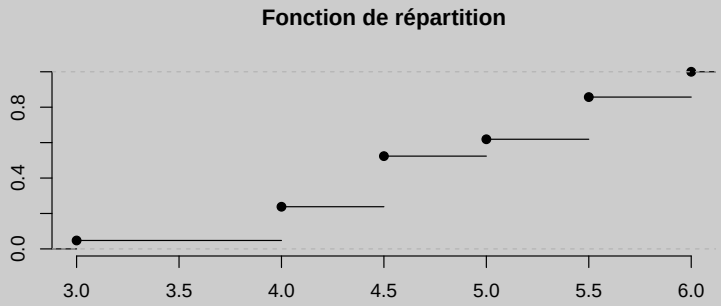


Diagramme en batonnet des effectifs
0
1
2
3
4
5
6
3
4
4.5
5
5.5
6


Fonction de répartition
0.0
0.4
0.8
3.0
3.5
4.0
4.5
5.0
5.5
6.0

Exercice 7.3. Poids d'élèves
On pèse les 50 élèves d'une classe et nous obtenons les résultats résumés dans le Tableau 7.4.

Table 7.4 – Poids des élèves

43	43	43	47	48	48	48	48	49	49
49	50	50	51	51	52	53	53	53	54
54	56	56	56	57	59	59	59	62	62
63	63	65	65	67	67	68	70	70	70
72	72	73	77	77	81	83	86	92	93

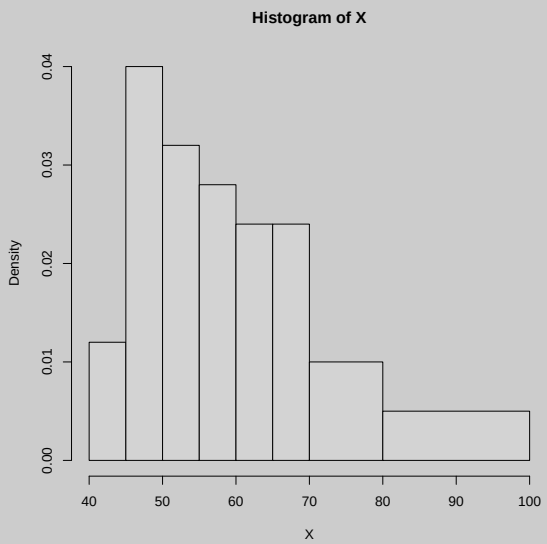
1. De quel type est la variable poids ?
2. Construisez le tableau statistique en adoptant les classes suivantes : [40; 45],]45; 50],]50; 55],]55; 60],]60; 65],]65; 70],]70; 80],]80; 100].
3. Construisez l'histogramme des fréquences.

Solution

1. La variable poids est de type quantitative continue.
2. Le tableau statistique est :

$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j	$f_j/(c_j^+ - c_j^-)$
[40; 45]	3	3	0.06	0.06	0.012
]45; 50]	10	13	0.20	0.26	0.040
]50; 55]	8	21	0.16	0.42	0.032
]55; 60]	7	28	0.14	0.56	0.028
]60; 65]	6	34	0.12	0.68	0.024
]65; 70]	6	40	0.12	0.80	0.024
]70; 80]	5	45	0.10	0.90	0.010
]80; 100]	5	50	0.10	1.00	0.005
	50		1		

3. L'histogramme est :



Il peut être construit grâce au code suivant :

En langage R

```
X=c(43,43,43,47,48,48,48,48,49,49,
```

```
49,50,50,51,51,52,53,53,53,54,
54,56,56,56,57,59,59,59,62,62,
63,63,65,65,67,67,68,70,70,70,
72,72,73,77,77,81,83,86,92,93)
classes=c(40,45,50,55,60,65,70,80,100)
plot(hist(X,breaks=classes))
```

Exercice 7.4. Variables et graphiques
Les Tableaux 7.5, 7.6, 7.7 et 7.8 contiennent quatre séries statistiques.
Pour la série statistique 7.8, on considérera le regroupement en classes suivant :

$$[500, 700[\ , [700, 900[\ , [900, 1100[\ , [1100, 1300[\ .$$

Pour chacune des séries :

1. définir les unités statistiques,
2. définir la variable,
3. définir le type de variable,
4. définir le domaine de la variable (l'ensemble de ses modalités),
5. construire le tableau statistique adéquat,
6. donner, si elle existe, l'expression de la fonction de répartition,
7. construire les représentations graphiques adéquates, à savoir :
   — Si la variable est *qualitative nominale* : diagramme en barres des effectifs.
   — Si la variable est *qualitative ordinale* : diagramme en barres des effectifs et diagramme en barres des effectifs cumulés.
   — Si la variable est *quantitative discrète* : diagramme en bâtonnets des effectifs et fonction de répartition.
   — Si la variable est *quantitative continue* : histogramme des fréquences et fonction de répartition.

Table 7.5 – Taux d'occupation professionnelle de 20 individus

S	S	S	S	Pa	Pa	Pa	Pa	Pa	Pa
Pl	Pl	Pl	Pl	Pl	Pl	Pl	Pl	Pl	Pl

(S=sans emploi, Pa=temps partiel, Pl=temps plein)

Table 7.6 – Nombre d'enfants par couple (40 couples)

0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4

Table 7.7 – Domicile des 20 élèves d'une classe

N	N	N	N	N	N	S	S	S	S
S	A	A	A	P	P	P	C	C	C

(N=Neuchâtel, S=Serrières, A=Auvernier, P=Peseux, C=Colombier)

Table 7.8 – Loyers mensuels en francs suisses de 25 appartements

589	591	612	754	771	775	786	821	884	913
918	952	967	998	1054	1103	1119	1174	1183	1191
1207	1211	1247	1277	1285					

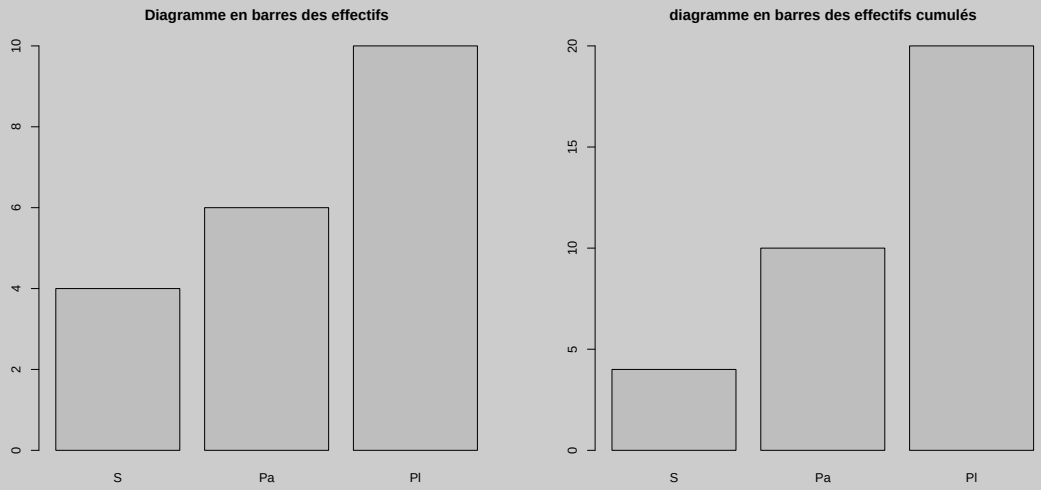
Solution

— **Série statistique 1 : Taux d'occupation professionnel de 20 individus**
   1. Les unités statistiques : individus.
   2. La variable : taux d'occupation professionnel.

3. La typologie de la variable : qualitative ordinale (S ≺ Pa ≺ Pl).
4. Le domaine : $\{S, Pa, Pl\}$.
5. Tableau du taux d'occupation professionnel de 20 individus

x_j	n_j	N_j	f_j	F_j
S	4	4	0.2	0.2
Pa	6	10	0.3	0.5
Pl	10	20	0.5	1.0
	n=20		1	

6. Pas de fonction de répartition, car la variable est qualitative.
7. Graphiques :



— **Série statistique 2 : Nombre d'enfants par couple**

1. Les unités statistiques : couples d'individus.
2. La variable : nombre d'enfants.
3. La typologie de la variable : quantitative discrète.
4. Le domaine : $\{0, 1, 2, 3, 4, \dots\}$.
5. Nombre d'enfants par couple.

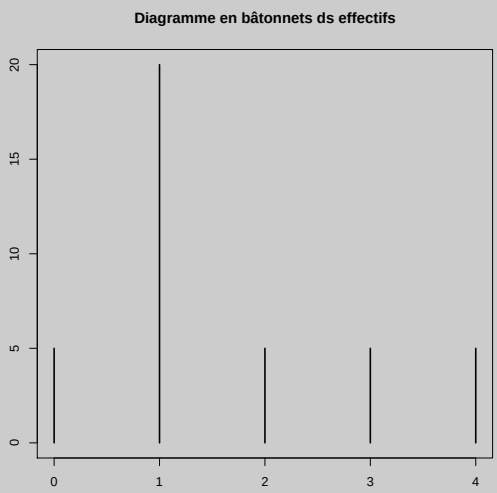
x_j	n_j	N_j	f_j	F_j
0	5	5	0.125	0.125
1	20	25	0.500	0.625
2	5	30	0.125	0.750
3	5	35	0.125	0.875
4	5	40	0.125	1.000
	n=40		1.000	

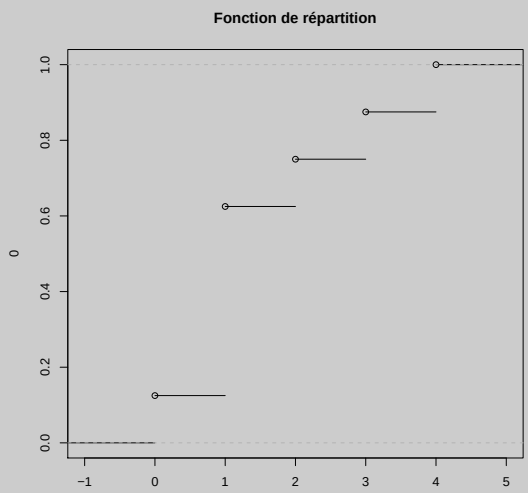
6. La fonction de répartition est donnée par :

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_j & x_j \leq x < x_{j+1} \\ 1 & x_J \leq x. \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.125 & 0 \leq x < 1 \\ 0.625 & 1 \leq x < 2 \\ 0.750 & 2 \leq x < 3 \\ 0.875 & 3 \leq x < 4 \\ 1 & 4 \leq x \end{cases}$$

7. Graphiques :



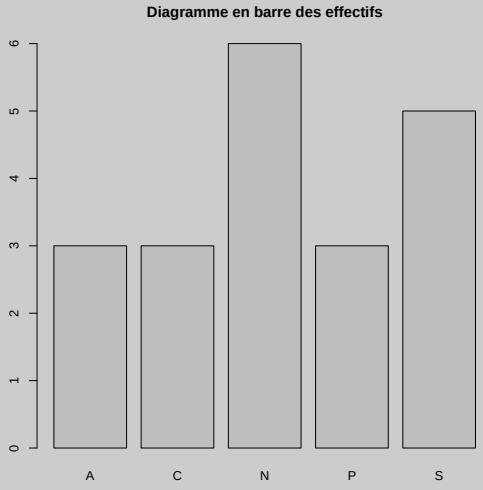


— **Série statistique 3 : Domicile des 20 élèves d'une classe**

1. Les unités statistiques : élèves.
2. La variable : domicile.
3. La typologie de la variable : qualitative nominale.
4. Le domaine : $\{N, S, A, P, C\}$.
5. Tableau du domicile des 20 élèves d'une classe.

x_j	n_j	f_j
N	6	0.30
S	5	0.25
A	3	0.15
P	3	0.15
C	3	0.15
	n=20	1.00

6. Pas de fonction de répartition, car c'est une variable qualitative.
7. Graphique :



— **Série statistique 4 : Loyers mensuels en francs suisses de 25 appartements**

1. les unités statistiques : appartements
2. la variable : loyer
3. la typologie de la variable : quantitative continue

4. le domaine : $[0, +\infty[$
5. Tableau des loyers mensuels en francs suisses de 25 appartements :

$\left[c_j^-, c_j^+\right)$	n_j	N_j	f_j	F_j	d_j	h_j
$[500, 700)$	3	3	0.12	0.12	0.0006	0.015
$[700, 900)$	6	9	0.24	0.36	0.0012	0.03
$[900, 1100)$	6	15	0.24	0.60	0.0012	0.03
$[1100, 1300)$	10	25	0.40	1.00	0.0020	0.05
	$n = 25$		1			

6. La fonction de répartition est donnée par :

$$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_J^+ \leq x \end{cases}$$

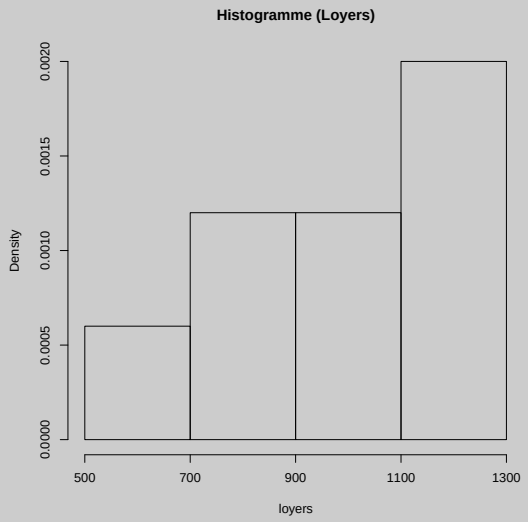
$$F(x) = \begin{cases} 0 & x < 500 \\ \frac{0.12}{200}(x - 500) & 500 \leq x < 700 \\ 0.12 + \frac{0.24}{200}(x - 700) & 700 \leq x < 900 \\ 0.36 + \frac{0.24}{200}(x - 900) & 900 \leq x < 1100 \\ 0.6 + \frac{0.4}{200}(x - 1100) & 1100 \leq x < 1300 \\ 1 & 1300 \leq x \end{cases}$$

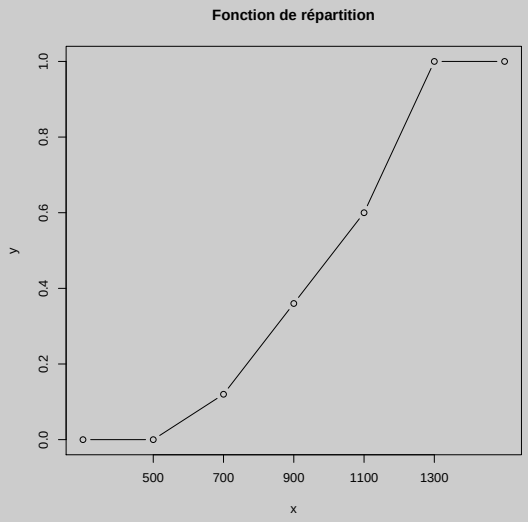
7. Histogramme :

$$d_j = \frac{f_j}{c_j^+ - c_j^-}, \text{ où } h_j = \frac{n_j}{c_j^+ - c_j^-}.$$

$$d_1 = \frac{0.12}{700 - 500} = 0.0006,$$

$$d_2 = \frac{0.24}{900 - 700} = 0.0012.$$





— Loyers mensuels en francs suisses de 25 appartements, regroupement des deux premières classes :

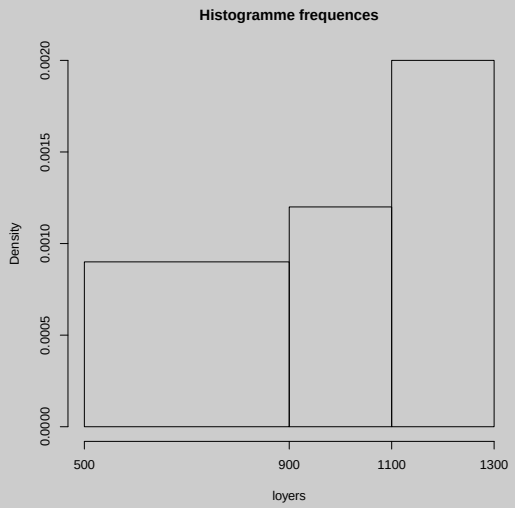
$$d_j = \frac{f_j}{c_j^+ - c_j^-}, \text{ ou } h_j = \frac{n_j}{c_j^+ - c_j^-}$$

$$d_1 = \frac{0.36}{900 - 500} = 0.0009,$$

$$d_2 = \frac{0.24}{1100 - 900} = 0.0012,$$

$$d_3 = \frac{0.4}{1300 - 1100} = 0.002.$$

$\left[c_j^-, c_j^+\right)$	n_j	N_j	f_j	F_j	h_j
$[500, 900)$	9	9	0.36	0.36	0.0225
$[900, 1100)$	6	15	0.24	0.60	0.03
$[1100, 1300)$	10	25	0.40	1.00	0.05
	n=25		1		



Exercice 7.5. Histogrammes
Soit la série suivante d'une variable continue donnée dans le Tableau 7.9 :

Table 7.9 – Série continue

712	712.22	729	731.5	732.7	745.74	749	752.8	789.3	789.82

Nous considérons le regroupement en classes suivant : $[710, 730[$, $[730, 750[$, $[750, 770[$, $[770, 790[$.

1. Dressez le tableau des fréquences/fréquence cumulée/densité.
2. Dessinez le graphe de la fonction de répartition
3. Dessinez l'histogramme des fréquences associé à nos données.
4. Maintenant nous regroupons les classes $[750, 770[$ et $[770, 790[$ en une seule classe $[750, 790[$. Dessinez à nouveau un histogramme des fréquences.

Solution
Le Tableau suivant contient la fréquence, la fréquence cumulée ainsi que la densité par classe.

Tableau statistique

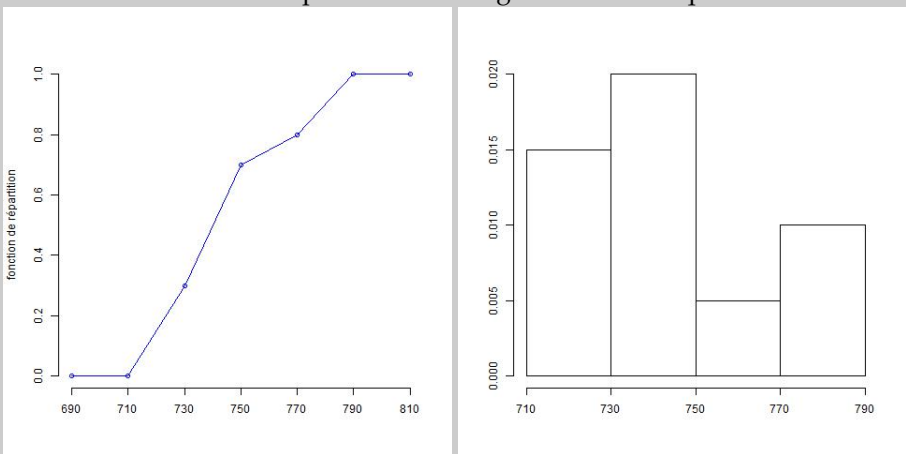
$\left[c_j^-, c_j^+\right)$	f_j	F_j	d_j
$[710, 730)$	0.3	0.3	0.015
$[730, 750)$	0.4	0.7	0.02
$[750, 770)$	0.1	0.8	0.005
$[770, 790)$	0.2	1.0	0.01

Fonction de répartition :

$$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_J^+ \leq x. \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 710 \\ \dfrac{0.3}{20}(x - 710) & 710 \leq x < 730 \\ 0.3 + \dfrac{0.4}{20}(x - 730) & 730 \leq x < 750 \\ 0.7 + \dfrac{0.1}{20}(x - 750) & 750 \leq x < 770 \\ 0.8 + \dfrac{0.2}{20}(x - 770) & 770 \leq x < 790 \\ 1 & 790 \leq x. \end{cases}$$

Fonction de répartition et histogramme des fréquences



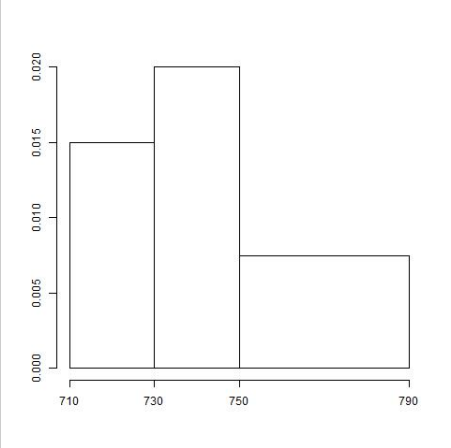
Si nous agrégeons les deux dernières classes, on obtient :

Tableau statistique

$\left[c_j^-, c_j^+\right)$	f_j	F_j	d_j
$[710, 730)$	0.3	0.3	0.015
$[730, 750)$	0.4	0.7	0.02
$[750, 790)$	0.3	1.0	0.0075

Nouvel histogramme des fréquences :

L'histogramme des fréquences 2



Exercice 7.6. Histogramme et classes
Le Tableau 7.10 présente l'âge des 50 employés d'une entreprise.

TABLE 7.10 – Âge des 50 employés d'une entreprise

23	23	24	24	25	25	26	26	28	28
28	29	30	31	31	32	33	33	34	34
35	35	36	36	36	37	37	37	38	38
39	40	41	41	42	43	44	44	45	45
47	48	48	49	49	54	55	56	57	58

1. Effectuer un regroupement en classe selon la règle de Sturge et donner le tableau statistique correspondant.
2. Tracer l'histogramme des fréquences et la fonction de répartition.
3. Quel découpage obtiendrait-on à l'aide de la règle de Yule ?

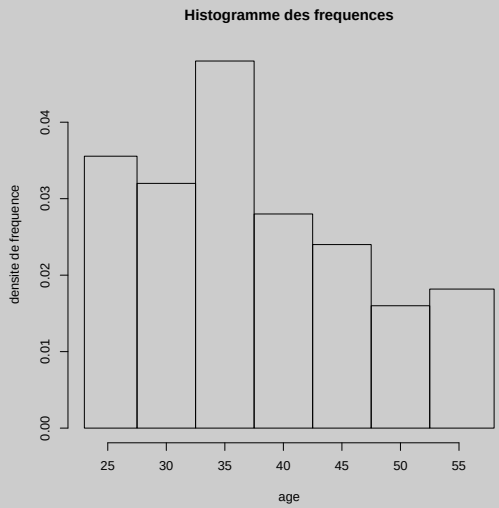
Solution

1. Règle de Sturge : $J = 1 + (3.3 \log n)$.
   Ici, $J = 1 + (3.3 \log 50) = 6.61$. On arrondi à l'entier le plus proche $\Rightarrow$ on fait 7 classes.
   Longueur de l'intervalle :
   $$a_j = \frac{x_{max} - x_{min}}{J} = \frac{58 - 23}{7} = 5.$$

   Âge des 50 employés d'une entreprise

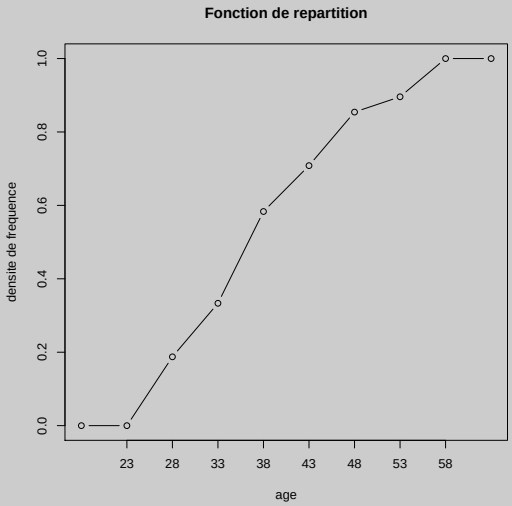
$\left[c_j^-, c_j^+\right)$	n_j	N_j	f_j	F_j
$[23, 28[$	8	8	0.16	0.16
$[28, 33[$	8	16	0.16	0.32
$[33, 38[$	12	28	0.24	0.56
$[38, 43[$	7	35	0.14	0.70
$[43, 48[$	6	41	0.12	0.82
$[48, 53[$	4	45	0.08	0.90
$[53, 58[$	5	50	0.10	1.00
	$n = 50$		1	

2. Pour tracer l'histogramme des fréquence :
   Densité de fréquence : $d_j = \frac{f_i}{a_i}$.
   On a alors $d_1 = 0.16/5 = 0.032$, $d_2 = 0.032$, $d_3 = 0.048$, $d_4 = 0.028$, $d_5 = 0.024$, $d_6 = 0.016$, $d_7 = 0.020$.



   Pour la fonction de répartition :
   $$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_J^+ \leq x \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 23 \\ \frac{0.16}{5}(x-23) & 23 \leq x < 28 \\ 0.16 + \frac{0.16}{5}(x-28) & 28 \leq x < 33 \\ 0.32 + \frac{0.24}{5}(x-33) & 33 \leq x < 38 \\ 0.56 + \frac{0.14}{5}(x-38) & 38 \leq x < 43 \\ 0.70 + \frac{0.12}{5}(x-43) & 43 \leq x < 48 \\ 0.82 + \frac{0.08}{5}(x-48) & 48 \leq x < 53 \\ 0.90 + \frac{0.10}{5}(x-53) & 53 \leq x < 58 \\ 1 & 58 \leq x \end{cases}$$



3. Règle de Yule : $J = 2.5\sqrt[4]{n}$.
   Ici : $J = 2.5\sqrt[4]{50} = 6.65$. On arrondi à l'entier le plus proche $\Rightarrow$, ce qui fait 7 classes. C'est le même découpage qu'avec la règle de Sturge.

Exercice 7.7. Variables, types et graphiques
Pour chacune des séries statistiques données dans les Tableaux 7.11, 7.12 et 7.13 :

1. Définir la variable.
2. De quel type de variable s'agit-il ?
3. Préciser les modalités de cette variable.
4. Donner le tableau statistique complet (calculer n_j, N_j, f_j et F_j).
5. Donner une représentation graphique des fréquences.

Table 7.11 – Nombre de jours de chômage pour 40 personnes

180	10	30	50	420	30	180	360	200	30
360	120	500	200	30	420	360	370	360	150
180	280	30	500	180	720	420	180	40	500
120	180	194	400	30	360	40	400	180	200

Table 7.12 – Répartition de la population d'un pays par groupe d'âge

classe d'âge	n_j
0 à 19 ans	1 621 600
20 à 39 ans	2 180 900
40 à 64 ans	2 147 100
65 à 79 ans	746 900
80 ans et plus	272 000
	6 968 500

Table 7.13 – Qualité de production de 30 produits (D = défectueux, Q = de bonne qualité)

Q	D	Q	D	Q	Q	Q	Q	Q	Q
D	Q	Q	D	Q	D	D	Q	Q	Q
D	D	D	Q	Q	Q	Q	Q	Q	D

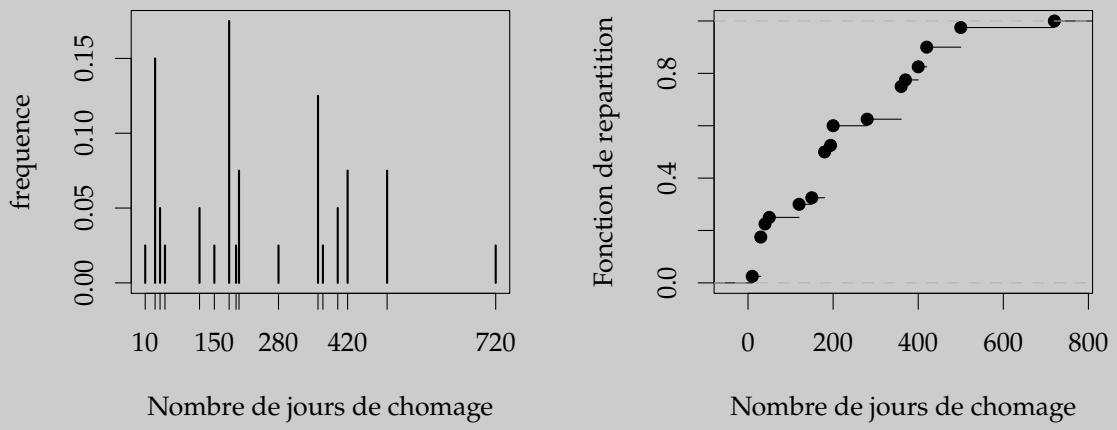
Solution

— La première variable est le nombre de jours de chômage de quarante personnes et elle est quantitative discrète. Les modalités sont $10, 30, 40, 50, \ldots, 720$.
— La deuxième variable est la classe d'âge et elle est qualitative ordinale. Les modalités sont '0 à 19 ans', '20 à 39 ans', . . . , '80 ans et plus'.
— La troisième variable est la qualité de production de trente produits et elle est qualitative nominale. Les modalités sont 'défectueux' et 'de bonne qualité'.
— Graphiques.

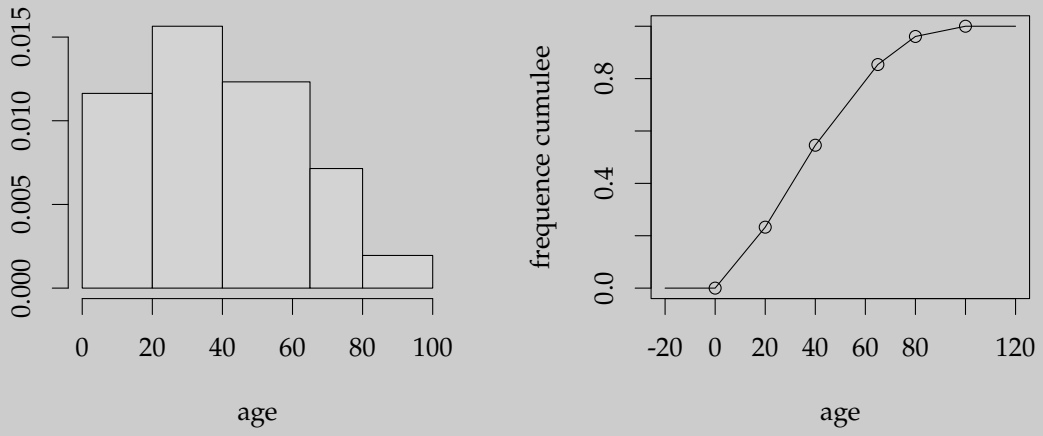
En langage R

```
# Variable~: Nombre de jours de chômage pour 40 personnes
install.packages("tikzDevice")
library(tikzDevice)
# Variable : Nombre de jours de chômage pour 40 personnes
X1=c(180,10,30,50,420,30,180,360,200,30,360,120,500,200,30,420,
360,370,360,150,180,280,30,500,180,720,420,180,40,500,120,180,
194,400,30,360,40,400,180,200)
freq=table(X1)/length(X1)
tikz(file="graph1.tex",width=3,height=3)
plot(freq,xlab="Nombre de jours de chomage",ylab="frequence")
dev.off()
tikz(file="graph2.tex",width=3,height=3)
plot(ecdf(X1),xlab="Nombre de jours de chomage",
ylab="Fonction de repartition",main="")
dev.off()
# Variable~: population d'un pays par groupe d'âge
breaks=c(0,20,40,65,80,100)
counts=c(1621600,2180900,2147100,746900,272000)
larg=breaks[2:6]-breaks[1: 5]
density=counts/sum(counts)/larg
X2=list(breaks=c(0,20,40,65,80,100),
counts=counts,
density=density,
mids=c(10,30,52.5,72.5,100),
name="age",
equidist=FALSE)
attr(X2,"class") ="histogram"
tikz(file="graph3.tex",width=3,height=3)
plot(X2,xlab="age",ylab="",main="")
dev.off()
T=cbind(breaks=c(-20,breaks,120),
ccounts=cumsum(c(0,0,counts,0)/sum(counts)) )
tikz(file="graph4.tex",width=3,height=3)
plot(T,type="l",xlab="age",ylab="frequence cumulee")
points(T[2:7,])
dev.off()
# Variable~: Qualité de production
X3=c("Q","D","Q","D","Q","Q","Q","Q","Q","Q","D","Q","Q","D","Q",
"D","D","Q","Q","Q","D","D","D","Q","Q","Q","Q","Q","Q","D")
F=table(X3)/length(X3)
tikz(file="graph5.tex",width=3,height=3)
barplot(F,main="")
dev.off()
tikz(file="graph6.tex",width=3,height=3)
pie(F,main="")
dev.off()
```

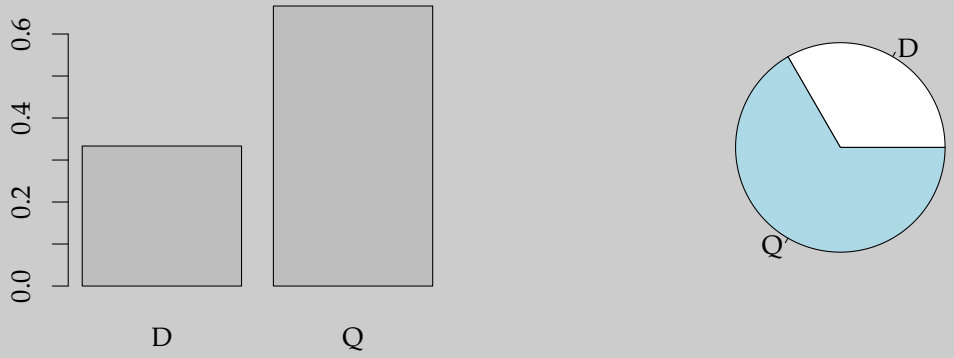
Nombre de jours de chômage pour 40 personnes



Répartition de la population d'un pays par groupe d'âge



Qualité de production de 30 produits (D = défectueux, Q = de bonne qualité)



Chapitre 8

Exercices : Statistique descriptive univariée

Exercice 8.1. Opérateur de sommation 1

1. Soit la série statistique $x_i \ : \ x_1 = 3, \ x_2 = 5, \ x_3 = 8, \ x_4 = 4$.
   À l'aide des propriétés du signe de sommation, calculer :
   (a) $\displaystyle\sum_{i=1}^{4} x_i$,
   (b) $\displaystyle\sum_{i=1}^{4} (6x_i - 5)$,
   (c) $\displaystyle\sum_{i=1}^{4} (3i + 2)$,
   (d) $\displaystyle\sum_{i=1}^{3} (x_i + 1)$.
2. Réécrivez, en utilisant le signe de sommation, les expressions suivantes :
   (a) $1 + 4 + 9 + 16 + 25 + 36$
   (b) $1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 + 19$
   (c) $(x_1 + x_2 + x_3 + \cdots + x_n)/n$,
   (d) $x_3 + x_6 + x_9 + x_{12} + \cdots + x_{60}$,
   (e) $x_1 + 2x_2 + 3x_3 + 4x_4 + \cdots + 8x_8$,
   (f) $5x_1 + 5x_2 + 5x_3 + 5x_4 + \cdots + 5x_8$.
3. Ecrire les expressions suivantes sous la forme d'une somme unique :
   (a) $\displaystyle\sum_{p=0}^{3} x_p^2 + \sum_{p=4}^{6} x_p^2$,
   (b) $\displaystyle\sum_{p=0}^{3} x_p^2 + \sum_{q=4}^{6} x_q^2$.
4. Soient les séries statistiques
   $x_i \ : x_1 = 2, \ x_2 = 3, \ x_3 = 5, \ x_4 = 7$, et $y_i \ : y_1 = 4, \ y_2 = 6, \ y_3 = 8, \ y_4 = 9$.
   Calculer :
   (a) $\displaystyle\sum_{i=1}^{4} x_i y_i$,
   (b) $\displaystyle\sum_{i=1}^{4} x_i \sum_{i=1}^{4} y_i$.
5. Soient $\bar{x}$ et $\bar{y}$, la moyenne arithmétique des x_i et y_i respectivement. Montrer que :

(a) $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$,

(b) $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$.

Solution

1. (a) $\sum_{i=1}^{4} x_i = 3 + 5 + 8 + 4 = 20$

   (b) $\sum_{i=1}^{4}(6x_i - 5) = \sum_{i=1}^{4} 6x_i - \sum_{i=1}^{4} 5 = 6\sum_{i=1}^{4} x_i - 4 \times 5 = 6 \times 20 - 20 = 100$

   (c) $\sum_{i=1}^{4}(3i + 2) = \sum_{i=1}^{4} 3i + \sum_{i=1}^{4} 2 = 3\sum_{i=1}^{4} i + 4 \times 2 = 3 \times \frac{4(4+1)}{2} + 4 \times 2 = 38$

   (d) $\sum_{i=1}^{3}(x_i + 1) = \sum_{i=1}^{3} x_i + \sum_{i=1}^{3} 1 = 3 + 5 + 8 + 3 \times 1 = 19$

2. Ecrire, à l'aide du signe de sommation, les expressions suivantes :

   (a) $1 + 4 + 9 + 16 + 25 + 36 = 1^2 + 2^2 + \cdots + 6^2 = \sum_{i=1}^{6} i^2$

   (b) $1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 + 19 = \sum_{i=1}^{10}(2i - 1)$ ou $\sum_{i=0}^{9}(2i + 1)$.

   (c) $(x_1 + x_2 + x_3 + \cdots + x_n)/n = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ (moyenne arithmétique).

   (d) $x_3 + x_6 + x_9 + x_{12} + \cdots + x_{60} = \sum_{i=1}^{20} x_{3i}$.

   (e) $x_1 + 2x_2 + 3x_3 + 4x_4 + \cdots + 8x_8 = \sum_{i=1}^{8} i x_i$.

   (f) $5x_1 + 5x_2 + 5x_3 + 5x_4 + \cdots + 5x_8 = 5\sum_{i=1}^{8} x_i$.

3. Ecrire les expressions suivantes sous la forme d'une somme unique :

   (a) $\sum_{p=0}^{3} x_p^2 + \sum_{p=4}^{6} x_p^2 = \sum_{p=0}^{6} x_p^2$.

   (b) $\sum_{p=0}^{3} x_p^2 + \sum_{q=4}^{6} x_q^2 = \sum_{p=0}^{3} x_p^2 + \sum_{p=4}^{6} x_p^2 = \sum_{p=0}^{6} x_p^2$.

4. Calculer :

   (a) $\sum_{i=1}^{4} x_i y_i = x_1 \times y_1 + x_2 \times y_2 + x_3 \times y_3 + x_4 \times y_4 = 2 \times 4 + 3 \times 6 + 5 \times 8 + 7 \times 9 = 129$,

   (b) $\sum_{i=1}^{4} x_i \sum_{i=1}^{4} y_i = (x_1 + x_2 + x_3 + x_4)\left(y_1 + y_2 + y_3 + y_4\right) = (2 + 3 + 5 + 7)(4 + 6 + 8 + 9) = 459$.

5. Soit $\bar{x}$ et $\bar{y}$, la moyenne arithmétique des x_i et y_i respectivement. Montrer que :

   (a) $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2$

   $= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x} \times n\bar{x} + n\bar{x}^2$, (car par définition, $\sum_{i=1}^{n} x_i = n\bar{x}$)

   $= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

(b) $\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n} y_i\bar{x} + \sum_{i=1}^{n}\bar{x}\bar{y}$

$\displaystyle= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}$, (car par définition, $\displaystyle\sum_{i=1}^{n} x_i = n\bar{x}$ et $\displaystyle\sum_{i=1}^{n} y_i = n\bar{y}$)

$\displaystyle= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$

Exercice 8.2. Opérateur de sommation 2

1. Soit la série statistique : $x_1 = 2$, $x_2 = 4$, $x_3 = 7$, $x_4 = 8$. Calculer :
   (a) $\sum_{i=1}^{4} x_i$
   (b) $\sum_{i=1}^{4} (3x_i - 2)$
   (c) $\sum_{i=1}^{3} (x_i + i)$
   (d) $\sum_{i=2}^{4} (i - 2)$
   (e) $\sum_{i=2}^{4} (x_i^2 - 2)$
2. Réecriver en utilisant le signe de sommation, les expressions suivantes :
   (a) $x_{40} + x_{42} + x_{44} + x_{46}$
   (b) $3x_2 + 4x_3 + 5x_4 + 6x_5$
   (c) $2x_1 + 2x_2 + 2x_3 + 2x_4$

Solution

1. (a) $\sum_{i=1}^{4} x_i = 2 + 4 + 7 + 8 = 21$
   (b) $\sum_{i=1}^{4} (3x_i - 2) = \sum_{i=1}^{4} 3x_i - \sum_{i=1}^{4} 2 = 3 \sum_{i=1}^{4} x_i - 4 \times 2 = 3 \times 21 - 8 = 55$
   (c) $\sum_{i=1}^{3} (x_i + i) = \sum_{i=1}^{3} x_i + \sum_{i=1}^{3} i = (2 + 4 + 7) + (1 + 2 + 3) = 13 + 6 = 19$
   (d) $\sum_{i=2}^{4} (i - 2) = \sum_{i=2}^{4} i - \sum_{i=2}^{4} 2 = (2 + 3 + 4) - (3 \times 2) = 9 - 6 = 3$
   (e) $\sum_{i=2}^{4} (x_i^2 - 2) = \sum_{i=2}^{4} (x_i^2) - \sum_{i=2}^{4} (2) = (4^2 + 7^2 + 8^2) - (3 \times 2) = 129 - 6 = 123$
2. (a) $x_{40} + x_{42} + x_{44} + x_{46} = \sum_{i=0}^{3} x_{40+2i}$
   (b) $3x_2 + 4x_3 + 5x_4 + 6x_5 = \sum_{i=3}^{6} i x_{i-1} = \sum_{i=2}^{5} (i + 1) x_i$ (les deux réponses sont justes)
   (c) $2x_1 + 2x_2 + 2x_3 + 2x_4 = 2 \sum_{i=1}^{4} x_i$

Exercice 8.3. Variance avec une double somme

1. Montrez que
$$s_x^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

2. Montrez que
$$s_x \leq E\sqrt{\frac{n-1}{2n}},$$
où E est l'étendue.

Solution

1. On développe l'expression :
$$\begin{aligned}
\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^2 + x_j^2 - 2x_ix_j) \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i^2 + \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_j^2 - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}2x_ix_j \\
&= \frac{1}{2n}\sum_{i=1}^{n}x_i^2 + \frac{1}{2n}\sum_{j=1}^{n}x_j^2 - \frac{1}{n}\sum_{i=1}^{n}x_i\frac{1}{n}\sum_{j=1}^{n}x_j \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}x_i\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = s_x^2.
\end{aligned}$$

2. Comme $|x_i - x_j| \leq E$ pour tout i, j, on a :
$$\begin{aligned}
s_x^2 &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(x_i - x_j)^2 \\
&\leq \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(x_{(1)} - x_{(n)})^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}E^2 \\
&= \frac{1}{2n^2}n(n-1)E^2 = \frac{n-1}{2n}E^2.
\end{aligned}$$

Donc,
$$s_x \leq E\sqrt{\frac{n-1}{2n}}.$$

Exercice 8.4. Classes d'élèves
Dans deux classes de niveau équivalent d'une même école, les notes, obtenues par les élèves à l'occasion d'une même épreuve, sont données dans les tableaux 8.1 et 8.2. Comparer les deux séries statistiques. Pour chaque série :

1. Faire le tableau statistique en calculant les effectifs et les effectifs cumulés.
2. Construire le diagramme en bâtonnets des effectifs.
3. Calculer la moyenne, le mode et l'étendue.
4. Calculer les premier, deuxième (la médiane) et troisiéme quartiles et ensuite la distance interquartile.
5. Calculer la variance et l'écart-type.

Table 8.1 – Classe A

9	15	15	7	11	12	14	10	11	8	8	11	11	14	8	10	11	11	10	11
7	15	12	6	14	9	15	8	8	14	15	10	11	13	11	11	15	12	15	10

Table 8.2 – Classe B

11	9	8	13	9	8	13	14	15	15	10	10	7	15	15	7	14	9	3	10
15	10	15	8	15	8	14	9	6	13	12	11	9	9	13	14	8	13	8	5

Solution

Le tableau statistique pour la Classe A ($n = 40,\ J = 10$)

x_j	n_j	N_j	$x_j n_j$	x_j^2	$x_j^2 n_j$
6	1	1	6	36	36
7	2	3	14	49	98
8	5	8	40	64	320
9	2	10	18	81	162
10	5	15	50	100	500
11	10	25	110	121	1210
12	3	28	36	144	432
13	1	29	13	169	169
14	4	33	56	196	784
15	7	40	105	225	1575
$\sum_{j=1}^{J}$			448		5286

— La moyenne : $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \dfrac{1}{n}\sum_{j=1}^{J} x_j n_j = \dfrac{1}{40} \times 448 = 11.2.$
— Le mode : $x_M = 11$.
— L'étendue : E = 15 − 6 = 9.
— Le premier quartile : Comme $np = \frac{1}{4} \times 40 = 10$ est un nombre entier, on a

$$x_{1/4} = \frac{1}{2}\left\{x_{(10)} + x_{(11)}\right\} = (9 + 10)/2 = 9.5.$$

— La médiane : Comme $np = \frac{1}{2} \times 40 = 20$ est un nombre entier, on a

$$x_{1/2} = \frac{1}{2}\left\{x_{(20)} + x_{(21)}\right\} = (11 + 11)/2 = 11.$$

— Le troisième quartile : Comme $np = \frac{3}{4} \times 40 = 30$ est un nombre entier, on a

$$x_{3/4} = \frac{1}{2}\left\{x_{(30)} + x_{(31)}\right\} = (14+14)/2 = 14.$$

— La distance interquartile : $IQ = x_{3/4} - x_{1/4} = 4.5$.
— La variance :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2,$$

mais encore (ce que on va utiliser) :

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} x_j^2 n_j - \bar{x}^2 = \frac{1}{40}\sum_{j=1}^{10} x_j^2 n_j - \bar{x}^2 = \frac{1}{40} \times 5286 - 11.2^2 = 132.15 - 125.44 = 6.71.$$

— L'écart-type : $s_x = \sqrt{s_x^2} = \sqrt{6.71} = 2.59$.

Le tableau statistique pour la Classe B ($n = 40,\ J = 12$)

y_j	n_j	N_j	$y_j n_j$	y_j^2	$y_j^2 n_j$
3	1	1	3	9	9
5	1	2	5	25	25
6	1	3	6	36	36
7	2	5	14	49	98
8	6	11	48	64	384
9	6	17	54	81	486
10	4	21	40	100	400
11	2	23	22	121	242
12	1	24	12	144	144
13	5	29	65	169	845
14	4	33	56	196	784
15	7	40	105	225	1575
$\sum_{j=1}^{J}$			430		5028

— La moyenne : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{j=1}^{J} y_j n_j = \frac{1}{40} \times 430 = 10.75$.
— Le mode : $y_M = 15$.
— L'étendue : E = 15 − 3 = 12.
— Le premier quartile : Comme $np = \frac{1}{4} \times 40 = 10$ est un nombre entier, on a

$$y_{1/4} = \frac{1}{2}\left\{y_{(10)} + y_{(11)}\right\} = (8+8)/2 = 8.$$

— La médiane : Comme $np = \frac{1}{2} \times 40 = 20$ est un nombre entier, on a

$$y_{1/2} = \frac{1}{2}\left\{y_{(20)} + y_{(21)}\right\} = (10+10)/2 = 10.$$

— Le troisième quartile : Comme $np = \frac{3}{4} \times 40 = 30$ est un nombre entier, on a

$$y_{3/4} = \frac{1}{2}\left\{y_{(30)} + y_{(31)}\right\} = (14+14)/2 = 14.$$

— La distance interquartile : $IQ = y_{3/4} - y_{1/4} = 6$
— La variance :

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2,$$

mais encore (ce que on va utiliser) :

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{J} y_j^2 n_j - \bar{y}^2 = \frac{1}{40}\sum_{j=1}^{12} y_j^2 n_j - \bar{y}^2 = \frac{1}{40} \times 5028 - 10.75^2 = 125.7 - 115.56 = 10.14.$$

— L'écart-type : $s_y = \sqrt{s_y^2} = \sqrt{10.14} = 3.18$.

Exercice 8.5. Moyennes arithmétique, géometrique et harmonique
Soit la série statistique : x_i : 7, 15, 6, 8, 11. Calculer les moyennes arithmétique, géometrique et harmonique.

Solution

1. $\bar{x} = \dfrac{\sum_{i=1}^{5} x_i}{5} = \dfrac{7+15+6+8+11}{5} = 9.4.$
2. Méthode 1 :

$$\begin{aligned} G &= \exp \frac{1}{n} \sum_{i=1}^{n} \log x_i = \exp \frac{1}{n}(\log 7 + \log 15 + \log 6 + \log 8 + \log 11) \\ &= \exp 2.18461 = e^{2.18461} = 8.88719. \end{aligned}$$

Méthode 2 : $G = \left(\prod_{i=1}^{n} x_i \right)^{1/n} = (7 \times 15 \times 6 \times 8 \times 11)^{1/5} = 8.88719.$

3. $H = \dfrac{5}{\sum_{i=1}^{5} \frac{1}{x_i}} = \dfrac{5}{\frac{1}{7} + \frac{1}{15} + \frac{1}{6} + \frac{1}{8} + \frac{1}{11}} = 8.44.$

Exercice 8.6. Changement d'origine
Soit un échantillon $x_i,\ i = 1, \ldots, n$.

1. En posant $y_i = ax_i + b,\ i = 1, \ldots, n$, vérifier algébriquement que, quels que soient a et b, $\bar{y} = a\bar{x} + b$.
2. À partir des données numériques suivantes : $x_i\ :\ 6,\ 12,\ 21,\ 0,\ 6,\ 9,\ 15,\ 3$, calculer $\sum_{i=1}^{n} x_i$ et ensuite la moyenne $\bar{x}$.
3. En utilisant les résultats (1) et (2), calculer, d'une façon simple la moyenne des y_i suivants : $1906,\ 1912,\ 1921,\ 1900,\ 1906,\ 1909,\ 1915,\ 1903$, en sachant que $a = 1$ et $b = 1900$.

Solution

— $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n} = \dfrac{\sum_{i=1}^{n}(ax_i + b)}{n} = \dfrac{a\sum_{i=1}^{n} x_i}{n} + \dfrac{nb}{n} = a\bar{x} + b.$

— $\displaystyle\sum_{i=1}^{8} x_i = 72$ and $\bar{x} = \dfrac{\sum_{i=1}^{8} x_i}{8} = \dfrac{72}{8} = 9.$

— Comme $y_1 = x_i + 1900$, pour tout $i = 1, \ldots, n$, $\bar{y} = a\bar{x} + b = 9 + 1900 = 1909$.

Exercice 8.7. Quelle moyenne ?
Le district de Neuchâtel se compose de 10 communes. En connaîssant le nombre d'habitants par commune et le nombre de véhicules par habitant (Tableau 8.3), déterminer le nombre moyen de véhicules par habitant. De quelle moyenne s'agit-il ?

Table 8.3 – Nombre de véhicules par habitant selon les communes

Localité	Habitants	Vhc/hab
Cornaux	1570	0.4694
Cressier	1701	0.4556
Enges	280	0.525
Hauterive	2357	0.5002
Le Landeron	4031	0.4646
Lignières	713	0.6437
Marin-Epagnier	3710	0.4396
Neuchâtel	31800	0.4495
St-Blaise	2961	0.5569
Thielle-Wavre	462	0.593

Solution
Le nombre total d'habitants est 49585. Donc, la moyenne pondérée est égale à :

$$\bar{x} = \frac{1570 \times 0.4694 + 1701 \times 0.4556 + \cdots + 462 \times 0.593}{49585} = 0.463.$$

Exercice 8.8. Calcul de paramètres
Soit la série statistique suivante : $2, 2, 4, 4, 4, 4, 6, 8, 8, 8, 11, 11$.

1. Calculer le mode.
2. Calculer la moyenne arithmétique.
3. Calculer la moyenne harmonique.
4. Calculer la médiane.
5. Calculer le premier et le troisième quartile.
6. Calculer l'étendue.
7. Calculer la distance interquartile.

Solution

1. Mode : 4.
2. Moyenne arithmétique : $\bar{x} = \dfrac{1}{12}\sum_{i=1}^{12} x_i = 6$.
3. Moyenne harmonique : $H = \dfrac{12}{\sum_{i=1}^{12} \frac{1}{x_i}} = 4.406$.
4. Médiane : n est pair donc $x_{1/2} = \dfrac{x_{(6)} + x_{(7)}}{2} = \dfrac{4+6}{2} = 5$.
5. Premier quartile : $n \times p = 12 \times 1/4 = 3$. $n \times p$ est un nombre entier. Donc, $x_{1/4} = x_{(3)} = 4$.
   Troisième quartile : $n \times p = 3/4 \times 12 = 9$. $n \times p$ est un nombre entier. Donc, $x_{3/4} = x_{(9)} = 8$.
6. $\mathrm{E} = x_{(n)} - x_{(1)} = 11 - 2 = 9$.
7. $IQ = x_{3/4} - x_{1/4} = 8 - 4 = 4$.

Exercice 8.9. Salaires hommes et femmes
Soient les séries statistiques suivantes : x_i : salaire mensuel de 5 hommes et y_i : salaire mensuel de 5 femmes contenues dans le Tableau 8.4.

Table 8.4 – Salaire mensuels de 5 hommes et 5 femmes

Hommes x_i	250	280	300	350	500

Femmes y_i	200	220	280	350	400

1. Calculer le salaire moyen des femmes.
2. Calculer le salaire moyen des hommes.
3. Calculer l'écart moyen absolu des x_i.
4. Calculer la variance des x_i et des y_i.
5. Calculer l'écart médian absolu de x_i.
6. Calculer le moment centré d'ordre 3 des x_i.
7. Calculer le coefficient d'asymétrie de Fischer des x_i. Que peut-on dire sur la distribution ? Interpréter.

Solution

1. $\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = 290.$
2. $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 336.$
3. $e_{moy} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| = \frac{1}{5}\sum_{i=1}^{5} |x_i - \bar{x}|.$

$x_i - \bar{x}$	-86	-56	-36	14	164
$\lvert x_i - \bar{x}\rvert$	86	56	36	14	164

   $e_{moy} = \frac{1}{5} \times 356 = 71.2.$
4. $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{5}603400 - 336^2 = 7784,$
   $s_y^2 = 5760.$
5. $e_{med} = \frac{1}{5}\sum_{i=1}^{5} |x_i - x_{1/2}| = \frac{1}{5}449300 - 290^2 = 64.$
6. $m_3 = \frac{1}{5}\sum_{i=1}^{5} (x_i - \bar{x})^3 = 711072.$
7. $g_1 = \frac{m_3}{s_x^3} = \frac{711072}{\sqrt{711070}^3} = \frac{711072}{686758.8123} = 1.035402804.$
   Le coefficient est positif, il y a donc une asymétrie à droite.
   Interprétation : Cette distribution est très classique pour les revenus. Cela signifie qu'il y a beaucoup de personnes qui gagnent peu et peu de personnes qui gagnent beaucoup.

Exercice 8.10. Nombre d'enfants
On dispose des données du Tableau 8.5 sur le nombre d'enfants par couple pour 45 couples :

TABLE 8.5 – Nombre d'enfants par couple

Nombre d'enfants	0	1	2	3	4	5
Effectif	8	10	15	7	4	1

1. Quelle est la variable mesurée ? De quel type de variable s'agit-il ?
2. Calculer le nombre moyen d'enfants par couple.
3. Calculer la médiane.
4. Calculer les quartiles Q_1, Q_2 et Q_3.
5. Calculer le mode.
6. Le couple ayant 5 enfants est remplacé par un couple ayant 20 enfants. Calculer la moyenne, la médiane et le mode de la série modifiée. Que remarque-t-on ?

Solution

1. Variable : nombre d'enfants.
   Type de la variable : quantitative discrète.
2. Moyenne calculée à partir des effectifs :
$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{0\times 8 + 1\times 10 + 2\times 15 + 3\times 7 + 4\times 4 + 5\times 1}{45} = \frac{82}{45} = 1.82.$$
3. (n impair) $x_{1/2} = x_{(\frac{n+1}{2})} = x_{23} = 2$.
4. Le premier quartile : $np = \frac{1}{4}\times 45 = 11.25$ n'est pas un nombre entier, on arrondi vers le haut
5. Mode : $x_M = 2$
6. Moyenne : $\bar{x} = \dfrac{97}{45} = 2.156$.
   Médiane : (n impair) $x_{1/2} = x_{(\frac{n+1}{2})} = x_{23} = 2$.
   Mode : $x_M = 2$.
   La moyenne est très influencée par la nouvelle valeur (elle est très sensible aux valeurs extrèmes).
   La médiane et le mode ne sont pas affectés (robustes, peu sensibles aux valeurs extrèmes).

Exercice 8.11. Moyennes géométrique, harmonique ou arithmétique
Le but de cet exercice est d'utiliser adéquatement les différents types de moyennes, à savoir *arithmétique*, *géométrique*, *harmonique* et *pondérée*.

1. Vous disposez des taux d'intérêts d'un investissement durant les années 2004 à 2008. Ces données se trouvent dans le Tableau 8.6.

Table 8.6 – Taux d'intérêt d'un investissement

	Taux d'intérêt
2004	+31%
2005	+51%
2006	-1%
2007	-22%
2008	-34%

On aimerait calculer le taux moyen de cet investissement.

(a) Calculer la moyenne géométrique et la moyenne arithmétique des taux.
Aide : Vous devez calculer les moyennes de $(1+0.31), (1+0.51), \ldots, (1-0.34)$ et non pas de $31, 51, \ldots, -34$.

(b) Vérifier, par calculs, que le taux moyen est bien la moyenne géométrique des taux successifs et non pas la moyenne arithmétique des taux successifs.
Aide : Supposer que vous placez 100 francs. Regarder ce que vous obtenez après les 5 ans avec les taux d'intérêts donnés et les deux moyennes calculées.

2. Vous avez fait une randonnée de 30km. Vous avez parcouru les 10 premiers km à une vitesse de 4km/h, les 10 suivants à une vitesse de 8km/h et les 10 derniers à une vitesse de 5km/h. Vous aimeriez connaître votre vitesse moyenne.

(a) Calculer la moyenne harmonique et la moyenne arithmétique des vitesses.

(b) Vérifier, par calculs, que la vitesse moyenne est la moyenne harmonique des vitesse et non pas la moyenne arithmétique des vitesses.
Aide : Calculer la vitesse moyenne par raisonnement et comparer cette dernière avec les résultats obtenus au point précédent.

3. Vous avez acheté ce matin trois emballages de café en grains :
— 1 emballage de 1kg au prix de 12 francs le kg,
— 1 emballage de 500g (=0.5kg) au prix de 8 francs le kg et
— 1 emballage de 250g (=0.25kg) au prix de 5 francs le kg.
Quel est le prix moyen (au kg) du café que vous avez acheté ?

Solution

1. (a) Moyenne géométrique

$$G = ((1+0.31)(1+0.51)(1-0.01)(1-0.22)(1-0.34))^{1/5} = 1.0016.$$

Moyenne arithmétique

$$\bar{x} = \frac{1.31 + 1.51 + 0.99 + 0.78 + 0.66}{5} = 1.05.$$

(b) Taux d'intérêts :

$$V_T = 100(1.31)(1.51)(0.99)(0.78)(0.66) = 100.8143.$$

Moyenne géométrique des taux d'intérêts :

$$V_G = 100 \times 1.0016 \times 1.0016 \times 1.0016 \times 1.0016 \times 1.0016 = 100 \times 1.0016^5 = 100.8026$$

Moyenne arithmétique des taux d'intérêts :

$$V_{\bar{x}} = 100 \times 1.05^5 = 127.6282$$

On remarque que $V_T \cong V_G$ (le fait qu'on ne trouve pas = est dû aux arrondis.) tandis que $V_{\bar{x}}$ est très différent de V_T. On en déduit que le taux moyen est bien la moyenne géométrique des taux.

2. (a) La moyenne harmonique

$$H = \frac{3}{\frac{1}{4} + \frac{1}{8} + \frac{1}{5}} = 5.2174.$$

La moyenne arithmétique

$$\bar{x} = \frac{4 + 8 + 5}{3} = 5.\overline{6}.$$

(b) Les 10 premiers km ont été parcourus en 2h30. En effet,

$$10\,[km] = 4\left[\frac{km}{h}\right] \times 2.5\,[h]\,.$$

Les 10 suivants km ont été parcourus en 1h15. En effet,

$$10\,[km] = 8\left[\frac{km}{h}\right] \times 1.25\,[h]\,.$$

Et enfin, les 10 derniers km ont été parcourus en 2h. En effet,

$$10\,[km] = 5\left[\frac{km}{h}\right] \times 2\,[h]\,.$$

Les 30 km ont donc été parcourus en 5h45 (ou en 5.75h). La vitesse moyenne est donc :

$$V = \frac{30}{5.75} = 5.2174.$$

On a que $V = H$ et V très différent de $\bar{x}$. La vitesse moyenne est donc la moyenne harmonique des vitesses et non pas la moyenne arithmétique des vitesses.

3. On calcule ici la moyenne pondérée :

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{1 \times 12 + 0.5 \times 8 + 0.25 \times 5}{1 + 0.5 + 0.25} = \frac{12 + 4 + 1.25}{1.75} = \frac{17.25}{1.75} = 9.857.$$

Le prix moyen du café acheté est de 9.857 francs/kg.

Exercice 8.12. Primes d'assurance
Soit x_i, une série statistique représentant les primes d'assurance accident payées par quatre femmes et y_i, une série statistique représentant les primes payées par quatre hommes. Ces séries sont présentées dans le Tableau 8.7

TABLE 8.7 – Primes d'assurance accident payées par quatre femme et quatre hommes

x_i	95	105	100	100
$x_i - \bar{x}$				
y_i	70	130	85	115
$y_i - \bar{y}$				
z_i	210	390	255	345

1. Calculer la moyenne des x_i.
2. Calculer la moyenne des y_i.
3. Compléter le tableau.
4. Calculer $\sum_{i=1}^{4}(x_i - \bar{x})$.
5. Calculer l'écart moyen absolu pour la série des x_i.
6. Calculer la variance des x_i.
7. Calculer la variance des y_i.
8. Comparer les deux séries en termes de moyenne et de variance.
9. Calculer l'écart-type des deux séries.
10. L'année suivante, la prime des quatre hommes est triplée (série statistique z_i dans le tableau). Calculer la moyenne et la variance de cette série en utilisant la relation qui lie z_i à y_i et commenter.

Solution

1. $\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = 100.$
2. $\bar{y} = \frac{1}{4}\sum_{i=1}^{4} y_i = 100.$
3. Tableau complété :

x_i	95	105	100	100
y_i	70	130	85	115
z_i	210	390	255	345
$x_i - \bar{x}$	-5	5	0	0
$y_i - \bar{y}$	-30	30	-15	15

4. $\sum_{i=1}^{4}(x_i - \bar{x}) = (-5) + 5 + 0 + 0 = 0$ (toujours vrai).
5. $e_{moy} = \frac{1}{4}\sum_{i=1}^{4}|x_i - \bar{x}| = \frac{1}{4}(|-5| + 5 + 0 + 0) = 2.5.$
6. $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4}[(-5)^2 + 5^2 + 0^2 + 0^2] = 12.5$
7. $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{4}\left\{(-30)^2 + 30^2 + (-15)^2 + 15^2\right\} = 562.5$

8. Les deux variables ont la même moyenne. Cependant, la variance des y_i est beaucoup plus grande. La série est plus dispersée autour de la moyenne.
9. $s_x = \sqrt{s_x^2} = \sqrt{12.5} = 3.536$ et $s_y = \sqrt{s_y^2} = \sqrt{562.5} = 23.717-$
10. L'année suivante, la prime des quatre hommes est triplée (série statistique z_i dans le tableau).

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} 3y_i = 3 \times \frac{1}{n}\sum_{i=1}^{n} y_i = 3\bar{y} = 300,$$

$$s_z^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2 - \bar{z}^2 = \frac{1}{n}\sum_{i=1}^{n}(3y_i)^2 - (3\bar{y})^2 = 9\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2\right) = 9s_y^2 = 5062.5.$$

Exercice 8.13. Distribution de salaires
Au sein d'une entreprise, la distribution des salaires est donnée par le Tableau 8.8.

TABLE 8.8 – Distribution des salaires

Salaire x_j	Effectif n_j
4000	12
5000	5
6000	2
10 000	1
	20

1. Donner :
   (a) Le mode.
   (b) Le salaire moyen.
   (c) La médiane.
   (d) Le quantile d'ordre 1/2.
   (e) Le premier et le troisième quartile.
   (f) L'étendue.
   (g) La distance interquartile.
   (h) La variance et l'écart-type.
2. Sachant que le moment centré d'ordre 3, $m_3 = 7181250000$, calculer les coefficients d'asymétrie de Fisher, Yule et Pearson.
3. La distribution est-elle symétrique, allongée à gauche ou allongée à droite ? Ce résultat est-il surprenant ?
4. Construire le diagramme en boîte (boxplot).
5. Eliminer l'observation 10 000 puis calculer
   (a) La moyenne.
   (b) La médiane, le premier et le dernier quartile.

Solution

1. (a) $x_M = 4000$.
   (b) $\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{12 \times 4000 + 5 \times 5000 + 2 \times 6000 + 1 \times 10\,000}{20} = 4750.$
   (c) $n = 20$ est pair. Donc, $x_{1/2} = \frac{1}{2}\left[x_{(10)} + x_{(11)}\right] = \frac{1}{2}[4000 + 4000] = 4000.$
   (d) $np = 20 \times \frac{1}{2} = 10$ qui est un nombre entier. Donc, $x_{1/2} = \frac{1}{2}\left[x_{(10)} + x_{(11)}\right] = 4000.$
   La médiane n'est autre que le quantile d'ordre 1/2.
   (e) Premier quartile : $np = 20 \times \frac{1}{4} = 5$, ce qui est un nombre entier. Donc,
   $$x_{1/4} = \frac{1}{2}\left[x_{(5)} + x_{(6)}\right] = \frac{1}{2}[4000 + 4000] = 4000.$$
   Troisième quartile : $np = 20 \times \frac{3}{4} = 15$, ce qui est un nombre entier. Donc,
   $$x_{3/4} = \frac{1}{2}\left[x_{(15)} + x_{(16)}\right] = \frac{1}{2}[5000 + 5000] = 5000.$$
   (f) $\mathrm{E} = x_{(20)} - x_{(1)} = 10\,000 - 4000 = 6000.$
   (g) $IQ = x_{3/4} - x_{1/4} = 5000 - 4000 = 1000.$

(h) $s_x^2 = \dfrac{1}{n}\displaystyle\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2$

$= \dfrac{1}{20}(12 \times 4000^2 + 5 \times 5000^2 + 2 \times 6000^2 + 1 \times 10\,000^2) - 4750^2 = 1887500.$

$s_x = \sqrt{s_x^2} = \sqrt{1887500} = 1373.863.$

2. $g_1 = \dfrac{m_3}{s_x^3} = \dfrac{7181250000}{1373.863^3} = 2.769.$

$A_Y = \dfrac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}} = \dfrac{5000 + 4000 - 2 \times 4000}{5000 - 4000} = 1.$

$A_p = \dfrac{\bar{x} - x_M}{s_x} = \dfrac{4750 - 4000}{1373.863} = 0.546,$

où x_M est le mode de la distribution, lorsqu'il est défini.

3. La distribution est allongée à droite. Tous les coefficients d'asymétrie le laissent penser parce qu'ils sont positifs. Ce résultat n'est pas surprenant : en effet, une seule personne gagne 10 000francs tandis que plus de la moitié (en fait 12/20) des employés gagnent moins que la moyenne.

4. Boxplot :

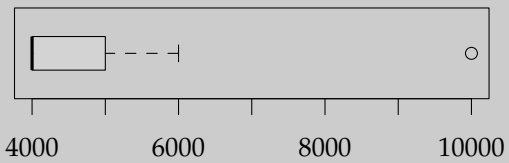
$b^- = x_{1/4} - 1.5IQ = 4000 - 1.5 \times 1000 = 2500 \Rightarrow$ valeur adjacente : 4000

$b^+ = x_{3/4} + 1.5IQ = 5000 + 1.5 \times 1000 = 6500 \Rightarrow$ valeur adjacente : 6000

Il y a une valeur extreme (10 000).

En langage R

```
library(tikzDevice)
X=c(rep(4000,12),rep(5000,5),rep(6000,2),rep(10\,000,1))
# boxplot
tikz("boxplotentr.tex",width=3.5,height=2)
boxplot(X, horizontal = TRUE)
dev.off()
```



5. On a alors

(a) $\bar{x} = \dfrac{12 \times 4000 + 5 \times 5000 + 2 \times 6000}{19} = 4473.684.$

(b) Médiane :

$$np = 19 \times \frac{1}{2} = 9.5,$$

qui n'est pas un entier. Donc,

$$x_{1/2} = x_{(\lceil np \rceil)} = x_{(\lceil 9.5 \rceil)} = x_{(10)} = 4000.$$

Premier quartile :

$$np = 19 \times \frac{1}{4} = 4.75,$$

qui n'est pas un entier. Donc,

$$x_{1/4} = x_{(\lceil np \rceil)} = x_{(\lceil 4.75 \rceil)} = x_{(5)} = 4000.$$

Dernier quartile :

$$np = 19 \times \frac{3}{4} = 14.25,$$

qui n'est pas un entier. Donc,

$$x_{3/4} = x_{(\lceil np \rceil)} = x_{(\lceil 14.25 \rceil)} = x_{(15)} = 5000.$$

On remarque que la moyenne a changé alors que les quantiles d'ordre 1/4,1/2 et 3/4 n'ont pas changé.

Exercice 8.14. Boxplot
Le Tableau 8.9 contient la série d'une variable discrète.

TABLE 8.9 – Série d'une variable discrète

1	2	3	4	4	7	11	14	15	17	23	57	57	57	63	72	101	107	200

1. Calculer la médiane, le premier quartile et le troisième quartile.
2. Construire un boxplot (boîte à moustache) en indiquant les calculs nécessaires à sa construction.

Solution

1. Calculons la médiane : $np = 19 \times 0.5 = 9.5$. Comme np n'est pas un entier, on arrondit à l'entier supérieur : $x_{1/2} = x_{(10)} = 17$.
   Calculons le premier quartile : $np = 19 \times 0.25 = 4.75$. Comme np n'est pas un entier, on arrondit à l'entier supérieur : $x_{1/4} = x_{(5)} = 4$.
   Calculons le troisième quartile : $np = 19 \times 0.75 = 14.25$. Comme np n'est pas un entier, on arrondit à l'entier supérieur : $x_{3/4} = x_{(15)} = 63$.
2. Calculons la distance interquartile, les bornes ainsi que les valeurs extrêmes : IQ= $63 - 4 = 59$.
   $b^- = x_{1/4} - 1.5\text{IQ}= 4 - 1.5 \times 59 = -84.5$ La valeur adjacente est : $x_{(5)} = 1$.
   $b^+ = x_{3/4} - 1.5\text{IQ}= 63 + 1.5 \times 59 = 151.5$. La valeur adjacente est : $x_{(18)} = 107$.
   Il nous reste une valeur extrême : $x_{(19)} = 200$. Nous pouvons maintenant contruire une boîte à moustache.

   En langage R

```
library(tikzDevice)
X=c(1,2,3,4,4,7,11,14,15,17,23,57,57,57,63,72,101,107,200)
# boxplot
tikz("boxplotdisc.tex",width=3.5,height=2)
boxplot(X, horizontal = TRUE)
dev.off()
```

Exercice 8.15. Variances
Montrer que la variance peut être exprimée à partir des effectifs et des valeurs distinctes de la manière suivante :

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2.$$

Aide : On remarque que

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2 = \frac{1}{n}\sum_{j=1}^{J} n_j\left(x_j - \bar{x}\right)^2.$$

Solution

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2 = \frac{1}{n}\sum_{j=1}^{J} n_j\left(x_j - \bar{x}\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{J} n_j\left(x_j^2 - 2x_j\bar{x} + \bar{x}^2\right) = \frac{1}{n}\sum_{j=1}^{J}\left(n_j x_j^2 - 2n_j x_j\bar{x} + n_j\bar{x}^2\right) \\
&= \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \frac{2\bar{x}}{n}\sum_{j=1}^{J} n_j x_j + \frac{\bar{x}^2}{n}\sum_{j=1}^{J} n_j = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - 2\bar{x}\frac{1}{n}\sum_{j=1}^{J} n_j x_j + \frac{\bar{x}^2}{n}n \\
&= \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2
\end{aligned}$$

Exercice 8.16. Âges dans les familles
Soit une série statistique x_i : $2, 5, 32, 34, 6, 8, 11, 41, 54$ représentant les âges des individus de deux familles, la famille A (les 4 premières valeurs de la série) et la famille B (les cinq dernières valeurs). Calculez :

1. la moyenne d'âge de la famille A,
2. la moyenne d'âge de la famille B,
3. la moyenne générale,
4. la variance de la famille A,
5. la variance de la famille B,
6. la variance totale à l'aide du théorème des probabilités totales (variance totale= variance intra-groupes + variance inter-groupes).

Solution

1. Moyenne d'âge de la famille A :

$$\bar{x}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} x_i = \frac{2+5+32+34}{4} = 18.25.$$

2. Moyenne d'âge de la famille B :

$$\bar{x}_B = \frac{1}{n_B}\sum_{i=n_A+1}^{n} x_i = \frac{6+8+11+41+54}{5} = 24.$$

3. Moyenne générale :

$$\bar{x} = \frac{1}{n}(n_A\bar{x}_A + n_B\bar{x}_B) = \frac{1}{9}(4\times 18.25 + 5\times 24) = 21.44.$$

4. Variance de la famille A :

$$s_A^2 = \frac{1}{n_A}\sum_{i=1}^{n_A} x_i^2 - \bar{x}_A^2 = \frac{1}{4}2209 - 18.25^2 = 219.19.$$

5. Variance de la famille B :

$$s_B^2 = \frac{1}{n_B}\sum_{i=n_A+1}^{n_B} x_i^2 - \bar{x}_B^2 = \frac{1}{5}4818 - 24^2 = 387.60.$$

6. Variance totale (intra + inter) :

$$\begin{aligned} s_x^2 &= \frac{n_A s_A^2 + n_B s_B^2}{n} + \frac{n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2}{n} \\ &= \frac{4\times 219.19 + 5\times 387.60}{9} + \frac{4(18.25-21.44)^2 + 5(24-21.44)^2}{9} = 312.75 + 8.16 = 320.91. \end{aligned}$$

Exercice 8.17. Spectateurs dans les stades
Le Tableau 8.10 donne le nombre de spectateurs que peuvent accueillir les stades des grandes équipes anglaises de football ainsi que le stade national de Wembley :

Table 8.10 – Principaux stades anglais

Equipe	Places	Equipe	Places
Southampton	15000	Derby County	34000
Charlton Athletic	15222	Middlesbrough	35000
Bradford City FC	18018	Tottenham Hotspur	36214
Barnsley FC	19073	Newcastle UTD	36610
Watford FC	22011	Arsenal	38500
Leicester City	22517	Aston villa	39339
Coventry City	23500	Sheffield Wednesday	39814
Bolton Wanderers	25000	Everton	40200
West Ham UTD	26014	Leeds UTD	40204
Crystal palace	26309	Sunderland	41600
Wimbledon	26309	Liverpool	45000
Nottingham Forest	30602	Manchester UTD	55400
Blackburn Rovers	31367	Wembley	78500
Chelsea	31791		

1. Calculez les premier, deuxième et troisième quartiles et ensuite la distance interquartile.
2. Construire le diagramme en tiges et feuilles.
3. Construisez le diagramme en boîte (boxplot).

Solution

1. Le premier quartile : $np = \frac{1}{4} \times 27 = 6.75$ n'est pas un nombre entier, on arrondit à l'entier supérieur
$$x_{1/4} = x_{(7)} = 23500.$$
La médiane (deuxième quartile) : Comme $np = \frac{1}{2} \times 27 = 13.5$ n'est pas un nombre entier, on arrondit à l'entier supérieur
$$x_{1/2} = x_{(14)} = 31791.$$
Le troisième quartile : Comme $np = \frac{3}{4} \times 27 = 20.25$ n'est pas un nombre entier, on arrondit à l'entier supérieur
$$x_{3/4} = x_{(21)} = 39814.$$
IQR :
$$IQR = x_{3/4} - x_{1/4} = 39814 - 23500 = 16314.$$
2. $b^- = x_{1/4} - 1.5IQ = 23500 - 1.5 \times 16314 = -971 \Rightarrow$ valeur adjacente : 15000 (Southampton)
$b^+ = x_{3/4} + 1.5IQ = 39814 + 1.5 \times 16314 = 64285 \Rightarrow$ valeur adjacente : 55400 (Man. UTD)
+ une valeur extreme (78500).

En langage R

```
library(tikzDevice)
X=c(15000,34000,15222,35000,18018,36214,19073,36610,22011,38500,22517,
39339,23500,39814,25000,40200,26014,40204,26309,41600,26309,45000,
30602,55400,31367,78500,31791)
# boxplot
tikz("boxplotstatdes.tex",width=3.5,height=2)
boxplot(X, horizontal = TRUE)
dev.off()
```

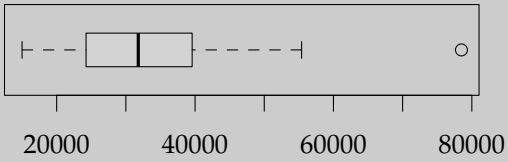


Diagramme tige-feuille :

```
1 | 5589
2 | 2345666
3 | 112456799
4 | 00025
5 | 5
6 |
7 | 9
```

Exercice 8.18. Paramètres dans une distribution
Calculez tous les paramètres (de position, de dispersion et de forme) à partir du Tableau 8.11. Il ne faut pas regrouper les données en classes.

Table 8.11 – Valeurs prises par une variable quantitative discrète

152	152	152	153	153	154	154	154	155	155
156	156	156	156	156	157	157	157	158	158
159	159	160	160	160	161	160	160	161	162
162	162	163	164	164	164	164	165	166	167
168	168	168	169	169	170	171	171	171	171

Solution

— Médiane : Comme n est pair,

$$x_{1/2} = \frac{1}{2}(x_{25} + x_{26}) = \frac{1}{2}(160 + 160) = 160.$$

— Quantiles
  — Premier quartile :

$$x_{1/4} = x_{13} = 156$$

  — Deuxième quartile :

$$x_{3/4} = x_{38} = 165$$

— Étendue :

$$E = 171 - 152 = 19.$$

— Distance interquartile :

$$IQ = x_{3/4} - x_{1/4} = 165 - 156 = 9.$$

— Variance :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{50} \times 1668 = 33.36.$$

— Écart type :

$$s_x = \sqrt{s_x^2} = 5.7758.$$

— Écart moyen absolu :

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{50} \times 245.2 = 4.904.$$

— Écart médian absolu :

$$e_{med} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_{1/2}| = \frac{1}{50} \times 242 = 4.84.$$

— Moment centré d'ordre trois :

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{50} \times 2743.2 = 54.864.$$

— Coefficient d'asymétrie

$$g_1 = \frac{m_3}{s_x^3} = \frac{54.864}{5.7758^3} = 0.28474.$$

— Moment centré d'ordre quatre :

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 = \frac{1}{50} \times 108032.2 = 2160.643.$$

— Coefficients d'applatissement

$$\beta_2 = \frac{m_4}{s_x^4} = \frac{2160.643}{5.7758^4} = 1.941471,$$

et

$$g_2 = \beta_2 - 3 = -1.058529.$$

Exercice 8.19. Séries et calcul de paramètres
Considérons les séries ordonnées données dans les tableaux 8.12 et 8.13.

Table 8.12 – Quotient intellectuel de 100 enfants

75	75	76	76	76	77	77	79	80	80	80	82	82	82	82	82	83	83	84	84
84	85	85	85	85	86	86	86	87	87	88	88	88	89	89	89	90	90	90	90
90	90	90	90	90	90	91	91	91	91	91	91	91	92	92	92	92	92	92	92
93	93	93	93	94	94	96	96	96	96	96	96	97	98	98	98	98	98	98	98
99	99	100	100	100	100	102	103	104	104	104	105	105	105	106	106	107	110	110	110

Table 8.13 – Nombre de jours de chômage pour 40 personnes

180	10	30	50	420	30	180	360	200	30	360	120	500	200	30	420	360	370	360	150
180	280	30	500	180	720	420	180	40	500	120	180	194	400	30	360	40	400	180	200

Le Tableau 8.14 contient les tableaux statistiques de ces séries. Pour chaque série :

Table 8.14 – Tableaux statistiques des variables x et y

quotient			chômage		
x_j	n_j	N_j	y_j	n_j	N_j
75	2	2	10	1	1
76	3	5	30	6	7
77	2	7	40	2	9
79	1	8	50	1	10
80	3	11	120	2	12
82	5	16	150	1	13
83	2	18	180	7	20
84	3	21	194	1	21
85	4	25	200	3	24
86	3	28	280	1	25
87	2	30	360	5	30
88	3	33	370	1	31
89	3	36	400	2	33
90	10	46	420	3	36
91	7	53	500	3	39
92	7	60	720	1	40
93	4	64			
94	2	66			
96	6	72			
97	1	73			
98	7	80			
99	2	82			
100	4	86			
102	1	87			
103	1	88			
104	3	91			
105	3	94			
106	2	96			
107	1	97			
110	3	100			

1. Calculer la moyenne et le mode.
2. Calculer les quantiles d'ordre $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ et la distance interquartile.

3. Calculer la variance et l'écart-type.
4. En sachant que les moments centrés d'ordre 3 et 4 sont : $m_3 = 68.01$ et $m_4 = 12984.07$ pour la première série et $m_3 = 2816704$ et $m_4 = 2331922331$ pour la deuxième série, calculer les coefficients d'asymétrie de Fisher, Yule et Pearson. Calculer les coefficients d'applatissement de Pearson et Fisher. Quelle est la conclusion sur la distribution des deux séries ?
5. Construire le diagramme en tiges et feuilles.
6. Construire le diagramme en boîte.

Solution

Quotient intellectuel

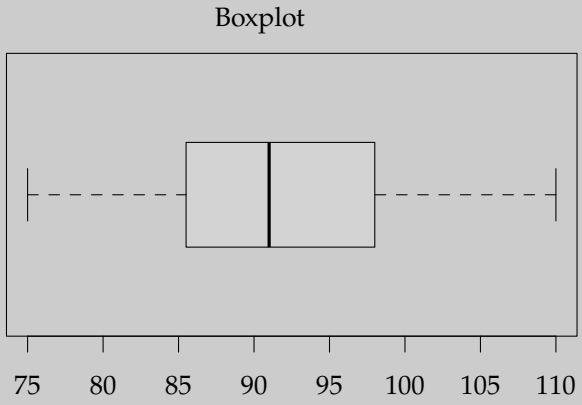
1. Moyenne $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{100} \times 9161 = 91.61$
   Mode : $x_M = 90$.
2. Le premier quartile : Comme $np = \frac{1}{4} \times 100 = 25$ est un nombre entier, on a
$$x_{1/4} = \frac{1}{2}\left\{x_{(25)} + x_{(26)}\right\} = (85 + 86)/2 = 85.5.$$
   La médiane : Comme $np = \frac{1}{2} \times 100 = 50$ est un nombre entier, on a
$$x_{1/2} = \frac{1}{2}\left\{x_{(50)} + x_{(51)}\right\} = (91 + 91)/2 = 91.$$
   Le troisième quartile : Comme $np = \frac{3}{4} \times 100 = 75$ est un nombre entier, on a
$$x_{3/4} = \frac{1}{2}\left\{x_{(75)} + x_{(76)}\right\} = (98 + 98)/2 = 98.$$
   $IQR = x_{3/4} - x_{1/4} = 98 - 85.5 = 12.5$
3. La variance :
$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{j=1}^{J} x_j^2 n_j - \bar{x}^2 = \frac{1}{100}\sum_{j=1}^{30} x_j^2 n_j - \bar{x}^2 \\ &= \frac{1}{100} \times [75^2 \times 2 + 76^2 \times 3 + 77^2 \times 2 + \cdots + 110^2 \times 3] - 91.61^2 = 71.9979. \end{aligned}$$
   Écart-type :
$$s_x = \sqrt{71.9979} = 8.485.$$
4. Asymétrie :
   Coefficient de Fisher :
$$g_1 = \frac{m_3}{s_x^3} = \frac{68.01}{610.88} = 0.11.$$
   Coefficient de Yule :
$$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}} = \frac{98 + 85.5 - 2 \times 91}{12.5} = 0.12.$$
   Coefficient de Pearson :
$$A_P = \frac{\bar{x} - x_M}{s_x} = \frac{91.61 - 90}{8.485} = 0.19.$$
   Aplatissement :
   Coefficient de Pearson :
$$\beta_2 = \frac{m_4}{s_x^4} = \frac{12984.07}{5183.31} = 2.50.$$
   Coefficient de Fisher :
$$g_2 = \beta_2 - 3 = 2.50 - 3 = -0.50.$$

5. **En langage R**

```
library(tikzDevice)
# calcul des parametres de la série quotient intellectuel
a=c(75,75,76,76,76,77,77,79,80,80,80,82,82,82,82,82,83,83,84,84,84,85,85,85,85,
86,86,86,87,87,88,88,88,89,89,89,90,90,90,90,90,90,90,90,90,90,91,91,91,91,
91,91,91,92,92,92,92,92,92,92,93,93,93,93,94,94,96,96,96,96,96,96,97,98,98,
98,98,98,98,98,99,99,100,100,100,100,102,103,104,104,104,105,105,105,106,106,
107,110,110,110)
# la longueur de la série
n=length(a)
n
# tableau statistique
t=table(a)
v=c(t)
# tableau statistique complet
data.frame(Eff=v,EffCum=cumsum(v))
# la moyenne
m=mean(a)
m
# les quartiles
quantile(a,type=2)
# la distance interquartile=12.5
# la variance
s2=sum((a-m)^2)/n
s2
# l'écart-type
s=sqrt(s2)
s
# le moment centré d'ordre 3
m3=sum((a-m)^3)/n
m3
# les coefficients d'asymétrie
# de Fisher
g1=m3/(s^3)
g1
# de Yule
y=(98+85.5-2*91)/12.5
y
# de Pearson
p=(m-90)/s
p
# paramètres d'applatissement
# le moment centré d'ordre 4
m4=sum((a-m)^4)/n
m4
# Pearson
beta2=m4/(s2^2)
beta2
# Fisher
g2=beta2-3
g2
# stem and leaf plot
stem(a)
# boxplot
tikz("boxplotquo.tex",width=4,height=3)
boxplot(a, horizontal = TRUE)
dev.off()
```

Diagramme en tige et feuilles :

```
7 | 55666779
8 | 0002222233444
8 | 555566677888999
9 | 00000000001111111222222233344
9 | 6666667888888899
10 | 000023444
10 | 555667
11 | 000
```



Variable : Chômage

1. Moyenne $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{40} \times 9494 = 237.35$.
   Mode : $x_M = 180$.
2. Le premier quartile : Comme $np = \frac{1}{4} \times 40 = 10$ est un nombre entier, on a
   $$x_{1/4} = \frac{1}{2}\left\{x_{(10)} + x_{(11)}\right\} = (50 + 120)/2 = 85.$$
   La médiane : Comme $np = \frac{1}{2} \times 40 = 20$ est un nombre entier, on a
   $$x_{1/2} = \frac{1}{2}\left\{x_{(20)} + x_{(21)}\right\} = (180 + 194)/2 = 187.$$
   Le troisième quartile : Comme $np = \frac{3}{4} \times 40 = 30$ est un nombre entier, on a
   $$x_{3/4} = \frac{1}{2}\left\{x_{(30)} + x_{(31)}\right\} = (360 + 370)/2 = 365.$$
   $IQR = x_{3/4} - x_{1/4} = 365 - 85 = 280$.
3. Variance
   $$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{j=1}^{J} x_j^2 n_j - \bar{x}^2 = \frac{1}{40}\sum_{j=1}^{16} x_j^2 n_j - \bar{x}^2 \\ &= \frac{1}{100} \times [10^2 \times 1 + 30^2 \times 6 + 40^2 \times 2 + \cdots + 720^2 \times 1] - 237.35^2 = 29360.88. \end{aligned}$$
   Écart-type : $s_x = \sqrt{29360.88} = 171.35$.
4. Asymétrie : Coefficient de Fisher :
   $$g_1 = \frac{m_3}{s_x^3} = \frac{2816704}{171.35^3} = 0.56.$$
   Coefficient de Yule :
   $$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}} = \frac{365 + 85 - 2 \times 187}{280} = 0.27.$$

Coefficient de Pearson :

$$A_P = \frac{\bar{x} - x_M}{s_x} = \frac{237.35 - 180}{171.35} = 0.33.$$

Aplatissement : Coefficient de Pearson :

$$\beta_2 = \frac{m_4}{s_x^4} = \frac{2331922331}{171.35^4} = 2.71.$$

Coefficient de Fisher :

$$g_2 = \beta_2 - 3 = 2.71 - 3 = -0.29.$$

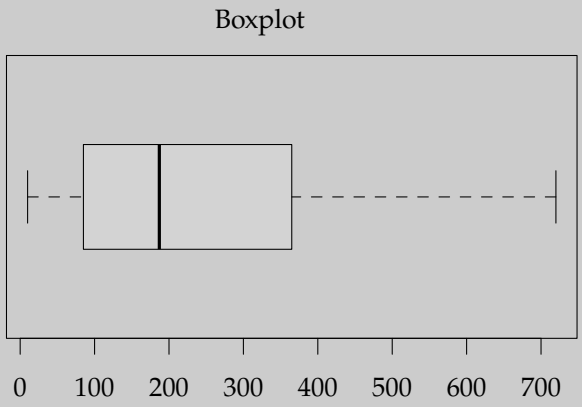
5. **En langage R**

```
# calcul des paramètres de la série nombre de jours de chômage
a=c(180,10,30,50,420,30,180,360,200,30,360,120,500,200,30,420,360,370,360,150,
180,280,30,500,180,720,420,180,40,500,120,180,194,400,30,360,40,400,180,200)
# longueur de la série
n=length(a)
n
# tableau statistique
t=table(a)
v=c(t)
# tableau statistique complet
data.frame(Eff=v,EffCum=cumsum(v))
# moyenne
m=mean(a)
m
# les quartiles
quantile(a,type=2)
# la distance interquartile=280
# la variance
s2=sum((a-m)^2)/n
s2
# l'écart-type
s=sqrt(s2)
s
# le moment centré d'ordre 3
m3=sum((a-m)^3)/n
m3
# les coefficients d'asymétrie
# de Fisher
g1=m3/(s^3)
g1
# de Yule
y=(365+85-2*187)/280
y
# de Pearson
p=(m-180)/s
p
# paramètres d'applatissement
# le moment centré d'ordre 4
m4=sum((a-m)^4)/n
m4
# Pearson
beta2=m4/(s2^2)
beta2
# Fisher
g2=beta2-3
g2
```

```
# stem and leaf plot
stem(a)
# boxplot
tikz("boxplotchom.tex",width=4,height=3)
boxplot(a, horizontal = TRUE)
dev.off()
```

Diagramme en tige et feuilles :

```
0 | 1333333445
1 | 22588888889
2 | 0008
3 | 666667
4 | 00222
5 | 000
6 |
7 | 2
```



Chapitre 9

Exercices : Statistique descriptive bivariée

Exercice 9.1. Années d'études et revenus
Considérons un échantillon de 10 employés du même âge, d'une entreprise. Soit X le nombre d'années d'études et Y le revenu mensuel (en milliers de francs) touché par chacun d'entre eux. Les observations sont contenues dans le Tableau 9.1.

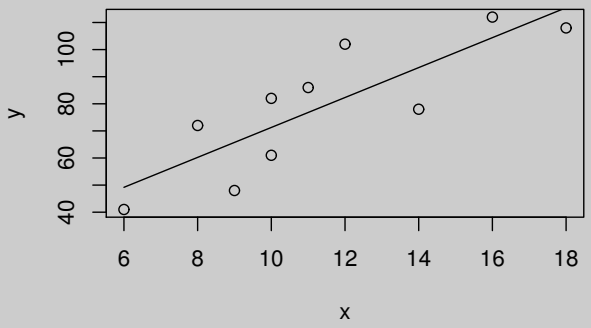
Table 9.1 – Revenu mensuel et nombre d'années

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	6	41			
	8	72			
	9	48			
	10	82			
	10	61			
	11	86			
	12	102			
	14	78			
	16	112			
	18	108			
Somme					
Moyenne					

1. Représenter cette série par un nuage de points.
2. Compléter les cellules vides du tableau des données.
3. Calculer les moyennes, les variances marginales et la covariance entre X et Y.
4. Déterminer le coefficient de corrélation entre les variables X et Y.
5. Déterminer l'équation de la droite de régression de Y en fonction de X.
6. Déterminer la qualité de cet ajustement.
7. Donner la valeur ajustée et le résidu pour la première observation du Tableau 9.1.
8. Établisser, sur base du modèle, le revenu pour un employé ayant fait 13 ans d'études.

Solution

1. Nuage de points :



2. Tableau :

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	6	41	36	1681	246
	8	72	64	5184	576
	9	48	81	2304	432
	10	82	100	6724	820
	10	61	100	3721	610
	11	86	121	7396	946
	12	102	144	10404	1224
	14	78	196	6084	1092
	16	112	256	12544	1792
	18	108	324	11664	1944
Somme	114	790	1422	67706	9682
Moyenne	11.4	79	142.2	6770.6	968.2

3. Paramètres :

$$\bar{x} = 11.4,\ \bar{y} = 79,$$
$$s_x^2 = 142.2 - 11.4^2 = 142.2 - 129.96 = 12.24,$$
$$s_y^2 = 6770.6 - 79^2 = 6770.6 - 6241 = 529.6,$$
$$s_{xy} = 968.2 - 11.4 \times 79 = 968.2 - 900.6 = 67.6.$$

4.
$$r_{xy} = \frac{67.6}{\sqrt{12.24 \times 529.6}} = \frac{67.6}{80.51} = 0.839.$$

5. Ajustement linéaire de y en x

$$D_{y|x}\ : y = a + b\ x,\ \text{avec } b = \frac{s_{xy}}{s_x^2} = 5.52 \text{ et } a = \bar{y} - \frac{s_{xy}}{s_x^2} x = 79 - 5.52 \times 11.4 = 79 - 62.93 = 16.07.$$

$$D_{y|x}\ : y = 5.52x + 16.07.$$

6. $r^2 = 0.839^2 = 0.704 \Rightarrow$ très bon ajustement.
7. $y_1^* = 5.52 \times 6 + 16.07 = 49.19,\ e_1 = 41 - 49.19 = -8.19.$
8. $y = 5.52 \times 13 + 16.07 = 87.83.$

Exercice 9.2. Ancienneté et absence
Le Tableau 9.2 contient un échantillon de 10 fonctionnaires (ayant entre 40 et 50 ans) d'un ministère. Soit X le nombre d'années de service et Y le nombre de jours d'absence pour raison de maladie (au cours de l'année précédente) déterminé pour chaque personne appartenant à cet échantillon.

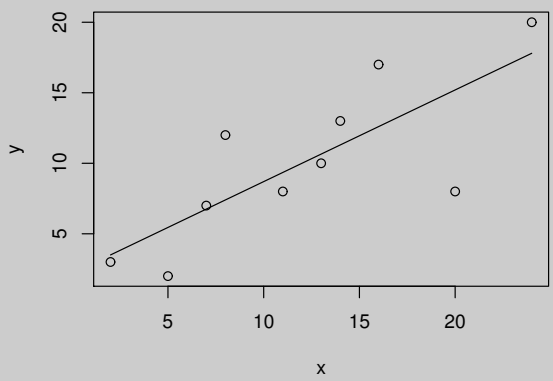
TABLE 9.2 – Ancienneté et absence

x_i	2	14	16	8	13	20	24	7	5	11
y_i	3	13	17	12	10	8	20	7	2	8

1. Représenter le nuage de points.
2. Calculer les coefficients de corrélation et de détermination entre X et Y.
3. Déterminer l'équation de la droite de régression de Y en fonction de X.
4. Déterminer la qualité de cet ajustement.
5. Établisser, sur base de ce modèle, le nombre de jours d'absence pour un fonctionnaire ayant 22 ans de service.
6. Démonstration de l'exercice en R.

Solution

1. Nuage des points et droite de régression



2. Tableau statistique :

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	2	3	4	9	6
	14	13	196	169	182
	16	17	256	289	272
	8	12	64	144	96
	13	10	169	100	130
	20	8	400	64	160
	24	20	576	400	480
	7	7	49	49	49
	5	2	25	4	10
	11	8	121	64	88
somme	120	100	1860	1292	1473
moyenne	12.00	10.00	186.00	129.20	147.30

Calcul des paramètres :

$$\sum_{i=l}^{n} x_i = 120 \Rightarrow \bar{x} = 120/10 = 12, \sum_{i=l}^{n} x_i = 100 \Rightarrow \bar{y} = 100/10 = 10,$$

$$\sum_{i=l}^{n} x_i^2 = 1860, \sum_{i=l}^{n} y_i^2 = 1292, \sum_{i=1}^{n} x_i y_i = 1473,$$

$$s_x^2 = (1860/10) - 12^2 = 42, s_y^2 = (1292/10) - 10^2 = 29.2,$$

$$s_{xy} = (1473/10) - (10.12) = 27.3,$$

$$r_{xy} = \frac{27.3}{\sqrt{42 \times 29.2}} = 0.78.$$

3. Droite de régression :

$$D_{y|x} \equiv y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) = y - 10 = \frac{27.3}{42}(x - 12).$$

Donc,

$$y = 0.65x + 2.2.$$

4. Qualité de l'ajustement :

$$r^2 = 0.608,$$

$$s_e^2 = s_y^2(1 - r^2) = 29.2 \times (1 - 0.608) = 11.43.$$

$$s_e^2 = 11.43 \text{ est beaucoup plus petit que } s_y^2 = 29.2 \Rightarrow \text{L'ajustement est bon.}$$

5. Valeur prédite :

$$y^* = 0.65 \times 22 + 2.2 = 16.5 \text{ jours.}$$

Exercice 9.3. Crèmes glacées
La consommation de crèmes glacées par individus a été mesurée pendant 30 périodes. L'objectif est de déterminer si la consommation dépend de la température (en Farenheit). Les données sont dans le Tableau 9.3. On

TABLE 9.3 – Consommation de crèmes glacées

consommation y	température x	consommation y	température x
386	41	381	63
374	56	470	72
393	63	443	72
425	68	386	67
406	69	342	60
344	65	319	44
327	61	307	40
288	47	284	32
269	32	326	27
256	24	309	28
286	28	359	33
298	26	376	41
329	32	416	52
318	40	437	64
381	55	548	71

sait en outre que

$$\sum_{i=i}^{n} y_i = 10783, \quad \sum_{i=i}^{n} x_i = 1473, \quad \sum_{i=i}^{n} y_i^2 = 4001293, \quad \sum_{i=i}^{n} x_i^2 = 80145, \quad \sum_{i=i}^{n} x_i y_i = 553747.$$

1. Donner les moyennes marginales, les variances marginales et la covariance entre les deux variables.
2. Calculer les coefficients de corrélation et de détermination entre X et Y.
3. Donner la droite de régression, avec comme variable dépendante la consommation de glaces et comme variable explicative la température.
4. Déterminer la qualité de cet ajustement.
5. Donner la valeur ajustée et le résidu pour la première observation du Tableau 9.3.
6. Démonstration de l'exercice en R.

Solution

1. Moyennes, variances et covariance :

$$\bar{y} = 10783/30 = 359.433, \ \bar{x} = 1473/30 = 49.1,$$
$$s_y^2 = 4001293/30 - 359.433^2 = 4184.112, \ s_x^2 = 80145/30 - 49.1^2 = 260.69,$$
$$s_{xy} = 553747/30 - 49.1 \times 359.433 = 810.057.$$

2. Corrélation :

$$r = \frac{810.057}{\sqrt{49.1}\sqrt{359.433}} = 0.776, \ r^2 = 0.602.$$

3. Droite de régression :

$$D_{y|x} \equiv y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}),$$
$$D_{y|x} \equiv y - 359.433 = \frac{810.057}{260.69}(x - 49.1),$$
$$D_{y|x} \equiv y = 3.107x + 206.862,$$
$$a = 206.862, \ b = 3.107.$$

4. $r^2 = 0.602 \Rightarrow$ bon ajustement.
5. $y_1^* = 334.264, \ e_1 = 51.736.$

Exercice 9.4. Poids et tailles
Considérons le poids (y_i) et la taille en cm (x_i) de 10 étudiants présentés dans le Tableau 9.4.

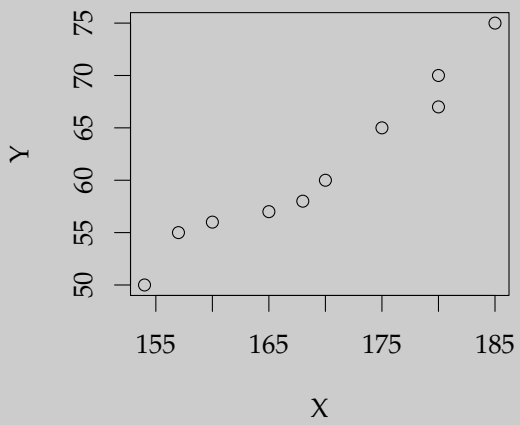
TABLE 9.4 – Poids et tailles

y_i	50	55	56	57	58	60	65	67	70	75
x_i	154	157	160	165	168	170	175	180	180	185

1. Dessiner le nuage de points
2. Calculer les variances marginales
3. Calculer la covariance
4. Calculer le coefficient de corrélation
5. Déterminer l'équation de la droite de régression de y en fonction de x
6. Dessiner la droite de régression
7. Calculer le coefficient de détermination

Solution

1. Dessiner le nuage de points :

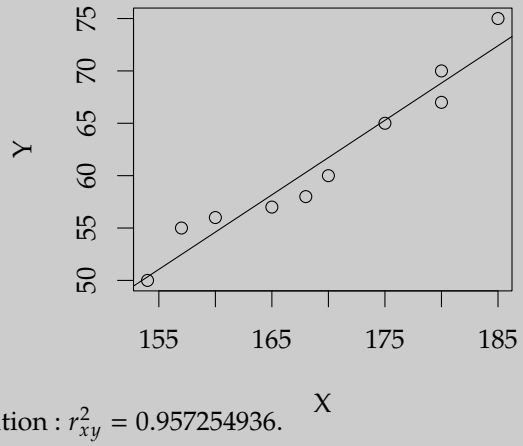


En langage R

```
library(tikzDevice)
Y=c(50,55,56,57,58,60,65,67,70,75)
X=c(154,157,160,165,168,170,175,180,180,185)
tikz(file="etudplot.tex",width=3,height=3)
plot(X,Y)
dev.off()
tikz(file="etudplotdr.tex",width=3,height=3)
plot(X,Y)
abline(coefficients(lm(Y~X)))
dev.off()
```

2. Variances marginales
Comme $\bar{x} = 169.40$ et $\bar{y} = 61.50$,
$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{10}287964 - 169.40^2 = 100.04.$$
$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{10}38113 - 61.50^2 = 53.61.$$
3. Covariance
$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} = \frac{1}{5}104554 - 169.40 \times 61.30 = 71.18.$$

4. Coefficient de corrélation
   $r_{xy} = \dfrac{s_{xy}^2}{s_x s_y} = \dfrac{71.18}{10.00 \times 7.32} = 0.97.$
5. Équation de la droite de régression
   $y = a + b\ x.$
   $b = \dfrac{s_{xy}}{s_x^2} = 0.71.$
   $a = \bar{y} - b\ \bar{x} = 61.30 - 0.71 \times 169.40 = -59.23.$
   $y = -59.23 + 0.71x$
6. Droite de régression



7. Coefficient de détermination : $r_{xy}^2 = 0.957254936.$

Exercice 9.5. Cours du dollar
Dans le Tableau 9.5, vous trouvez la moyenne annuelle du cours du dollar américain en francs suisses (cours d'achat s'appliquant aux banques) de 1982 à 2007. On aimerait savoir s'il y a une relation linéaire entre l'année X et le cours du dollar Y. Ces observations ont été représentées sous forme d'un nuage de points dans la Figure 9.1. (Source : Office fédéral de la statistique, 2008).

TABLE 9.5 – Cours du dollar en Suisse (en CHF) de 1982 à 2007.

Année x_i	Cours du \$ y_i	Année x_i	Cours du \$ y_i	Année x_i	Cours du \$ y_i	Année x_i	Cours du \$ y_i
1982	2.032	1989	1.636	1996	1.235	2003	1.345
1983	2.099	1990	1.388	1997	1.451	2004	1.242
1984	2.348	1991	1.435	1998	1.448	2005	1.246
1985	2.452	1992	1.404	1999	1.503	2006	1.253
1986	1.796	1993	1.477	2000	1.689	2007	1.200
1987	1.490	1994	1.366	2001	1.687		
1988	1.464	1995	1.182	2002	1.556		

On a déjà les résultats suivants :

$$\sum_{i=1}^{26} x_i = 51\,857, \sum_{i=1}^{26} x_i^2 = 103\,430\,249, \sum_{i=1}^{26} y_i = 40.424, \sum_{i=1}^{26} y_i^2 = 65.74761, \sum_{i=1}^{26} x_i y_i = 80\,581.101.$$

1. Calculer les moyennes, les variances et la covariance des deux variables.
2. Serait-il adéquat de faire une régression linéaire de X en Y (c'est-à-dire de considérer Y comme la variable explicative et X comme la variable expliquée) ?
3. En observant le nuage de points, déterminer si la pente de la droite de régression est positive, négative ou nulle. Vérifier votre résultat en calculant le coefficient de corrélation.
4. Donner l'équation de la droite de régression.
5. Représenter cette droite dans la Figure 9.1.
6. Donner les résidus correspondants aux années 1982 et 2000.
7. Donner, à l'aide de la droite de régression, une prédiction du cours du dollar en 2010 et en 2011.
8. Déterminer la qualité de l'ajustement au moyen du coefficient de détermination.

Solution

1. Paramètres :

$$\bar{x} = \frac{1}{26}\sum_{i=1}^{26} x_i = \frac{1}{26}51\,857 = 1994.5$$

$$s_x^2 = \frac{1}{26}\sum_{i=1}^{26} x_i^2 - \bar{x}^2 = \frac{1}{26}103\,430\,249 - 1994.5^2 = 56.25$$

$$\bar{y} = \frac{1}{26}\sum_{i=1}^{26} y_i = \frac{40.424}{26} = 1.554769$$

$$s_y^2 = \frac{1}{26}\sum_{i=1}^{26} y_i^2 - \bar{y}^2 = \frac{65.748}{26} - 1.554769^2 = 0.111$$

$$s_{xy} = \frac{1}{26}\sum_{i=1}^{26} x_i y_i - \bar{x}\,\bar{y} = \frac{80\,581.1}{26} - 1994.5 \times 1.554769 = -1.714.$$

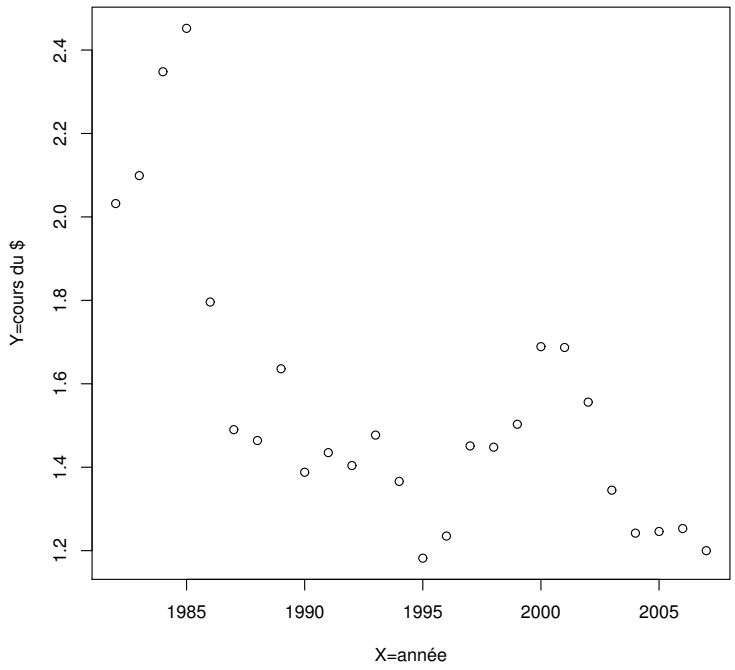


FIGURE 9.1 – Nuage de points

2. Non.
   En effet, cela n'aurait pas de sens de supposer que le cours du dollar explique le temps qui passe, mais cela peut avoir un sens de supposer que le cours du dollar est une fonction du temps qui passe.

3. La pente est négative. En effet, on remarque que Y (le cours du dollar) baisse à mesure que X (le temps) augmente.
   Coefficient de corrélation :

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-1.714}{\sqrt{56.25}\sqrt{0.111}} = -0.6846.$$

   On a $r_{xy} < 0$. Donc, la droite de régression est bien une droite décroissante.

4. On cherche a et b tels que $y = a + b\ x$:

$$b = \frac{s_{xy}}{s_x^2} = \frac{-1.714}{56.25} = -0.3047,$$

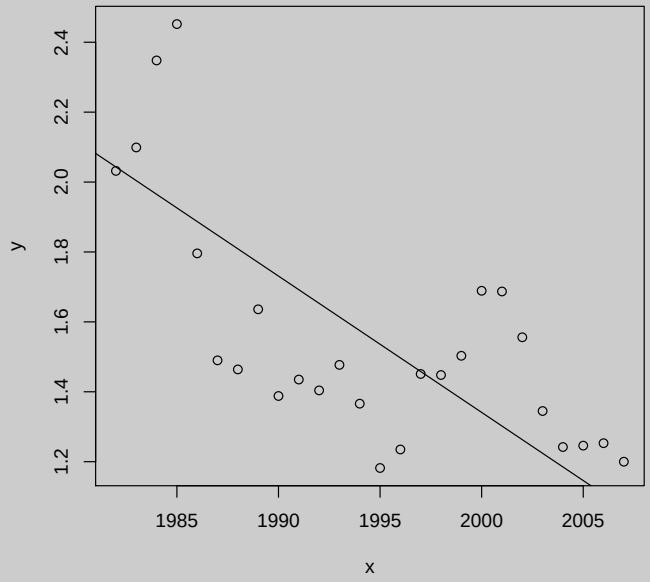
   et

$$a = \bar{y} - b\ \bar{x} = 1.555 + 0.3047 \times 1994.5 = 62.333.$$

   Donc, la droite de régression est donnée par :

$$y = 62.333 - 0.3047\ x.$$

5. Nuage de points avec la droite de régression :



6. On sait que le i-ème résidu est donné par $e_i = y_i - y_i^*$, où $y_i^* = 62.333 - 0.3047 \times x_i$,

$$y_{1982}^* = 62.333 - 0.3047 \times 1982 = 1.936,$$

$$y_{2000}^* = 62.333 - 0.3047 \times 2000 = 1.387,$$

$$e_{1982} = y_{1982} - y_{1982}^* = 2.032 - 1.936 = 0.096,$$

$$e_{2000} = y_{2000} - y_{2000}^* = 1.689 - 1.387 = 0.302.$$

Remarque : les résidus peuvent être aussi bien positifs que négatifs.

7. Prédictions pour 2010 et 2011 :

$$y_{2010}^* = 62.333 - 0.3047 \times 2010 = 1.08879$$

$$y_{2011}^* = 62.333 - 0.3047 \times 2011 = 1.05832$$

8. $r_{xy}^2 = \left(r_{xy}\right)^2 = (-0.6846121)^2 = 0.469$. Donc, 46.9% de la variance de Y est expliquée par le modèle.

Exercice 9.6. Avis pédagogiques
Neuf étudiants émettent un avis pédagogique vis-à-vis d'un professeur selon une échelle d'appréciation de 1 à 20. On relève par ailleurs la note obtenue par ces étudiants l'année précédente auprès du professeur. Les résultats sont dans le Tableau 9.6.

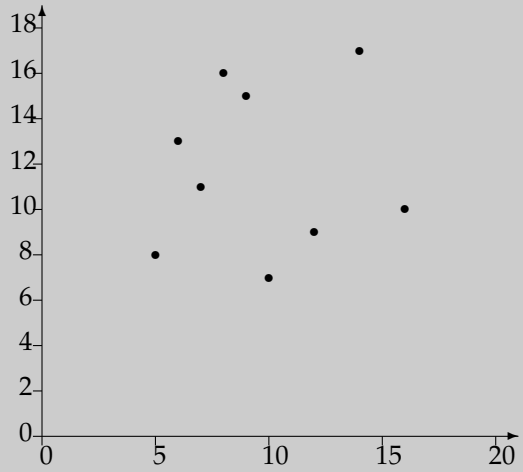
TABLE 9.6 – Avis et notes des étudiants

y = Avis	5	7	16	6	12	14	10	9	8
x = Résultat	8	11	10	13	9	17	7	15	16

1. Représentez graphiquement les deux variables.
2. Déterminez le coefficient de corrélation entre les variables X et Y. Ensuite, donnez une interprétation de ce coefficient.
3. Déterminez la droite de régression Y en fonction de X.
4. Établissez, sur base du modèle, l'avis pour un étudiant ayant obtenu 12/20.
5. Calculez la variance résiduelle et le coefficient de détermination.

Solution

1. Graphique des deux variables :



2. Tableau statistique

y_i	x_i	y_i^2	x_i^2	$x_i y_i$
5	8	25	64	40
7	11	49	121	77
16	10	256	100	160
6	13	36	169	78
12	9	144	81	108
14	17	196	289	238
10	7	100	49	70
9	15	81	225	135
8	16	64	256	128
87	106	951	1354	1034

Moyennes, variances, covariance et coefficient de corrélation entre les variables x et x.

$$\bar{y} = \frac{87}{9} = 9.667,$$

$$s_y^2 = \frac{951}{9} - 9.667^2 = 12.22,$$

$$\bar{x} = \frac{106}{9} = 11.78,$$

$$s_x^2 = \frac{1354}{9} - 11.78^2 = 11.73,$$

$$s_{xy} = \frac{1034}{9} - 9.667 \times 11.78 = 1.037,$$

$$r_{xy} = \frac{1.037}{\sqrt{12.22 \;\; 11.73}} = 0.087.$$

3. Droite de régression Y en fonction de X :

$$D_{y|x} \; : y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x},)$$

$$D_{y|x} \; : y = 0.088x + 8.625.$$

4. Valeur ajustée pour une cote de 12/20, (x=12)

$$y = 0.088 \times 12 + 8.625 = 9.686$$

5. Mesure de la qualité du modèle au moyen de la variance résiduelle et du coefficient de détermination :

$$\begin{aligned} s_{y|x}^2 &= s_y^2(1 - r^2) \\ &= 12.22(1 - 0.087^2) \\ &= 12.13 \text{ à comparer avec } s_y^2 = 12.22 \end{aligned}$$

Coefficient de détermination :

$$r^2 = 0.087^2 = 0.008.$$

Ce coefficient représente la proportion de variance expliquée par le modèle (ici 0.8% faible).

Exercice 9.7. Consommation de médicaments
Lors d'une enquête épidemiologique portant sur la consommation de médicaments psychotropes, on a demandé à des hommes et à des femmes de différentes catégories d'âge de dire si au cours des deux précédentes semaines, ils ou elles avaient consommé des médicaments de ce type. Parmi les personnes de la catégorie d'âge des 33 à 44 ans, on observe la répartition des effectifs suivante :

Table 9.7 – Consommation de médicaments, source : Murray et al. (1981, pp. 551–560).

	Consommation de psychotropes	
	Oui	Autres réponses
Hommes	32	596
Femmes	89	700

1. Etablir le tableau des fréquences.
2. Etablir le tableau des profils lignes.
3. Etablir le tableau des profils colonnes.
4. Calculez le tableau des fréquences et des effectifs théoriques sous hypothèse d'indépendance entre le sexe et les réponses.
5. Calculez les écarts à l'indépendance, le khi-carré, le phi-deux et le V de Cramer.

Solution

1. Tableau des effectifs et tableau des fréquences (j lignes, k colonnes, $f_{jk} = n_{jk}/n$) :

Tableau des effectifs

	Oui	Autres réponses	Total
Hommes	32	596	628
Femmes	89	700	789
Total	121	1296	1417

Tableau des fréquences

	Oui	Autres réponses	Total
Hommes	0.022	0.421	0.443
Femmes	0.063	0.494	0.557
Total	0.085	0.915	1.000

2. Profils lignes $f_k^{(j)} = n_{jk}/n_{j.}$:

Tableau des profils lignes

	Oui	Autres réponses	Total
Hommes	0.051	0.949	1.000
Femmes	0.113	0.887	1.000
Total	0.085	0.915	1.000

3. Profils colonnes $f_j^{(k)} = n_{jk}/n_{.k}$:

Tableau des profils colonnes

	Oui	Autres réponses	Total
Hommes	0.264	0.460	0.443
Femmes	0.736	0.540	0.557
Total	1.000	1.000	1.000

4. Effectifs théoriques $n^*_{jk} = \frac{n_{j.}n_{.k}}{n}$:
$n^*_{jk} = \frac{n_{j.}n_{.k}}{n}$ ex. cellule $n_{11} = 121 \times 628/1417 = 53.626$

Tableau des effectifs théoriques

	Oui	Autres réponses	Total
Hommes	53.626	574.374	628
Femmes	67.374	721.626	789
Total	121	1296	1417

Fréquence théorique $f^*_{jk} = f_{j.} f_{.k}$

Tableau des fréquences théoriques

	Oui	Autres réponses	Total
Hommes	0.038	0.405	0.443
Femmes	0.047	0.510	0.557
Total	0.085	0.915	1.000

5. Tableau des écarts à l'indépendance $e_{jk} = n_{jk} - n^*_{jk}$:

Tableau des écarts à l'indépendance :

	Oui	Autres réponses	Total
Hommes	-21.626	21.626	0
Femmes	21.626	-21.626	0
Total	0	0	0

Tableau des e^2_{jk}/n^*_{jk} :

Tableau des e^2_{jk}/n^*_{jk}

	Oui	Autres réponses	Total
Hommes	8.721	0.814	9.535
Femmes	6.942	0.648	7.590
Total	15.663	1.462	17.125

Le khi-carré observé vaut $\chi^2 = 17.125$.
Le phi-deux vaut $\phi^2 = 17.125/1417 = 0.0121$.

$$V = \sqrt{\frac{\phi^2}{\min(j-1, k-1)}} = \sqrt{0.0121/1} = 0.110$$

(dépendance faible).

Exercice 9.8. Daltonie
Mille personnes ont été classés selon leur sexe et selon le fait d'être daltonien ou pas. Les résultats sont donné dans le tableau de contingence suivant :

TABLE 9.8 – Tableau des effectifs observés n_{jk}

	homme	femme	total
normal	442	514	956
daltonien	38	6	44
total	480	520	1000

1. Calculer le tableau des fréquences.
2. Calculer le tableau des profils lignes.
3. Calculer le tableau des profils colonnes.
4. Calculer le tableau des effectifs théoriques n^*_{jk}.
5. Calculer le tableau des écarts à l'indépendance e_{jk}.
6. Calculer le tableau des e^2_{jk}/n^*_{jk}.
7. Calculer le χ^2_{obs}, le ϕ^2 et le V de Cramer.

Solution
Tableau des fréquences f_{jk}

	homme	femme	total
normal	0.442	0.514	0.956
daltonien	0.038	0.006	0.044
total	0.48	0.52	1

Tableau des profils lignes

	homme	femme	total
normal	0.462	0.538	1
daltonien	0.864	0.136	1

Tableau des profils colonnes

	homme	femme
normal	0.921	0.988
daltonien	0.079	0.012
total	1	1

Tableau des effectifs théoriques n^*_{jk}

	homme	femme	total
normal	458.88	497.12	956
daltonien	21.12	22.88	44
total	480	520	1000

Tableau des écarts à l'indépendance e_{jk}

	homme	femme	total
normal	−16.88	16.88	0
daltonien	16.88	−16.88	0
total	0	0	0

Tableau des e^2_{jk}/n^*_{jk}

	homme	femme	total
normal	0.621	0.573	1.194
daltonien	13.491	12.453	25.944
total	14.112	13.026	27.138

$\chi^2_{obs} = 27.138$, $\phi^2 = 0.027$ et V de Cramer $= 0.027$.

Exercice 9.9. Loyers et cantons
On considère les deux variables *Loyer mensuel* et *Canton* et les valeurs respectives *inférieur à 1000 francs suisses*, *supérieur ou égal à 1000 francs suisses* et *Vaud*, *Valais*, *Neuchâtel*, *Genève*. On s'intéresse à une éventuelle dépendance entre ces deux variables. On a donc effectué des observations. Le Tableau 9.9 comporte les contingences des données observées. Ces deux variables sont-elles indépendantes ?

Table 9.9 – Tableau des contingences

	Vaud	Valais	Neuchâtel	Genève
loyer < 1000	40	25	35	60
loyer ≥ 1000	80	25	15	120

Solution

Tableau des effectifs

	Vaud	Valais	Neuchâtel	Genève	Total
loyer < 1000	40	25	35	60	160
loyer ≥ 1000	80	25	15	120	240
Total	120	50	50	180	400

Effectifs théoriques

	Vaud	Valais	Neuchâtel	Genève	Total
loyer < 1000	48	20	20	72	160
loyer ≥ 1000	72	30	30	108	240
Total	120	50	50	180	400

Écarts à l'indépendance

	Vaud	Valais	Neuchâtel	Genève	Total
loyer < 1000	-8	5	15	-12	0
loyer ≥ 1000	8	-5	-15	12	0
Total	0	0	0	0	0

$$\chi^2_{obs} = \frac{(-8)^2}{48} + \frac{5^2}{20} + \frac{15^2}{20} + \frac{(-12)^2}{72} + \frac{8^2}{72} + \frac{(-5)^2}{30} + \frac{(-15)^2}{30} + \frac{12^2}{108} = 26.389.$$

$$V = \sqrt{\frac{\chi^2_{obs}}{n \min(J-1, K-1)}} = \sqrt{\frac{26.389}{400 \times \min(1,3)}} = \sqrt{\frac{26.389}{400}} = 0.257.$$

(dépendance modérée).

Exercice 9.10. Bières
Pour comparer deux bières, on fait une experience avec 100 amateurs de chaque marque. Chaque groupe affirme connaître la différence entre les deux et préférer nettement la sienne. On demande à chaque sujet d'identifier sa préférence, après avoir goûté les deux. Les résultats sont donnés dans le Tableau 9.10. Est-ce que l'habitude et la préférence sont des caractères indépendants ?

TABLE 9.10 – Tableau des effectifs observés n_{jk}

		boivent d'habitude		
		A	B	total
ont	A	65	45	110
préféré	B	35	55	90
total		100	100	200

Solution
Tableau des fréquences f_{jk}

		boivent d'habitude		
		A	B	total
ont	A	0.325	0.225	0.55
préféré	B	0.175	0.275	0.45
total		0.5	0.5	1

Tableau des profils lignes

		boivent d'habitude		
		A	B	total
ont	A	0.591	0.409	1
préféré	B	0.389	0.611	1

Tableau des profils colonnes

		boivent d'habitude	
		A	B
ont	A	0.65	0.45
préféré	B	0.35	0.55
total		1	1

Tableau des effectifs théoriques n^*_{jk}

		boivent d'habitude		
		A	B	total
ont	A	55	55	110
préféré	B	45	45	90
total		100	100	200

Tableau des écarts à l'indépendance e_{jk}

		boivent d'habitude		
		A	B	total
ont	A	10	−10	0
préféré	B	−10	10	0
total		0	0	0

Tableau des e^2_{jk}/n^*_{jk}

		boivent d'habitude		
		A	B	total
ont	A	1.818	1.818	3.636
préféré	B	2.222	2.222	4.444
total		4.040	4.040	8.080

$\chi^2_{obs} = 8.080$, $\phi^2 = 0.040$ et V de Cramer $= 0.040$.

Exercice 9.11. État civil et nationalité
On s'est intéressé à l'état civil (C = Célibataire, M = marié, D = divorcé, V = veuf) ainsi qu'à la nationalité (E = de nationalité étrangère, S = de nationalité suisse) de 300 individus. Ces données se trouvent dans le Tableau 9.11. On aimerait savoir si ces deux variables sont dépendantes.

Table 9.11 – Tableau des effectifs

	C	M	D	V	Total
E	30	–	15	15	120
S	–	90	45	15	180
Total	–	–	–	–	–

1. Certaines cases ont été effacées par mégarde dans le Tableau 9.11. Remplir ce tableau.
2. Quel pourcentage des individus sont mariés ? Quel pourcentage des individus sont de nationalité étrangère ?
3. Quel pourcentage des individus étrangers sont mariés ?
4. Quel pourcentage des individus divorcés sont de nationalité suisse ?
5. Les variables *État civil* et *Nationalité* sont-elles dépendantes ?

Solution

1. Tableau des effectifs complété :

Tableau des effectifs (complet)

	C	M	D	V	Total
E	30	**60**	15	15	120
S	**30**	90	45	15	180
Total	**60**	**150**	**60**	**30**	**300**

2. Fréquences : $f_{jk} = \dfrac{n_{jk}}{n}$.
Pourcentage d'individus mariés :

$$f_{.2} = \frac{n_{.2}}{n} = \frac{150}{300} = 0.5 = 50\%$$

Pourcentage d'individus de nationalité étrangère :

$$f_{1.} = \frac{n_{1.}}{n} = \frac{120}{300} = 0.4 = 40\%$$

Tableau des fréquences

	C	M	D	V	Total
E	0.1	0.2	0.05	0.05	**0.4**
S	0.1	0.3	0.15	0.05	0.6
Total	0.2	**0.5**	0.2	0.1	1.0

3. Profils lignes : $f_k^{(j)} = \frac{n_{jk}}{n_{j.}}$.
Pourcentage des individus étrangers qui sont mariés :

$$f_2^{(1)} = \frac{n_{12}}{n_{1.}} = \frac{60}{120} = 0.5 = 50\%$$

Profils lignes

	C	M	D	V	Total
E	0.25	**0.5**	0.125	0.125	1
S	0.167	0.5	0.25	0.083	1
Total	0.2	0.5	0.2	0.1	1

4. Profils colonnes : $f_j^{(k)} = \frac{n_{jk}}{n.k}$.
   Pourcentage des divorcés qui sont de nationalité suisse :

$$f_2^{(3)} = \frac{n_{23}}{n.3} = \frac{45}{60} = 0.75 = 75\%$$

Profils colonnes

	C	M	D	V	Total
E	0.5	0.4	0.25	0.5	0.4
S	0.5	0.6	**0.75**	0.5	0.6
Total	1.0	1.0	1.0	1.0	1.0

5. Effectifs théoriques : $n_{jk}^* = \frac{n_{j.}n_{.k}}{n}$

$$n_{11}^* = \frac{n_{1.}n_{.1}}{n} = \frac{120 \times 60}{300} = 24$$

$$n_{12}^* = \frac{n_{1.}n_{.2}}{n} = \frac{120 \times 150}{300} = 60$$

$$n_{23}^* = \frac{n_{2.}n_{.3}}{n} = \frac{180 \times 60}{300} = 36$$

Effectifs théoriques

	C	M	D	V	Total
E	24	60	24	12	120
S	36	90	36	18	180
Total	60	150	60	30	300

Écarts à l'indépendance : $e_{jk} = n_{jk} - n_{jk}^*$.

$$e_{11} = n_{11} - n_{11}^* = 30 - 24 = 6$$

$$e_{12} = n_{12} - n_{12}^* = 60 - 60 = 0$$

$$e_{23} = n_{23} - n_{23}^* = 45 - 36 = 9$$

Écarts à l'indépendance

	C	M	D	V	Total
E	6	0	-9	3	0
S	-6	0	9	-3	0
Total	0	0	0	0	0

Tableau des e_{ij}^2/n_{ij}^*

χ^2 par cellule, avec signes

	C	M	D	V	Total
E	1.5 (+)	0	3.375 (-)	0.75 (+)	0
S	1 (-)	0	2.25 (+)	0.5 (-)	0
Total	0	0	0	0	0

$$\chi_{obs}^2 = \sum_{k=1}^{K}\sum_{j=1}^{J} \frac{e_{jk}^2}{n_{jk}^*} = \frac{6^2}{24} + \frac{(-6)^2}{36} + \frac{(-9)^2}{24} + \frac{9^2}{36} + \frac{3^2}{12} + \frac{(-3)^2}{18} = 9.375.$$

La valeur du χ^2 vient principalement de la colonne 'D'. En regardant les signes des écarts, entre parenthèses, on constate que le nombre de divorcés parmi les étrangers est plus faible qu'attendu si les variables étaient indépendantes et que le nombre de divorcés parmi les suisses est lui plus élevé.

$$\phi^2 = \frac{\chi_{obs}^2}{n} = \frac{9.375}{300} = 0.03125.$$

$$V = \sqrt{\frac{\phi^2}{\min(J-1, K-1)}} = \sqrt{\frac{0.03125}{\min(1,3)}} = 0.177.$$

On remarque que V est proche de 0. Donc, la dépendance entre l'état civil et la nationalité est plutôt faible.

Exercice 9.12. Dépendance entre variables dichotomiques
Considérons deux variables ne pouvant prendre que 2 valeurs, $0, 1$ pour la première et A, B pour la seconde. On a affectué 200 observations afin de déterminer s'il y a une dépendance entre ces deux variables. On sait que, parmi ces 200 observations, 50 prenaient la valeur 0 en la première variable et que 110 prenaient la valeur A en la seconde variable.

1. Reconstruire le tableau des effectifs à partir du Tableau 9.12 des profils lignes.

Table 9.12 – Tableau des profils lignes

	A	B	Total
0	0.4	0.6	1
1	0.6	0.4	1
	0.55	0.45	1

2. Calculez le tableau des effectifs théoriques en cas d'indépendance.
3. Déterminez s'il y a une dépendance entre les deux variables.

Solution

1. On a

	A	B	Total
0	$0.4 = f_1^{(1)}$	$0.6 = f_2^{(1)}$	1
1	$0.6 = f_1^{(2)}$	$0.4 = f_2^{(2)}$	1
	0.55	0.45	1

On sait que $f_k^{(j)} = \frac{n_{jk}}{n_{j.}}$ pout $k = 1, 2$ et $j = 1, 2$. De plus, les indications nous donnent une partie du tableau des effectifs :

	A	B	Total
0	n_{11}	n_{12}	$n_{1.} = 50$
1	n_{21}	n_{22}	$n_{2.}$
	$n_{.1} = 110$	$n_{.2}$	$n = 200$

$$n_{.1} + n_{.2} = n \Rightarrow n_{.2} = 200 - 110 = 90$$

$$n_{1.} + n_{2.} = n \Rightarrow n_{2.} = 200 - 15 = 150$$

De plus, on a :

$$f_k^{(j)} = \frac{n_{jk}}{n_{j.}}.$$

$$n_{jk} = f_k^{(j)} \times n_{j.}.$$

Ainsi,

$$n_{11} = f_1^{(1)} \times n_{1.} = 0.4 \times 50 = 20,$$

$$n_{12} = f_2^{(1)} \times n_{1.} = 0.6 \times 50 = 30,$$

$$n_{21} = f_1^{(2)} \times n_{2.} = 0.6 \times 150 = 90,$$

$$n_{22} = f_2^{(2)} \times n_{2.} = 0.4 \times 150 = 60.$$

On a donc trouvé :

	A	B	Total
0	$n_{11} = 20$	$n_{12} = 30$	$n_{1.} = 50$
1	$n_{21} = 90$	$n_{22} = 60$	$n_{2.} = 150$
	$n_{.1} = 110$	$n_{.2} = 90$	$n = 200$

2. On déduit du tableau des effectifs que $f_{1.} = 0.25$ et $f_{2.} = 0.75$. Cela permet de construire le tableau des fréquences théoriques :

	A	B	Total
0	$f^*_{11} = 0.1375$	$f^*_{12} = 0.1125$	$f_{1.} = 0.25$
1	$f^*_{21} = 0.4125$	$f^*_{22} = 0.3375$	$f_{2.} = 0.75$
	$f_{.1} = 0.55$	$f_{.2} = 0.45$	1

Et celui des effectifs théoriques :

	A	B	Total
0	$n^*_{11} = 27.5$	$n^*_{12} = 22.5$	$n_{1.} = 50$
1	$n^*_{21} = 82.5$	$n^*_{22} = 67.5$	$n_{2.} = 150$
	$n_{.1} = 110$	$n_{.2} = 90$	$n = 200$

3. Le tableau des écarts est :

	A	B	Total
0	$e_{11} = -7.5$	$e_{12} = +7.5$	0
1	$e_{21} = +7.5$	$e_{22} = -7.5$	0
	0	0	0

Le χ^2 vaut 6.06. Le ϕ^2 vaut 0.03 et le V vaut 0.173. Il ne semble donc pas y avoir de dépendance entre les deux variables, étant donné que le V est proche de 0.

Exercice 9.13. Manque de sommeil
Des chercheurs ont mené une étude sur le sommeil auprès de 105 femmes et 86 hommes. L'étude porte sur le manque de sommeil. Les participants devaient indiquer s'ils ont ressenti un manque de sommeil au cours de la semaine précédente. Les résultats sont dans le Tableau 9.13.

Table 9.13 – Sensation de manque de sommeil selon le sexe

	Sensation de manque de sommeil	
	oui	non
Homme	36	50
Femme	54	51

1. Etablir le tableau des effectifs complet.
2. Etablir le tableau des profils lignes.
3. Etablir le tableau des profils colonnes.
4. À partir de ces données, peut-on dire que les différences au niveau de la perception de la quantité du sommeil est liée à la variable sexe ?

Solution

1. Tableau des effectifs

	oui	non	total
Homme	36	50	86
Femme	54	51	105
Total	90	101	191

2. Tableau des profils lignes

	oui	non	total
Homme	0.42	0.58	1
Femme	0.51	0.49	1
Total	0.47	0.53	1

3. Tableau des profils colonnes

	oui	non	total
Homme	0.40	0.49	0.45
Femme	0.60	0.51	0.55
Total	1	1	1

4. Tableau des effectifs théoriques : $n^*_{jk} = (n_{j.}n_{.k})/n$
Par exemple, $n^*_{11} = (86 \times 90)/191 = 40.52$

	oui	non	total
Homme	40.52	45.48	86
Femme	49.48	55.52	105
Total	90	101	191

Tableau des écarts à l'indépendance : $e_{jk} = n_{jk} - n^*_{jk}$
Par exemple $e_{11} = 36 - 40.52 = -4.52$

	oui	non	total
Homme	-4.52	4.52	0.00
Femme	4.52	-4.52	0.00
Total	0.00	0.00	0.00

Tableau des e^2_{jk}/n^*_{jk}
Par exemple, $e^2_{11}/n^*_{11} = (-4.52)^2/40.52 = 0.50$

	oui	non	total
Homme	0.50	0.45	0.95
Femme	0.41	0.37	0.78
Total	0.91	0.82	1.73= χ^2_{obs}

phi-deux : $\phi^2 = \frac{\chi^2_{obs}}{n} = \frac{1.73}{191} = 0.009$

V de Cramer : $V = \sqrt{\frac{\phi^2}{\min(J-1,K-1)}} = \sqrt{\frac{0.009}{\min(1,1)}} = \sqrt{\frac{0.009}{1}} = 0.095$

Comme le V de Cramer est proche de 0, les 2 variables très peu dépendantes

Exercice 9.14. Habitudes alimenaires selon le sexe
Une enquête sur les habitudes alimentaires de 968 étudiants a été réalisée. La première variable d'intérêt est le sexe de la personne et la deuxième variable d'intérêt est l'endroit le plus fréquenté par la personne sur l'heure du midi (cafétéria, restaurant, chez soi). En tout, 400 étudiants ont l'habitude de manger à la cafétéria, 304 étudiants mangent plus régulièrement au restaurant. On sait aussi que 116 hommes ont l'habitudes de manger à la maison.

Le Tableau 9.14 contient les profils colonnes.

Table 9.14 – Profils colonnes

	cafétéria	restaurant	chez soi	total
Homme	0.32	0.52	0.44	0.415
Femme	0.68	0.48	0.56	0.585
Total	1	1	1	1

1. Etablir le tableau des effectifs
2. À partir de ces données, peut-on dire que les habitudes alimenaires sont dépendantes de la variable sexe ?

Solution

1. Tableau des effectifs

	cafétéria	restaurant	chez soi	total
Homme	128	158	116	402
Femme	272	146	148	566
Total	400	304	264	968

2. Dépendance
Tableau des effectifs théoriques : $n^*_{jk} = (n_{j.}n_{.k})/n$
Par exemple, $n^*_{11} = (400 \times 402)/968 = 166.11$

	cafétéria	restaurant	chez soi	total
Homme	166.11	126.25	109.64	402
Femme	233.89	177.75	154.36	566
Total	400	304	264	968

Tableau des écarts à l'indépendance : $e_{jk} = n_{jk} - n^*_{jk}$
Par exemple, $e_{11} = n_{11} - n^*_{11} = 128 - 166.11 = -38.11$

	cafétéria	restaurant	chez soi	total
Homme	-38.11	31.75	6.36	0.00
Femme	38.11	-31.75	-6.36	0.00
Total	0.00	0.00	0	0.00

Tableau des e^2_{jk}/n^*_{jk}
Par exemple, $e^2_{11}/n^*_{11} = (-38.11)^2/166.11 = 8.74$

	cafétéria	restaurant	chez soi	total
Homme	8.74	7.98	0.37	17.10
Femme	6.21	5.67	0.26	12.15
Total	14.95	13.65	0.63	29.25= χ^2_{obs}

Phi-deux : $\phi^2 = \dfrac{\chi^2_{obs}}{n} = \dfrac{29.25}{968} = 0.03.$

V de Cramer : $V = \sqrt{\frac{\phi^2}{\min(J-1,K-1)}} = \sqrt{\frac{0.03}{1}} = 0.17.$

Comme le V de Cramer est proche de 0, les 2 variables sont très peu dépendantes.

Chapitre 10

Exercices : Théorie des indices, mesures d'inégalité

Exercice 10.1. Indices simples
Le Tableau 10.1 contient l'indice simple $I(t/t')$ pour $t, t' = 0, 1, 2$ d'un bien. Remplir les cases vides.

Table 10.1 – Indice simple d'un bien

	$t = 0$	$t = 1$	$t = 2$
$t' = 0$	100	120	–
$t' = 1$	–	–	140
$t' = 2$	–	–	–

Solution

— La diagonale ne doit contenir que des 100. En effet, on a $t = t'$ sur la diagonale et ainsi

$$I(t/t') = I(t/t) = 100\frac{p_t}{p_t} = 100.$$

	$t = 0$	$t = 1$	$t = 2$
$t' = 0$	100	120	–
$t' = 1$	–	**100**	140
$t' = 2$	–	–	**100**

— On sait que

$$I(t/t') = 100\frac{p_t}{p_{t'}}$$

et donc

$$p_t = \frac{I(t/t')}{100}p_{t'}.$$

Ainsi, on a

$$p_1 = \frac{I(1/0)}{100}p_0 = \frac{120}{100}p_0 = 1.2p_0$$

$$p_2 = \frac{I(2/1)}{100}p_1 = \frac{140}{100}p_1 = 1.4p_1 = 1.4 \times 1.2p_0 = 1.68p_0.$$

On obtient :

$$I(0/1) = 100\frac{p_0}{p_1} = 100\frac{p_0}{1.2p_0} = \frac{100}{1.2} = 83.333,$$

$$I(0/2) = 100\frac{p_0}{p_2} = 100\frac{p_0}{1.68p_0} = \frac{100}{1.68} = 59.524,$$

$$I(1/2) = 100\frac{p_1}{p_2} = 100\frac{p_1}{1.4p_1} = \frac{100}{1.4} = 71.429,$$

$$I(2/0) = 100\frac{p_2}{p_0} = 100\frac{1.68p_0}{p_0} = 100 \times 1.68 = 168.$$

On complète le Tableau :

	$t = 0$	$t = 1$	$t = 2$
$t' = 0$	100	120	**168**
$t' = 1$	**83.333**	**100**	140
$t' = 2$	**59.524**	**71.429**	**100**

Exercice 10.2. Indice de Laspeyres
Les prix d'actions boursières de quatre compagnies d'automobiles sont indiquées, pour deux semaines consécutives, dans le Tableau 10.2, qui présente aussi les quantités d'actions échangées :

Table 10.2 – Tableau des prix et quantités

Compagnie	Semaine 1		Semaine 2	
	Prix	Quantité	Prix	Quantité
A	20	8	23	13
B	4	4	4	4
C	5	14	6	18
D	46	27	52	25

1. Calculer les indices élémentaires de l'évolution hebdomadaire des prix d'actions de chaque compagnie et interpréter les résultats obtenus.
2. Déterminer l'indice de Laspeyres de l'évolution hebdomadaire des prix d'actions des quatre compagnies dans leur ensemble.

Solution

1. $I_{PA}(2/1) = 100 \times \dfrac{P_A(2)}{P_A(1)} = 100 \times \dfrac{23}{20} = 115,$

   $I_{PB}(2/1) = 100 \times \dfrac{4}{4} = 1,$

   $I_{PC}(2/1) = 100 \times \dfrac{6}{5} = 120,$

   $I_{PD}(2/1) = 100 \times \dfrac{52}{46} = 113.$

   Individuellement, les prix ont augmentés.
2. Indice de Laspeyre des prix :

$$\begin{aligned} L_p(2/1) &= 100 \times \frac{q_A(1)p_A(2) + q_B(1)p_B(2) + q_C(1)p_C(2) + q_D(1)p_D(2)}{q_A(1)p_A(1) + q_B(1)p_B(1) + q_C(1)p_C(1) + q_D(1)p_D(1)} \\ &= 100 \times \frac{8 \times 23 + 4 \times 4 + 14 \times 6 + 27 \times 52}{20 \times 8 + 4 \times 4 + 5 \times 14 + 46 \times 27} = 100 \times \frac{1688}{1488} = 113.44. \end{aligned}$$

Exercice 10.3. Indice de Laspeyres et de Paasche

Les dépenses de Monsieur Durand en 2006 liées à son activité professionnelle ont été les suivantes :

— *Repas* : 15.- par repas. Monsieur Durand a pris 1 repas à l'extérieur par jour de travail.
— *Abonnement de train* : 100.- par mois. Monsieur Durand a payé son abonnement pour les 12 mois de l'année.
— *Cafés* : 2.50 par pièce. Monsieur Durand s'est offert deux cafés par jour de travail.

Monsieur Durand a travaillé 46 semaines en 2006, ce qui représente 230 jours de travail.

En 2007, les dépenses de Monsieur Durand et les prix ont changé, à savoir :

— *Repas* : 16.- par repas. Monsieur Durand a pris 1 repas à l'extérieur par jour de travail.
— *Abonnement de train* : 100.- par mois. Monsieur Durand a payé cette année son abonnement 11 mois de l'année.
— *Cafés* : 3.- par pièce. Monsieur Durand s'est offert un cafés par jour de travail.

En revanche, Monsieur Durand a travaillé 45 semaines en 2007. Il a donc effectué 225 jours de travail.

1. Remplir le Tableau 10.3
2. Quel indice allez vous calculer si vous désirez :
   (a) calculer le prix d'un café en 2007 par rapport au prix d'un café en 2006 ?
   (b) calculer la dépense totale de Monsieur Durand en 2007 par rapport à 2006 en considérant
      i. les quantités de 2006 ?
      ii. les quantités de 2007 ?
3. Calculer pour 2007, en considérant 2006 comme l'année de référence,
   (a) l'indice de Laspeyres et
   (b) l'indice de Paasche.

Table 10.3 – Dépenses de Monsieur Durand

Temps Année	0 2006		1 2007	
	Prix p_{0i}	Quantité q_{0i}	Prix p_{1i}	Quantité q_{1i}
Repas Abonnement Cafés				

Solution

1. La tableau complété :

Temps Année	0 2006		1 2007	
	Prix p_{0i}	Quantité q_{0i}	Prix p_{1i}	Quantité q_{1i}
Repas	15	230	16	225
Abonnement	100	12	100	11
Cafés	2.50	460	3	225

2. (a) Prix d'un café en 2007 par rapport au prix d'un café en 2006 $\rightsquigarrow$ indice simple $I(2007/2006)$.
   (b) Dépense totale en 2007 par rapport à 2006. En considérant
      i. les quantités de 2006 $\rightsquigarrow$ indice de Laspeyres $L(2007/2006)$,
      ii. les quantités de 2007 $\rightsquigarrow$ indice de Paasche $P(2007/2006)$.
3. Pour 2007, en considérant 2006 comme l'année de référence,
   (a) l'indice de Laspeyres :

$$L(1/0) = 100\frac{\sum_{i=1}^{n} q_{0i}p_{1i}}{\sum_{i=1}^{n} q_{0i}p_{0i}} = 100\frac{230\times 16 + 12\times 100 + 460\times 3}{230\times 15 + 12\times 100 + 460\times 2.5} = 107.931.$$

(b) l'indice de Paasche :

$$P(1/0) = 100\frac{\sum_{i=1}^{n} q_{1i}p_{1i}}{\sum_{i=1}^{n} q_{1i}p_{0i}} = 100\frac{225 \times 16 + 11 \times 100 + 225 \times 3}{225 \times 15 + 11 \times 100 + 225 \times 2.5} = 106.7.$$

Exercice 10.4. Indices composites
Soit le tableau suivant contenant les prix et les quantités de trois biens sur trois périodes :

Table 10.4 – Tableau des prix et quantités

Période	0		1		2	
Bien	Prix	Quantité	Prix	Quantité	Prix	Quantité
Bien A	100	14	150	10	200	8
Bien B	60	10	50	12	40	14
Bien C	160	4	140	5	140	5

1. Calculer l'indice de Laspeyres des prix $L(1/0)$, $L(2/0)$ et $L(2/1)$.
2. Quelle est la valeur de l'indice chaîne de Laspeyres $CL(2/0)$?
3. Calculer l'indice de Paasche des prix $P(1/0)$, $P(2/0)$ et $P(2/1)$.
4. Calculer les trois indices correspondants de Fisher et Sidgwick.

Solution

— Indice de Laspeyre des prix :

$$\begin{aligned} L(1/0) &= 100 \times \frac{q_A(0)p_A(1) + q_B(0)p_B(1) + q_C(0)p_C(1)}{q_A(0)p_A(0) + q_B(0)p_B(0) + q_C(0)p_C(0)} \\ &= 100 \times \frac{14 \times 150 + 10 \times 50 + 4 \times 140}{14 \times 100 + 10 \times 60 + 4 \times 160} = 100 \times \frac{3160}{2640} = 119.70. \end{aligned}$$

$$\begin{aligned} L(2/0) &= 100 \times \frac{q_A(0)p_A(2) + q_B(0)p_B(2) + q_C(0)p_C(2)}{q_A(0)p_A(0) + q_B(0)p_B(0) + q_C(0)p_C(0)} \\ &= 100 \times \frac{14 \times 200 + 10 \times 40 + 4 \times 140}{14 \times 100 + 10 \times 60 + 4 \times 160} = 100 \times \frac{3760}{2640} = 142.42. \end{aligned}$$

$$\begin{aligned} L(2/1) &= 100 \times \frac{q_A(1)p_A(2) + q_B(1)p_B(2) + q_C(1)p_C(2)}{q_A(1)p_A(1) + q_B(1)p_B(1) + q_C(1)p_C(1)} \\ &= 100 \times \frac{10 \times 200 + 12 \times 40 + 5 \times 140}{10 \times 150 + 12 \times 50 + 5 \times 140} = 100 \times \frac{3180}{2800} = 113.57. \end{aligned}$$

— Indice chaîne de Laspeyres :

$$\begin{aligned} CL(2/0) &= 100 \times \frac{L_p(2/1)}{100} \times \frac{L_p(1/0)}{100} \\ &= 100 \times \frac{113.57}{100} \times \frac{119.70}{100} = \frac{13594.33}{100} = 135.94. \end{aligned}$$

— Indice de Paashe des prix :

$$\begin{aligned} P(1/0) &= 100 \times \frac{q_A(1)p_A(1) + q_B(1)p_B(1) + q_C(1)p_C(1)}{q_A(1)p_A(0) + q_B(1)p_B(0) + q_C(1)p_C(0)} \\ &= 100 \times \frac{10 \times 150 + 12 \times 50 + 5 \times 140}{10 \times 100 + 12 \times 60 + 5 \times 160} = 100 \times \frac{2800}{2520} = 111.11. \end{aligned}$$

$$\begin{aligned} P(2/0) &= 100 \times \frac{q_A(2)p_A(2) + q_B(2)p_B(2) + q_C(2)p_C(2)}{q_A(2)p_A(0) + q_B(2)p_B(0) + q_C(2)p_C(0)} \\ &= 100 \times \frac{8 \times 200 + 14 \times 40 + 5 \times 140}{8 \times 100 + 14 \times 60 + 5 \times 160} = 100 \times \frac{2860}{2440} = 117.21. \end{aligned}$$

$$\begin{aligned} P(2/1) &= 100 \times \frac{q_A(2)p_A(2) + q_B(2)p_B(2) + q_C(2)p_C(2)}{q_A(2)p_A(1) + q_B(2)p_B(1) + q_C(2)p_C(1)} \\ &= 100 \times \frac{8 \times 200 + 14 \times 40 + 5 \times 140}{8 \times 150 + 14 \times 50 + 5 \times 140} = 100 \times \frac{2860}{2600} = 110. \end{aligned}$$

— Indice de Fisher des prix :

$$F(1/0) = \sqrt{L_p(1/0) \times P_p(1/0)} = \sqrt{119.70 \times 111.11} = \sqrt{13299.87} = 115.33.$$

$$F(2/0) = \sqrt{L_p(2/0) \times P_p(2/0)} = \sqrt{142.42 \times 117.21} = \sqrt{16693.05} = 129.20.$$

$$F(2/1) = \sqrt{L_p(2/1) \times P_p(2/1)} = \sqrt{113.57 \times 110} = \sqrt{12492.7} = 111.77.$$

— Indice de Sidgwick des prix :

$$S(1/0) = (L_p(1/0) + P_p(1/0))/2 = (119.70 + 111.11)/2 = 115.41.$$

$$S(2/0) = (L_p(2/0) + P_p(2/0))/2 = (142.42 + 117.21)/2 = 129.82.$$

$$Sp(2/1) = (L_p(2/1) + P_p(2/1))/2 = (113.57 + 110)/2 = 111.79.$$

Exercice 10.5. Indices de prix et montres
Vous travaillez au sein d'une fabrique de montres haut de gamme dont les ventes se trouvent dans le Tableau 10.5.

Table 10.5 – Prix (en milliers de francs suisses) et quantité de montres vendues

Année	2003		2004		2005		2006		2007	
	Prix	Qté	Prix	Qté	Prix	Qté	Prix	Qté	Prix	Qté
Montre A	12	10	12	10	14	12	14	15	15	12
Montre B	15	12	16	10	16	10	18	12	20	8
Montre C	16	8	18	5	20	8	22	10	24	12

On vous a chargé de comparer certaines grandeurs.

1. Calculer un indice simple pour le prix d'une montre C en 2006 par rapport au prix de cette même montre en 2003.
2. Calculer un indice pour le prix d'un ensemble de montres vendues en 2007 par rapport à l'année 2003 :
   (a) en considérant les quantités du temps de référence (c'est-à-dire de l'année 2003 dans ce cas). Dans ce cas, on veut travailler avec les quantités fixées à l'année de référence et on utilise donc l'indice de Laspeyres.
   (b) en considérant les quantités de l'année par rapport à laquelle on veut calculer l'indice (2007 dans ce cas). Dans ce cas, on veut travailler avec les quantités observées à l'année d'étude et on utilise donc l'indice de Paasche.
3. Même question mais en considérant les années 2006 et 2004.
4. Calculer l'indice de Fisher $F(2006/2004)$.
5. Calculer l'indice de Sidgwick $S(2006/2004)$.
6. On a déjà calculé :
$$L(2007/2006) = 109.133, \quad L(2006/2005) = 107.377,$$
$$L(2005/2004) = 108.108, \quad L(2004/2003) = 106.542.$$
   Calculer l'indice chaîne CL(2007/2003).

Solution

1. On calcule l'*indice simple* :
$$I(t/t') = 100 \times \frac{P_t}{P'_t}.$$
   Donc,
$$I(2006/2003) = 100 \times \frac{22}{16} = 137.5.$$
2. (a) On calcule ici l'*indice de Laspeyres* :
$$L(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_{0i} p_{ti}}{\sum_{i=1}^{n} q_{0i} p_{0i}}.$$
   On a
$$L(2007/2003) = 100 \times \frac{10 \times 15 + 12 \times 20 + 8 \times 24}{10 \times 12 + 12 \times 15 + 8 \times 16} = 135.981.$$
   (b) On calcule ici l'*indice de Paasche* :
$$P(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_{ti} p_{ti}}{\sum_{i=1}^{n} q_{ti} p_{0i}}.$$
   On a
$$P(2007/2003) = 100 \times \frac{12 \times 15 + 8 \times 20 + 12 \times 24}{12 \times 12 + 8 \times 15 + 12 \times 16} = 137.719.$$

3. (a)

$$L(2006/2004) = 100 \times \frac{10 \times 14 + 10 \times 18 + 5 \times 22}{10 \times 12 + 10 \times 16 + 5 \times 18} = 116.216.$$

(b)

$$P(2006/2004) = 100 \times \frac{15 \times 14 + 12 \times 18 + 10 \times 22}{15 \times 12 + 12 \times 16 + 10 \times 18} = 117.029.$$

4.

$$F(2006/2004) = \sqrt{L(2006/2004) \times P(2006/2004)} = \sqrt{116.216 \times 117.029} = 116.622.$$

5.

$$S(2006/2004) = \frac{L(2006/2004) + P(2006/2004)}{2} = \frac{116.216 + 117.029}{2} = 116.623.$$

6.

$$\begin{aligned} & CL(2007/2003) \\ = \quad & 100 \times \frac{L(2007/2006)}{100} \times \frac{L(2006/2005)}{100} \times \frac{L(2005/2004)}{100} \times \frac{L(2004/2003)}{100} \\ = \quad & 100 \times \frac{109.133}{100} \times \frac{107.377}{100} \times \frac{108.108}{100} \times \frac{106.542}{100} = 134.973. \end{aligned}$$

C'est un indice chaine avec la propriété de circularité. Par exemple,

$$100 \times CL(2007/2003) = CL(2007/2005) \times CL(2005/2003).$$

Exercice 10.6. Indice et lait
Soit le Tableau 10.6 contenant l'évolution des prix du lait et des quantités vendues (en litre) sur 5 ans, par un petit point de vente.

Table 10.6 – Évolution des prix du lait

	1980		1981		1982		1983		1984	
	Prix	Qté	Prix	Qté	Prix	Qté	Prix	Qté	Prix	Qté
Lait pasteurisé	1.50	2500	1.50	2000	1.55	2600	1.55	2800	1.58	2800
Lait UHT	1.40	3000	1.40	4000	1.45	3800	1.45	3900	1.46	3500
Lait bio	1.70	200	1.65	1000	1.65	1200	1.65	1400	1.68	1200

UHT = Upérisation Hautes températures

1. Calculer l'indice simple pour le prix du lait UHT en 1982 par rapport au prix de ce lait en 1980.
2. Calculer l'indice de Paasche, $P(1982/1980)$.
3. Calculer l'indice de Laspeyres pour la même période. Commenter la différence entre l'indice de Paasche et l'indice de Laspeyres.
4. Calculer l'indice de Fisher pour la même période.
5. Calculer l'indice de Sidgwick pour la même période.
6. Calculer l'indice en chaîne $CL(1984/1980)$.

Solution

1. Indice simple :
   $I(t/t') = 100 \times \dfrac{p_t}{p_{t'}}$.
   $I(1982/1980) = 100 \times \dfrac{\text{prix en } 1982}{\text{prix en } 1980} = 100 \times \dfrac{1.45}{1.40} = 100 \times 1.035 = 103.5.$
2. Indice de Paasche :
   $P(t/0) = 100 \times \dfrac{\sum_{i=1}^{n} q_{ti} p_{ti}}{\sum_{i=1}^{n} q_{ti} p_{0i}}$.
   $P(1982/1980) = 100 \times \dfrac{2600 \times 1.55 + 3800 \times 1.45 + 1200 \times 1.65}{2600 \times 1.50 + 3800 \times 1.40 + 1200 \times 1.70} = 100 \times \dfrac{11520}{11260} = 102.3.$
3. Indice de Laspeyres :
   $L(t/0) = 100 \times \dfrac{\sum_{i=1}^{n} q_{0i} p_{ti}}{\sum_{i=1}^{n} q_{0i} p_{0i}}$.
   $L(1982/1980) = 100 \times \dfrac{2500 \times 1.55 + 3000 \times 1.45 + 200 \times 1.65}{2500 \times 1.50 + 3000 \times 1.40 + 200 \times 1.70} = 100 \times \dfrac{8555}{8290} = 103.2.$

   En général l'indice de Laspeyres est plus grand que l'indice de Paasche, car l'indice de Laspeyres est une moyenne arithmétique et l'indice de Paasche une moyenne harmonique.
4. Indice de Fisher :
   $F(1/0) = \sqrt{L(1/0) \times P(1/0)}$.
   $F(1982/1980) = \sqrt{103.5 \times 102.31} = \sqrt{10589.085} = 102.9.$
5. Indice de Sidgwick :
   $S(t/0) = \dfrac{L(1/0) + P(1/0)}{2}$.
   $S(1982/1980) = \dfrac{103.5 + 102.31}{2} = \dfrac{205.81}{2} = 102.9.$
6. Indice chaîne :
   $$CL(t/0) = 100 \times \frac{L(t/t-1)}{100} \times \frac{L(t-1/t-2)}{100} \times \cdots \times \frac{L(1/0)}{100}.$$
   $$CL(1984/1980) = 100 \times \frac{L(1984/1983)}{100} \times \frac{L(1983/1982)}{100} \times \frac{L(1982/1981)}{100} \times \frac{L(1981/1980)}{100}.$$

$$
\begin{aligned}
L(1984/1983) &= 100 \times \frac{2800 \times 1.58 + 3900 \times 1.46 + 1400 \times 1.68}{2800 \times 1.55 + 3900 \times 1.45 + 1400 \times 1.65} \\
&= 100 \times \frac{4424 + 5694 + 2352}{4340 + 5655 + 2310} = 100 \times \frac{12470}{12305} = 101.3.
\end{aligned}
$$

$$
\begin{aligned}
L(1983/1982) &= 100 \times \frac{2600 \times 1.55 + 3800 \times 1.45 + 1200 \times 1.65}{2600 \times 1.55 + 3800 \times 1.45 + 1200 \times 1.65} \\
&= 100 \times \frac{4030 + 5510 + 1980}{4030 + 5510 + 1980} = 100 \times \frac{11520}{11520} = 100.0.
\end{aligned}
$$

$$
\begin{aligned}
L(1982/1981) &= 100 \times \frac{2000 \times 1.55 + 4000 \times 1.45 + 1000 \times 1.65}{2000 \times 1.50 + 4000 \times 1.40 + 1000 \times 1.65} \\
&= 100 \times \frac{3100 + 5800 + 1650}{3000 + 5600 + 1650} = 100 \times \frac{10550}{10250} = 102.9.
\end{aligned}
$$

$$
\begin{aligned}
L(1981/1980) &= 100 \times \frac{2500 \times 1.50 + 3000 \times 1.40 + 200 \times 1.65}{2500 \times 1.50 + 3000 \times 1.40 + 200 \times 1.70} \\
&= 100 \times \frac{3750 + 4200 + 330}{3750 + 4200 + 340} = 100 \times \frac{8280}{8290} = 99.9.
\end{aligned}
$$

$$
\begin{aligned}
CL(1984/1980) &= 100 \times \frac{101.3}{100} \times \frac{100.0}{100} \times \frac{102.9}{100} \times \frac{99.9}{100} \\
&= 100 \times \frac{101.3 \times 100.0 \times 102.9 \times 99.9}{100^4} \\
&= 100 \times \frac{104\,142\,968.292}{100\,000\,000} = 104.1.
\end{aligned}
$$

Exercice 10.7. Indice chaîne
On considère l'indice chaîne suivant :

$$CL(t/t') = 100 \times \frac{L(t/t-1)}{100} \times \frac{L(t-1/t-2)}{100} \times \cdots \times \frac{L(t'+1/t')}{100},$$

où L désigne l'indice de Laspeyres.

1. Montrer que pour $t > t_1 > t_2$ on a $CL(t/t_1) \times CL(t_1/t_2) = 100 \times CL(t/t_2)$.
2. On considère $t > t' > 0$. On sait que $CL(t/t') = 125$ et que $CL(t/0) = 150$. Calculer $CL(t'/0)$ en utilisant ce que vous avez démontré au point précédent.

Solution

1. On a
$$CL(t/t_1) = 100 \times \frac{L(t/t-1)}{100} \times \cdots \times \frac{l(t_1+1/t_1)}{100}$$
et
$$CL(t_1/t_2) = 100 \times \frac{L(t_1/t_1-1)}{100} \times \cdots \times \frac{l(t_2+1/t_2)}{100}.$$
Ainsi, on a
$$\begin{aligned} & CL(t/t_1) \times CL(t_1/t_2) \\ = \quad & 100 \times \underbrace{100 \times \frac{L(t/t-1)}{100} \times \cdots \times \frac{l(t_1+1/t_1)}{100} \times \frac{L(t_1/t_1-1)}{100} \times \cdots \times \frac{l(t_2+1/t_2)}{100}}_{CL(t/t_2)} \\ = \quad & 100 \times CL(t/t_2). \end{aligned}$$
2. On sait par le point précédent que : $CL(t/t') \times CL(t'/0) = 100 \times CL(t/0)$. Ainsi, on a
$$CL(t'/0) = 100 \times \frac{CL(t/0)}{CL(t/t')} = 100 \times \frac{150}{125} = 120.$$

Exercice 10.8. Courbe de Lorenz et inégalités
On trouve dans le Tableau 10.7 les revenus annuels en milliers de francs de 200 personnes travaillant dans une entreprise qui comporte une grille de salaire très simple, avec seulement cinq montants différents. On note x_j le revenu de la classe j.

Table 10.7 – Revenus annuels en milliers de francs de 200 personnes

Revenu x_j	Effectif n_j	Fréquence f_j	Fréquence cumulée F_j	Revenu total de la classe j $x_j n_j$
10	82			
24	68			
40	26			
50	20			
100	4			
–	200	1	–	

On a également les informations suivantes : $\bar{x} = 24.46$, $\sum_{i=1}^{n} i \times x_{(i)} = 660\,098$, $\sum_{i=1}^{n} |x_i - \bar{x}| = 2434$.

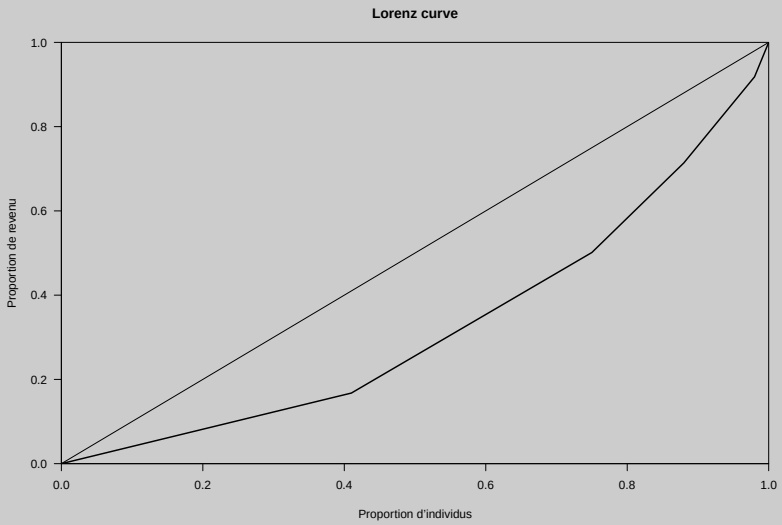
1. Completer le tableau ci-dessus.
2. Représenter graphiquement la courbe de Lorenz.
3. Calculer les indices de Gini et de Hoover et analyser le niveau d'inégalité dans la répartition des revenus.

Solution

1. Tableau complété :

x_j	n_j	f_j	F_j	$x_j n_j$	$\sum_{k=1}^{j} x_k n_k$	q_j
10	82	0.41	0.41	820	820	0.168
24	68	0.34	0.75	1632	2452	0.502
40	26	0.13	0.88	1040	3492	0.714
50	20	0.10	0.98	1000	4492	0.918
100	4	0.02	1	400	4892	1
	200	1	–	4892	–	–

2. Calcul des $q_j = \dfrac{\sum_{k=1}^{j} x_k n_k}{\sum_{k=1}^{n} x_k n_k}$ (proportion du revenu total gagné par les employés des classes 1 à j) et dessin de la courbe de Lorenz : relier les points (F_j, q_j), en commençant au point $(0, 0)$.



3. L'indice de Gini :

$$G = \frac{1}{n-1}\left(\frac{2\sum_{i=1}^{n} i \times x_{(i)}}{n\bar{x}} - (n+1)\right) = \frac{1}{199}\left(\frac{2 \times 660098}{200 \times 24.46} - 201\right) = 0.346$$

G plus proche de 0 que de 1 donc pas trop inégalitaire. Pas très loin du niveau d'un pays comme la Suisse.

$$H = \frac{\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|}{2\bar{x}} = \frac{\frac{1}{200}2434}{2 \times 24.46} = 0.249$$

Les conclusions sont les mêmes que pour G.

Exercice 10.9. Indice de Gini
Un institut de statistique fait annuellement une étude sur les emplois dans le secteur de Tourisme. Le Tableau 10.8 contient le nombre annuel moyen d'heures travaillées et le salaire horaire moyen (en dollars par heure) pour l'année 2002 pour 5 secteurs de tourisme :

TABLE 10.8 – Tableau année 2002

Secteur	Heures travaillées	Salaire par heure
Restauration	1447	10.31
Hébergement	1724	12.87
Loisirs	1514	17.66
Agences de voyages	1787	18.06
Transport	1800	25.46

1. Calculer le revenu total par secteur.
2. Calculer l'indice de Gini et l'indice de Hoover pour le revenu. La formule à utiliser est :

$$G_1 = \frac{1}{n-1}\left[\frac{2\sum_{i=1}^{n} i x_{(i)}}{\sum_{i=1}^{n} x_{(i)}} - (n+1)\right].$$

3. Démonstration en R - courbe de Lorenz et indice de Gini. À noter que le package ineq de R calcule le coefficient de Gini en utilisant la formule :

$$G_2 = \left[\frac{2\sum_{i=1}^{n} i x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}} - 1 - \frac{1}{n}\right] = G_1\frac{n-1}{n}.$$

Solution

1. Le tableau complet :

Tableau année 2002

Secteur	Heures	Salaire p.h.	Revenu total	$i \times$ Revenu	$\lvert x_i - \bar{x}\rvert$
Restauration	1447	10.31	14918.57	14918.57	13470.412
Hébergement	1724	12.87	22187.88	44375.76	6201.102
Loisirs	1514	17.66	26737.24	80211.72	1651.742
Agences	1787	18.06	32273.22	129092.88	3884.238
Transport	1800	25.46	45828	229140	17439.018
Somme			141944.91	497738.93	42646.512
Moyenne			28388.982		8529.3024

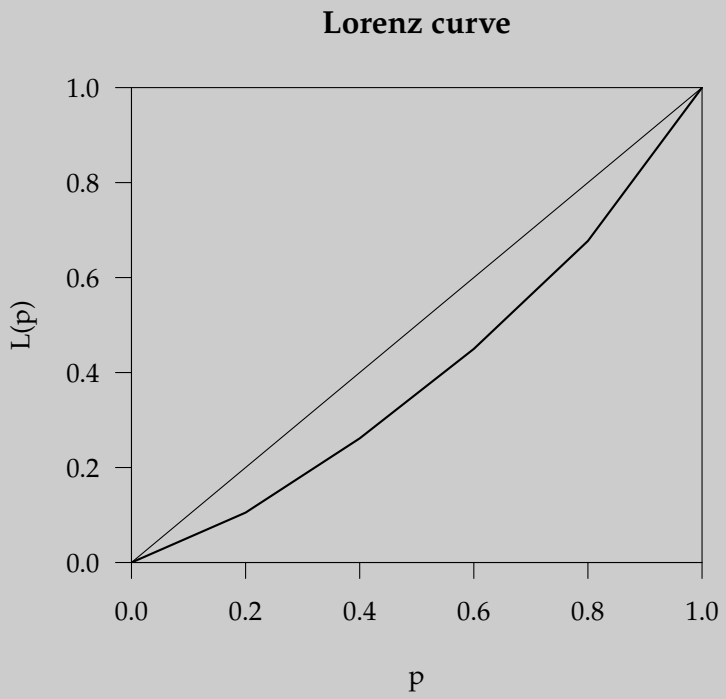
2.
$$G = \frac{1}{4}\left(\frac{2\sum_{i=1}^{5} i x_{(i)}}{\sum_{i=1}^{5} x_{(i)}} - 6\right) = \frac{1}{4}\left(\frac{2\times 497738.93}{141944.91} - 6\right) = \frac{1}{4}\times 1.013 = 0.253.$$

3.
$$H = \frac{\frac{1}{5}\sum_{i=1}^{5}\lvert x_i - \bar{x}\rvert}{2\bar{x}} = \frac{\frac{1}{5}\times 42646.512}{2\times 28388.982} = \frac{8529.302}{56777.964} = 0.15.$$

4. **En langage R**

```
install.packages("ineq")
library(ineq)
X=rbind(Restauration=c(1447,10.31),Hébergement=c(1724,12.87),
Loisirs=c(1514,17.66),Agences_de_voyages=c(1787,18.06),
Transport=c(1800,25.46))
X=as.data.frame(X)
colnames(X)=c("Heures","Salaire_heure")
```

```
Revenu_total=X[,1]*X[,2]
tikz(file="Lorenz.tex",width=4,height=4)
plot(Lc(Revenu_total))
dev.off()
G2=Gini(Revenu_total)
n=length(Revenu_total)
G1=G2*n/(n-1)
G1
G2
```



Exercice 10.10. Indices d'inégalité et courbe de Lorenz
Vous trouverez dans le Tableau 10.9 les revenus annuels en milliers de francs suisses de 100 personnes et dans la Figure 10.1 la courbe de Lorenz associée.
Notons X le revenu. On sait que :

$$n = 100, \bar{x} = 64.6, \sum_{i=1}^{n} i \times x_{(i)} = 384\,027, \sum_{i=1}^{n} |x_i - \bar{x}| = 1744.8.$$

En se basant sur la courbe de Lorenz, dire si les revenus sont distribués de manière égalitaire.
Confirmer votre réponse en calculant l'indice de Gini et l'indice de Hoover.

Table 10.9 – Revenu annuel en milliers de francs suisses de 100 personnes

Revenu	Effectif
13	5
39	24
65	49
78	18
144	4

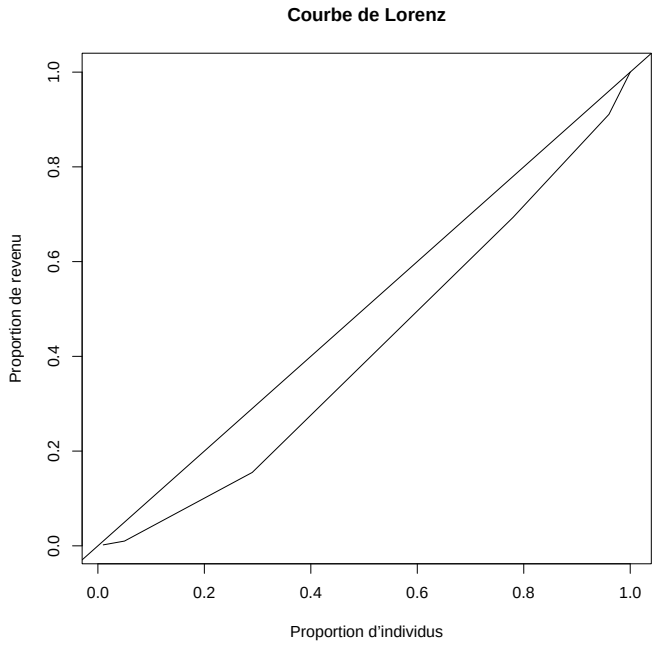


Figure 10.1 – Courbe de Lorenz

Solution
La courbe de Lorenz est relativement proche de la diagonale. Cela signifie que les revenus sont relativment équitablement répartis.

$$G = \frac{1}{n-1}\left(\frac{2\sum_{i=1}^{n} i \times x_{(i)}}{n\bar{x}} - (n+1)\right) = \frac{1}{99}\left(\frac{2 \times 384027}{100 \times 64.6} - 101\right) = 0.181$$

G est relativement proche de 0. Donc, les revenus sont plutôt également distribués

$$H = \frac{\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|}{2\bar{x}} = \frac{\frac{1}{100}1744.8}{2 \times 64.6} = 0.135.$$

Même conclusion que pour G.

Exercice 10.11. Comparaison de courbes de Lorenz
La Figure 10.2 représente les courbes de Lorenz des revenus de trois pays, à savoir A, B et C.

1. Classer les indices de Gini et de Hoover de ces 3 pays dans l'ordre croissant en se basant sur cette figure.
2. Dans quel pays les revenus sont-ils plus inégalement répartis ?
3. Quel pays est le plus riche en termes de revenu moyen par individu ?
4. La représentation graphique des courbes de Lorenz permet-elle toujours d'affirmer que la distribution du revenu dans une population est plus inégalitaire que dans une autre ?

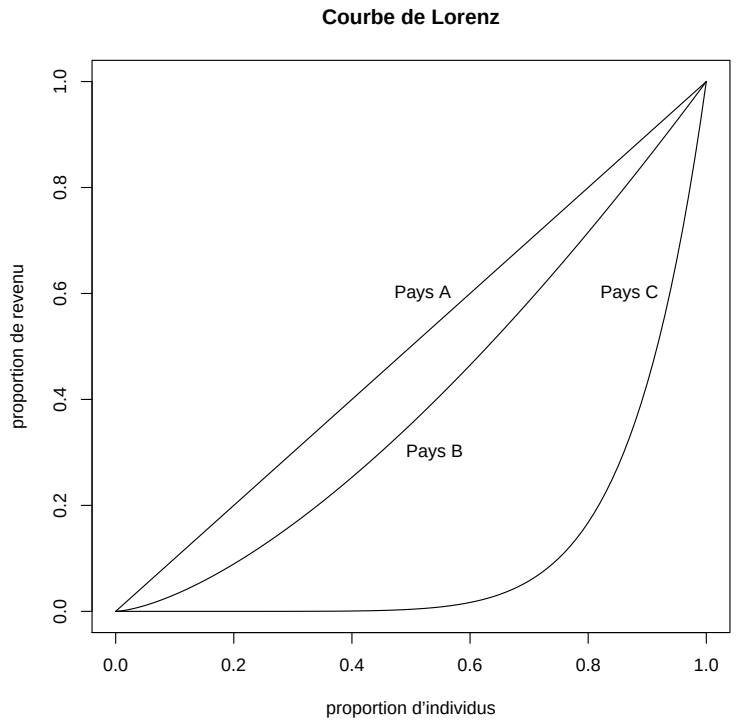


FIGURE 10.2 – Courbes de Lorenz

Solution

1. Indice de Gini = 2 fois la surface comprise entre la courbe de Lorenz et la diagonale.

$$G_A < G_B < G_C$$

   Indice de Hoover = la plus grande distance verticale entre la courbe de Lorenz et la diagonale.

$$H_A < H_B < H_C$$

2. Les revenus sont le plus inégalement répartis dans le pays C.
3. Impossible à dire. La courbe de Lorenz ne donne aucune indication sur le revenu absolu.
4. Non. Les courbes de Lorenz peuvent être confondues ou, surtout, se croiser.

Exercice 10.12. Calcul d'indices d'inégalité
Soit le Tableau 10.10, qui représente l'argent à disposition par mois, pour un ensemble de 100 étudiants.

TABLE 10.10 – Argent à disposition par mois pour 100 étudiants

Revenu x_j	Effectif n_j	Fréquence f_j	Fréquence cumulé F_j	Revenu total de la classe $x_j n_j$
200	30			
300	10			
500	30			
800	20			
1000	10			

Nous avons les résultats suivants :

$$\sum_{i=1}^{n} i x_i = 3275000, \frac{1}{n}\sum_{i=1}^{n} x_i = 500, \sum_{i=1}^{n} |x_i - \bar{x}| = 22000$$

1. Compléter le tableau.
2. Dessiner la courbe de Lorenz.
3. Calculer l'indice de Gini.
4. Calculer l'indice de Hoover.
5. Calculer le Decile Share Ratio (DSR).

Solution

1. Le tableau :

Revenu x_j	Effectif n_j	Fréquence f_j	Fréquence cumulé F_j	Revenu total de la classe $x_j n_j$
200	30	0.30	0.30	6 000
300	10	0.10	0.40	3 000
500	30	0.30	0.70	15 000
800	20	0.20	0.90	16 000
1000	10	0.10	10	10 000

2. Courbe de Lorenz :

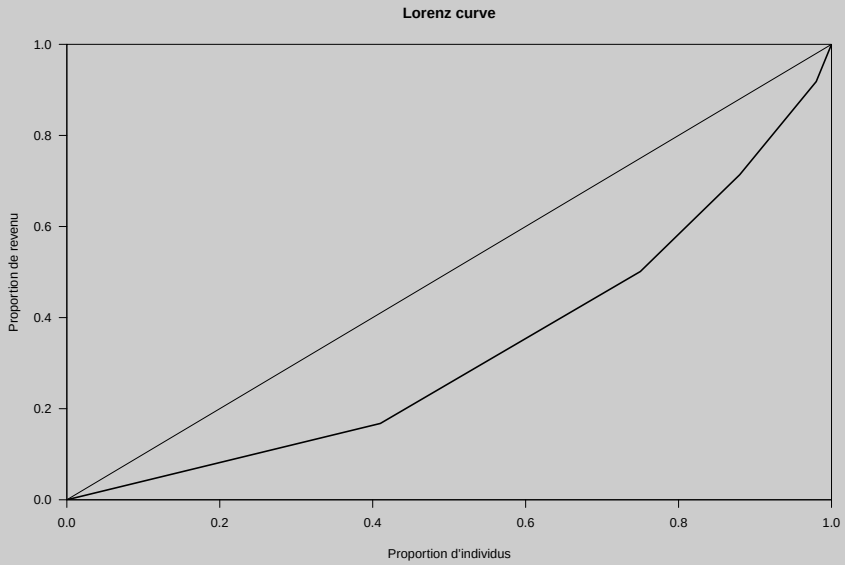
$$q_j = \frac{\text{revenu total cumulé de la classe}}{\text{somme des revenus totaux cumulés}}$$

Exemple pour la classe de revenu de $x_j = 300$:

$$q_j = \frac{\sum_{j=1}^{2} x_j n_j}{\sum_{j=1}^{5} x_j n_j} = \frac{6000 + 3000}{6000 + 3000 + 15000 + 16000 + 10000} = \frac{9000}{50000} = 0.18.$$

Revenu x_j	Fréquence cummulé F_j	Revenus totals cumulés $\sum_{k=1}^{j} x_k n_k$	Proportion revenus cumulés q_j
200	0.30	6000	0.12
300	0.40	9000	0.18
500	0.70	24000	0.48
800	0.90	40 000	0.80
1000	1.00	50 000	1.00

Pour dessiner la courbe de Lorenz, il faut relier les points (F_j, q_j).



3. Indice de Gini

$$G = \frac{1}{n-1}\left(\frac{2\sum_{i=1}^{n} i x_i}{n\bar{x}} - (n+1)\right) = \frac{1}{100-1}\left(\frac{2\times 3275000}{100\times 500} - (100+1)\right) = 0.30$$

4. Indice de Hoover

$$H = \frac{\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|}{2\bar{x}} = \frac{\frac{1}{100}\times 22000}{2\times 500} = 0.22.$$

5. Decile share ratio :

$$DSR = \frac{S_{90}}{S_{10}}.$$

S_{10} : $np = \frac{1}{10}\times n = \frac{1}{10}\times 100 = 10$. np est un nombre entier. Donc,

$$x_{1/10} = \frac{1}{2}\times(x_{(10)} + x_{(11)}) = \frac{1}{2}\times(200+200) = 200.$$

On cherche ensuite toutes les personnes ayant un revenu inférieur à $x_{1/10}$:

$$S_{10} = \frac{1}{9}\times\sum_{i=1}^{9} x_{(i)} = \frac{1}{9}\times 1800 = 200.$$

S_{90} : $np = \frac{9}{10}\times n = \frac{9}{10}\times 100 = 90$. np est un nombre entier. Donc,

$$x_{9/10} = \frac{1}{2}\times(x_{(90)} + x_{(91)}) = \frac{1}{2}\times(800+1000) = 900.$$

On cherche ensuite toutes les personnes ayant un revenu supérieur à $x_{9/10}$:

$$S_{90} = \frac{1}{10}\times\sum_{i=91}^{100} x_{(i)} = \frac{1}{10}\times 10000 = 1000.$$

Finalement,

$$\frac{S_{90}}{S_{10}} = \frac{1000}{200} = 5.00.$$

Cela signifie que le revenu moyen des 10% les plus riches est 5 fois plus grand que celui des 10% les plus pauvres.

Exercice 10.13. Revenus dans les États
Le Bureau d'analyse économique du Service de commerce des États-Unis donne des statistiques annuelles sur le revenu des personnes dans les différents États. Dans le Tableau 10.11, on a le revenu moyen par personne et par État (en dollars), en ordre croissant, pour l'année 2004. Imaginez que les États-Unis aient à l'époque adopté une loi qui impose au gouverneur d'un État de gagner exactement le salaire moyen de cet État. Nous nous intéressons aux indices d'inégalité dans la population des gouverneurs d'État.

Table 10.11 – Tableau des revenus 2004

État	Revenu	État	Revenu	État	Revenu
Mississippi	24518	Maine	30046	Alaska	34000
West Virginia	25792	Indiana	30204	Rhode Island	34207
Arkansas	25814	South Dakota	30209	Wyoming	34279
New Mexico	26184	Missouri	30475	Illinois	34721
Utah	26603	Oregon	30561	Washington	35041
Idaho	26877	Texas	30732	California	35219
South Carolina	27185	Iowa	31058	Delaware	35728
Kentucky	27265	Kansas	31078	Colorado	36113
Louisiana	27297	Ohio	31161	Virginia	36160
Montana	27657	Florida	31469	Minnesota	36184
Alabama	27695	Vermont	31780	New Hampshire	36616
Oklahoma	27840	Michigan	32079	New York	38264
Arizona	28658	Wisconsin	32166	Maryland	39631
North Carolina	29322	Nebraska	32341	New Jersey	41626
North Dakota	29494	Hawaii	32625	Massachusetts	42176
Georgia	29782	Pennsylvania	33312	Connecticut	45318
Tennessee	29844	Nevada	33787	District of Columbia	51155

1. En sachant que

$$\sum_{i=1}^{n} x_{(i)} = 1\,649\,348, \quad \sum_{i=1}^{n} i x_{(i)} = 46\,454\,336, \text{ et } \sum_{i=1}^{n} |x_i - \bar{x}| = 198\,719.412,$$

   calculer les indices de Gini et de Hoover pour le revenu.
2. Calculer les *QSR* et *DSR* et donnez leur interprétation.
3. Démonstration en R.

Solution

1. Pour l'indice de Gini, on peut utiliser la formule :

$$G = \frac{1}{n-1}\left[\frac{2\sum_{i=1}^{n} i x_{(i)}}{\sum_{i=1}^{n} x_{(i)}} - (n+1)\right].$$

$$G = \frac{1}{50}\left(\frac{2\sum_{i=1}^{51} i x_{(i)}}{\sum_{i=1}^{51} x_{(i)}} - 52\right) = \frac{1}{50}\left(\frac{2\times 46454336}{1649348} - 52\right) = \frac{1}{50}\times 4.331 = 0.087.$$

2. L'indice de Hoover :

$$H = \frac{\frac{1}{51}\sum_{i=1}^{51}|x_i - \bar{x}|}{2\bar{x}} = \frac{\frac{1}{51}\times 198719.412}{2\times 32340.157} = \frac{3896.46}{64680.314} = 0.06.$$

3. Quantile et decile share ratio :
   Le premier décile :
   Comme $np = \frac{1}{10}\times 51 = 5.1$ n'est pas un nombre entier, on a

$$x_{1/10} = x_{(\lceil 5.1 \rceil)} = x_{(6)} = 26877.$$

Donc,

$$S_{10} = \frac{1}{5}\sum_{i=1}^{5} x_{(i)} = \frac{1}{5} \times 128911 = 25782.2$$

Le deuxième décile :
Comme $np = \frac{1}{5} \times 51 = 10.2$ n'est pas un nombre entier, on a

$$x_{1/5} = x_{(\lceil 10.2 \rceil)} = x_{(11)} = 27695.$$

Donc,

$$S_{20} = \frac{1}{10}\sum_{i=1}^{10} x_{(i)} = \frac{1}{10} \times 265192 = 26519.2.$$

Le huitième décile :
Comme $np = \frac{8}{10} \times 51 = 40.8$ n'est pas un nombre entier, on a

$$x_{8/10} = x_{(\lceil 40.8 \rceil)} = x_{(41)} = 35728.$$

Donc,

$$S_{80} = \frac{1}{10}\sum_{i=42}^{51} x_{(i)} = \frac{1}{10} \times 403243 = 40324.3.$$

Le neuvième décile :
Comme $np = \frac{9}{10} \times 51 = 45.9$ n'est pas un nombre entier, on a

$$x_{9/10} = x_{(\lceil 45.9 \rceil)} = x_{(46)} = 38264.$$

Donc,

$$S_{90} = \frac{1}{5}\sum_{i=47}^{51} x_{(i)} = \frac{1}{5} \times 219906 = 43981.2.$$

Donc,

$$QSR = \frac{S_{80}}{S_{20}} = \frac{40324.3}{26519.2} = 1.521.$$

Le revenu moyen de 20% des plus riches états est 1.521 fois plus grand que le revenu moyen de 20% des plus pauvres.

$$DSR = \frac{S_{90}}{S_{10}} = \frac{43981.2}{25782.2} = 1.706.$$

Le revenu moyen de 10% des plus riches états est 1.706 fois plus grand que le revenu moyen de 10% des plus pauvres.

Chapitre 11

Exercices : Séries temporelles

Exercice 11.1. Opérateurs de décalage 1
Le Tableau 11.1 représente le nombre de téléphones portables utilisés dans un ménage au fil des ans.

Table 11.1 – Nombre de téléphones portables

Année t	Nombre téléph. Y_t	Ly_t	F^2y_t
0	-		
1	1		
2	1		
3	2		
4	3		
5	4		
6	-		

1. Compléter le Tableau 11.1.
2. On suppose que cette serie présente une tendance linéaire, modélisée par : $y_t = a + bt + E_t$. Créer une nouvelle colonne de données en enlevant cette tendance linéaire.

Solution

1. $Ly_t = y_{t-1}$ et $F^2y_t = y_{t+2}$.
2. Pour enlever la tendance linéaire, on utilise l'opérateur différence d'ordre 1 : $\nabla = I - L$:

$$\nabla y_t = Iy_t - Ly_t = y_t - y_{t-1}$$

Année t	Nombre téléph. Y_t	Ly_t	F^2y_t	$y_t - y_{t-1}$
0	0	-	1	-
1	1	0	2	1
2	1	1	3	0
3	2	1	4	1
4	3	2	-	1
5	4	3	-	1
6	-	4	-	-

Exercice 11.2. Opérateurs de décalage 2
Dans le Tableau 11.2 se trouve la série temporelle représentant le nombre d'élèves dans le degré secondaire 1, dans le canton de Neuchâtel (Source : OFS 2009).

TABLE 11.2 – Nombre d'élèves dans le degré secondaire 1

Année	t	y_t	$L^2 y_t$	$F y_t$
1997	0	-		
1998	1	7122		
1999	2	7226		
2000	3	7417		
2001	4	7629		
2002	5	7832		
2003	6	8023		
2004	7	8190		
2005	8	8114		
2006	9	8050		
2007	10	7950		
2008	11	-		
2009	12	-		

1. Remplir ce tableau.
2. Ajouter dans ce tableau les deux séries temporelles $L^2 y_t$ et $F y_t$.

Solution

Année	t	y_t	$L^2 y_t$	$F y_t$
1997	0	-	-	7122
1998	1	7122	-	7226
1999	2	7226	-	7417
2000	3	7417	7122	7629
2001	4	7629	7226	7832
2002	5	7832	7417	8023
2003	6	8023	7629	8190
2004	7	8190	7832	8114
2005	8	8114	8023	8050
2006	9	8050	8190	7950
2007	10	7950	8114	-
2008	11	-	8050	-
2009	12	-	7950	-

Exercice 11.3. Villas vendues
Le Tableau 11.3 indique le nombre de villas vendues par an par une agence immobilière en 10 ans.

Table 11.3 – Tableau des ventes

Année t	1	2	3	4	5	6	7	8	9	10
Ventes y_t	49	53	58	58	59	62	64	67	69	70

1. Représenter le nuage de points.

2. Faire un ajustement linéaire en fonction du temps.

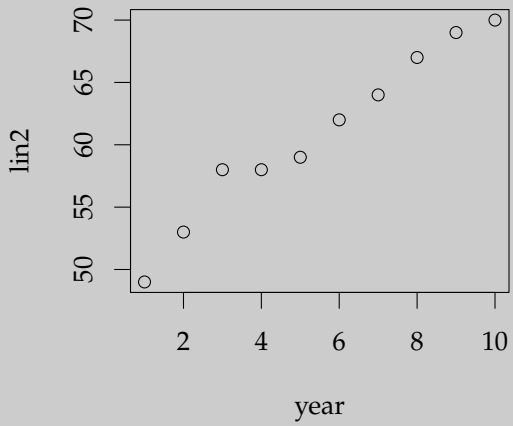
3. Appliquer un filtre linéaire (différence) sur la série et sur les résidus de la séries.

4. Représenter graphiquement la différence d'ordre un de la série.

5. Appliquer un filtre linéaire les résidus de la régression.

Solution

1. Le nuage de points :



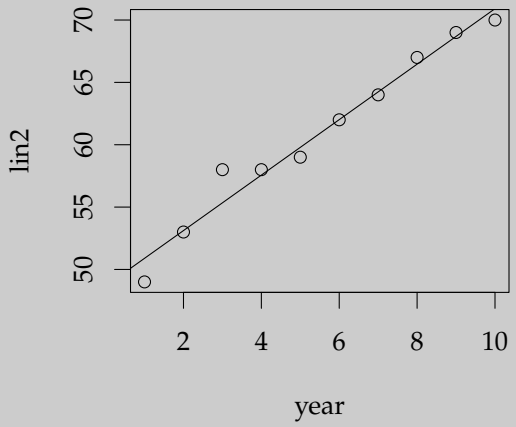
2. Régression linéaire (moindres carrés) : $y_t = a + bt$

$$\bar{t} = 5.5, \bar{y}_t = 60.9, s_t^2 = 8.25, s_{y_t}^2 = 42.09, s_{ty_t} = 18.35$$

$$r_{ty_t} = \frac{s_{ty_t}}{s_{y_t} s_t} = 0.985, r_{ty_t}^2 = 0.970,$$

$$b = \frac{s_{ty_t}}{s_t^2} = 2.224 \text{ et } a = \bar{y}_t - b\,\bar{t} = 60.9 - 5.5 \times 2.224 = 48.667.$$

Donc, $y_t = 48.667 + 2.224 \times t$. On peut dessiner la droite de régression sur le nuage de points.

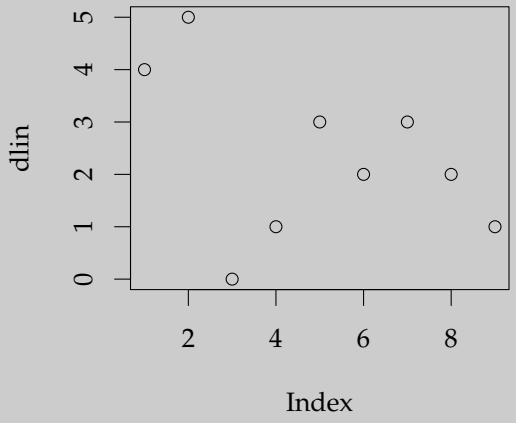


3. Filtre linéaire : $\nabla y_t = b + E_t - E_{t-1}$ ou $\nabla y_t = Iy_t - Ly_t = y_t - y_{t-1}$ (voir tableau).

Tableau des ventes

Année t	Ventes y_t	Résidus E_t	Filtre linéaire $\nabla y_t = b + E_t - E_{t-1}$
1	49	−1.891	–
2	53	−0.115	4
3	58	2.661	5
4	58	0.436	0
5	59	−0.788	1
6	62	−0.012	3
7	64	−0.236	2
8	67	0.539	3
9	69	0.315	2
10	70	−0.909	1

4. Différence d'ordre un de la série :



En langage R

```
library(tikzDevice)
#
lin2=c(49, 53, 58, 58, 59, 62, 64, 67, 69,70)
year=c(1,2,3,4,5,6,7,8,9,10)
#
tbar=mean(year)
```

```
Ttbar=mean(lin2)
s2t=var(year)*9/10
s2Tt=var(lin2)*9/10
covar=cov(lin2,year)*9/10
#
tikz(file="lin1.tex",width=3,height=3)
plot(year,lin2)
dev.off()
#
fit=lm(lin2~year)
summary(fit)
tikz(file="lin1dr.tex",width=3,height=3)
plot(year,lin2)
abline(coefficients(fit))
dev.off()
#
dlin=diff(lin2)
tikz(file="lin1diff.tex",width=3,height=3)
plot(dlin)
dev.off()
b=coefficients(fit)[2]
E=residuals(fit)
Edif=rep(0,times=9)
#
diff(E)+b
diff(lin2)
```

Exercice 11.4. Désaisonnalisation
Désaisonnalisez la série suivante (c'est une série trimestrielle sur 3 années)

$$2417, 1605, 1221, 1826, 2367, 1569, 1176, 1742, 2804, 1399, 1063, 1755$$

par la méthode additive, en utilisant une moyenne mobile d'ordre 4.

Solution
Il s'agit de

$$MA(4) = \frac{L^2 + 2L + 2I + 2F + F^2}{8}.$$

Nr.	Série	Trim.	MM(4)	Série-MM(4)	Desaison.		S	S'
1	2417	1			1589.53125	1	832.375	827.46875
2	1605	2			1864.71875	2	-254.8125	-259.71875
3	1221	3	1761	-540	1791.96875	3	-566.0625	-570.96875
4	1826	4	1750.25	75.75	1822.78125	4	8.125	3.21875
5	2367	1	1740.125	626.875	1539.53125	Total	19.625	0
6	1569	2	1724	-155	1828.71875			
7	1176	3	1768.125	-592.125	1746.96875			
8	1742	4	1801.5	-59.5	1738.78125			
9	2804	1	1766.125	1037.875	1976.53125			
10	1399	2	1753.625	-354.625	1658.71875			
11	1063	3			1633.96875			
12	1755	4			1751.78125			

Exercice 11.5. Décomposition et désaisonnalisation
Le Tableau 11.4 contient des données sur le nombre d'entrées de travailleurs saisonniers dans un pays (Y_t, en milliers). Les données concernent 16 trimestres consécutifs de 1991 à 1994. On pense que la série est de la forme

$$Y_t = T_t + S_t + e_t,$$

où T_t est la tendance, S_t est la composante saisonnière, qui ne dépend que du numéro de trimestre dans l'année et telle que la somme des S_t sur quatre trimestres consécutifs vaut 0 et les e_t sont des résidus.

Table 11.4 – Nombre d'entrées de travailleurs saisonniers dans un pays

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Y_t	51	111	123	59	78	162	146	54	83	175	174	88	95	185	197	75

1. Estimer la tendance T_t par une moyenne mobile $MM(4)$.
2. Désaisonnaliser la série par la méthode additive.
3. Représenter graphiquement les séries Y_t, T_t, $Y_t - T_t$ et $Y_t - S_t$.

Solution

1. On applique la moyenne mobile $MM(4) = (L^2 + 2L + 2I + 2F + F^2)/8$. Par exemple,

$$MM(4)y_3 = \frac{1}{8}(51 + 2 \times 111 + 2 \times 123 + 2 \times 59 + 78) = 89.375.$$

Nr	Y_t	$T_t = MM(4)Y_t$	$Y_t - T_t$
1	51	-	-
2	111	-	-
3	123	89.375	33.625
4	59	99.125	-40.125
5	78	108.375	-30.375
6	162	110.625	51.375
7	146	110.625	35.375
8	54	112.875	-58.875
9	83	118	-35
10	175	125.75	49.25
11	174	131.5	42.5
12	88	134.25	-46.25
13	95	138.375	-43.375
14	185	139.625	45.375
15	197	-	-
16	75	-	-

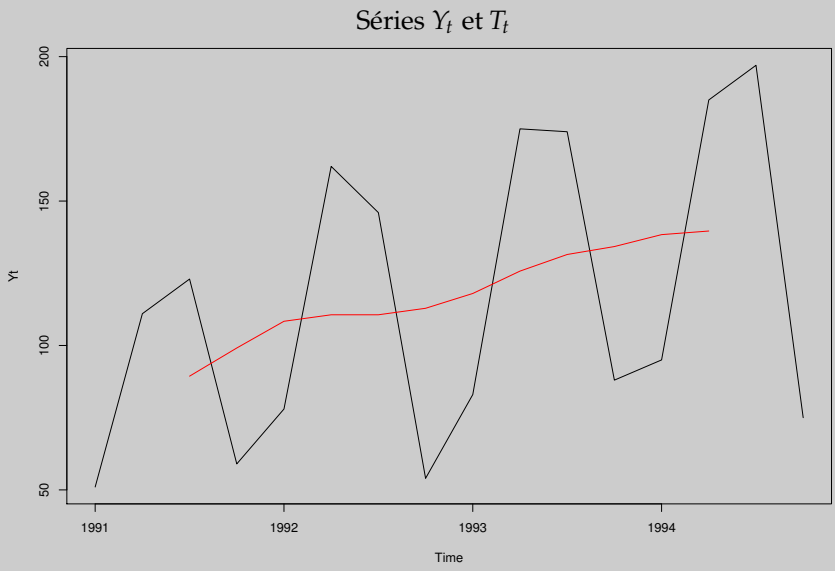
2. On cherche la composante saisonnière. On commence par enlever la tendance en calculant $Y_t - T_t$, puis, pour chaque trimestre, on fait la moyenne des valeurs observées. Les observations numérotées, $1, 5, 9, 13$ correspondent à un trimestre, le premier de l'année, les observations numérotées $2, 6, 10, 14$ au deuxième trimestre, les observations numérotées $3, 7, 11, 15$ au troisième et celles numérotées $4, 8, 12, 16$ au quatrième. Pour chaque trimestre on a trois observations disponibles de $Y_t - T_t$. On trouve :

$$S'_{1,5,9,13} = (-30.375 - 35 - 43.375)\frac{1}{3} = -36.25,$$

$$S'_{2,6,10,14} = 48.667,$$

$$S'_{3,7,11,15} = 37.167,$$

$$S'_{4,8,12,16} = -48.417.$$



On fait ensuite traditionnellement un ajustement de la série pour que la somme des composantes saisonnières soit nulle.

$$S_t = S'_t - \frac{1}{4}\sum_{i=1}^{4} S_i$$

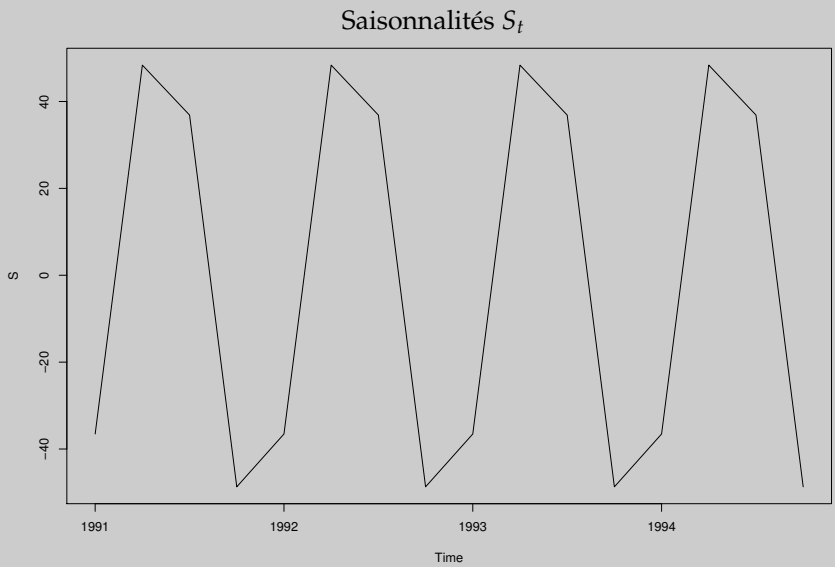
exemple :

$$S_1 = S'_1 - \frac{1}{4}\sum_{i=1}^{4} S'_t = -36.250 - \frac{1}{4} \times (-36.25 + 48.667 + 37.167 - 48.417) = -36.542.$$

On trouve aussi,

$$S_2 = 48.375, \; S_3 = 36.875, \text{ et } S_4 = -48.708.$$

Les S_t sont représentés dans la figure suivante :



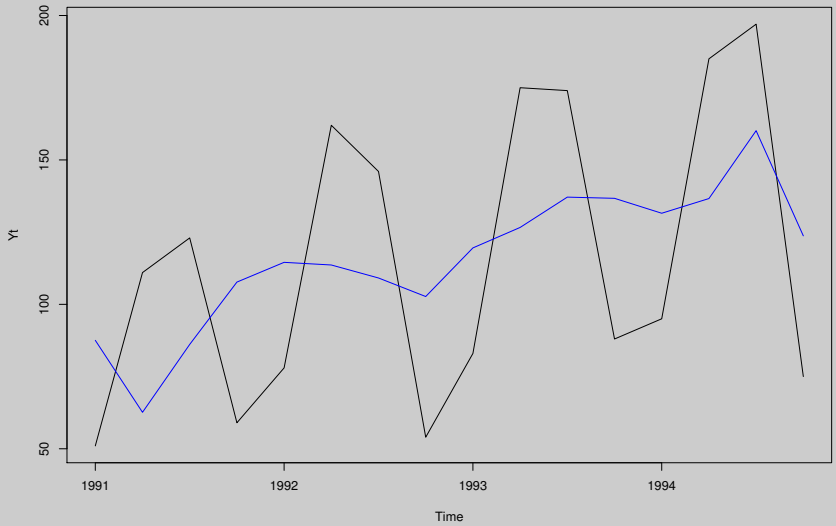
La série désaisonnalisée est la série obtenue en ôtant la composante saisonnière : $\tilde{Y}_t = Y_t - S_t$. Par exemple,

$$\tilde{Y_1} = Y_1 - S_1 = 51 - (-36.542) = 87.542.$$

Nr	Y_t	S_t	$Y_z - S_t$
1	51	-36.542	87.542
2	111	48.375	62.625
3	123	36.875	86.125
4	59	-48.708	107.708
5	78	-36.542	114.542
6	162	48.375	113.625
7	146	36.875	109.125
8	54	-48.708	102.708
9	83	-36.542	119.542
10	175	48.375	126.625
11	174	36.875	137.125
12	88	-48.708	136.708
13	95	-36.542	131.542
14	185	48.375	136.625
15	197	36.875	160.125
16	75	-48.708	123.708

La série Y_t et la série désaisonnalisée sont représentées dans la figure suivante :

Séries Y_t et $Y_t - S_t$



Chapitre 12

Exercices : Calcul des probabilités et variables aléatoires

Exercice 12.1. Cluedo
Cluedo est un jeu de société dans lequel les joueurs doivent découvrir parmi eux qui est le meurtrier d'un crime commis dans un manoir anglais : 6 suspects, 8 armes et 10 lieux sont possibles. De combien de manières différentes le meurtre a-t-il pu être commis ?

Solution
$6 \times 8 \times 10 = 480$.

Exercice 12.2. Cadenas
Paul a oublié le code de son cadenas à 4 boucles (nombre allant de 0 à 9 sur chaque boucle). il décide d'essayer tous les codes possibles. Combien de temps cherchera-t-il s'il teste toutes les possibilités avant de trouver le bon code si il a besoin de deux secondes pour tester un code ?

Solution

1. Nombre de codes à tester $10 \times 10 \times 10 \times 10 = 9^4 = 10000$.
2. Temps $10\,000 \times 2 = 20000$ secondes $= 333.3333$ minutes $= 5.555556$ heures $= 5h33\,\text{min}\,20$ sec.

Exercice 12.3. Cartes
On tire deux cartes dans un jeu de 36 cartes (qui comporte donc toutes les cartes du six à l'as). On considère les événements suivants :
$A = \{$ les deux cartes tirées sont rouges $\}$,
$B = \{$ les deux cartes tirées sont un valet et un dix$\}$,
$C = \{$ les deux cartes tirées sont des figures (valet, dame ou roi) $\}$.

1. Décrire en mots ce que représente les événements suivants :
   (a) $\overline{A}$,
   (b) $A \cap B \cap \overline{C}$,
   (c) $(A \cap \overline{C}) \cap (B \cap \overline{C})$,
   (d) $(A \cap B) \cap C$.
2. Ecrire, à l'aide des événements A, B, C, les événements :
   (a) $F = \{$les deux cartes tirées sont des figures et ne sont pas toutes les deux rouges$\}$.
   (b) $G = \{$on obtient au plus une figure$\}$.

Solution

1. (a) Au moins une des cartes tirées n'est pas rouge.
   (b) La réalisation des évènements A et B impliquent que les deux cartes tirées sont un valet rouge et un dix rouge. D'autre part, la réalisation de $\overline{C}$ implique qu'au moins une des deux cartes tirées n'est pas une figure, ce qui est le cas car on a pioché un "dix". Par conséquent la réalisation de l'évènement $A \cap B \cap \overline{C}$ est équivalente à la pioche d'un valet rouge et d'un dix rouge.
   (c) La réalisation de$A \cap \overline{C}$ signifie que l'on a tiré deux cartes rouges dont au moins une n'est pas une figure. Ensuite la réalisation de l'évènement $B \cap \overline{C}$ signifie que l'on a pioché un valet et un dix. (la condition $\overline{C}$ est redondante). Par conséquent, la réalisation de $(A \cap \overline{C}) \cap (B \cap \overline{C})$ signifie que l'on a pioché un valet rouge et un dix rouge.
   (d) L'évènement $(A \cap B) \cap C$ est impossible car la réalisation de $A \cap B$ implique la pioche d'un valet rouge et d'un dix rouge tandis que la réalisation de C implique la pioche de deux figures, ce qui est contradictoire avec $A \cap B$.
2. $F = C \cap \overline{A}$ et $G = \overline{C}$.

Exercice 12.4. Opinion
Une étude auprès de 210 personnes a été menée afin de connaître l'opinion sur deux quotidiens et de voir si le sexe a un effet sur cet opinion. On demandait à ces 210 individus s'ils préféraient le quotidien q_1 ou le quotidien q_2. Les résultats de cette étude étaient les suivants :

— 50 hommes ont déclaré préférer le quotidien q_1,
— 70 hommes ont déclaré préférer le quotidien q_2,
— 50 femmes ont déclaré préférer le quotidien q_1 et
— 40 femmes ont déclaré préférer le quotidien q_2.

On choisit au hasard une personne parmis les 210 interrogées. Considérons alors les évènements suivants :

— A = "la personne choisie est une femme",
— B = "la personne choisie préfère le quotidien q_2",
— C = "la personne choisie est un homme et préfère le quotidien q_1"et
— D = "la personne choisie est une femme et préfère le quotidien q_2".

1. Que signifient les évènements
   (a) $A \cup B$,
   (b) $A \cap B$,
   (c) $A \setminus B$,
   (d) $D \setminus A$,
   (e) $\overline{B}$ et
   (f) $\overline{C}$.
2. Ecrire à l'aide des événements A, B, C et D les événements :
   (a) La personne choisie est un homme et préfère le quotidien q_1,
   (b) La personne choisie est un homme.
3. (a) Donner la probabilité des évènements A, B et $A \cap B$.
   (b) Utiliser le point précédent pour donner la probabilité de l'évènement $A \cup B$.
   (c) Trouver un système complet d'évènements.

Solution

1. Donner les évènements
   (a) $A \cup B$
   = "la personne choisie est une femme ou la personne choisie préfère le quotidien q_2".
   Remarque : $D \subset A \cup B$.
   (b) $A \cap B$
   = "la personne choisie est une femme et préfère le quotidien q_2"
   =D
   (c) $A \setminus B$
   = "la personne choisie est une femme et ne préfère pas le quotidien q_2"
   = "la personne choisie est une femme et préfère le quotidien q_1"
   (d) $D \setminus A$
   = "la personne choisie est une femme et préfère le quotidien q_2 et la personne choisie n'est pas une femme"
   =$\emptyset$
   (e) $\overline{B}$
   = "la personne choisie ne préfère pas le quotidien q_2"
   = "la personne choisie préfère le quotidien q_1"
   (f) $\overline{C}$
   = "la personne choisie est une femme ou préfère le quotidien q_2"
   =$A \cup B$
2. (a) $\overline{A} \cap \overline{B}$
   (b) $\overline{A}$
3. (a) — $\Pr(A) = \dfrac{90}{210} = \dfrac{3}{7} = 0.43$
   — $\Pr(B) = \dfrac{110}{210} = \dfrac{11}{21} = 0.52$

— $\Pr(A \cap B) = \dfrac{40}{210} = \dfrac{4}{21} = 0.19$

(b)

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= \frac{3}{7} + \frac{11}{21} - \frac{4}{21} = \frac{9}{21} + \frac{11}{21} - \frac{4}{21} = \frac{16}{21} = 0.76\end{aligned}$$

(c) Systèmes complets d'évènements :
— E_1= "la personne choisie est une femme"
E_2= "la personne choisie est un homme"
— F_1= "la personne choisie préfère le quotidien q_1"
F_2= "la personne choisie préfère le quotidien q_2"
Ainsi, dans chaque système, les événements forment une partition de Ω. En effet,
— les couples d'événements sont mutuellement exclusifs, i.e. $E_1 \cap E_2 = \emptyset$,
— l'union des couples forment l'ensemble Ω, i.e. $E_1 \cup E_2 = \Omega$.

Exercice 12.5. Tirage de jetons
Dans une boite, il y a quatre jetons numérotés de 1 à 4. On tire simultanément au hasard deux jetons. On note l'évènement A= {les deux jetons sont pairs}.

1. Donner tous les tirages possibles.
2. Quels sont les tirages constituant les événements suivants :
   (a) A
   (b) $\overline{A}$
   (c) $A \cup \overline{A}$
   (d) $A \cap \overline{A}$
3. On considère l'événement C= {la somme des chiffres notés sur les deux jetons est pair}. Quels sont les tirages constituant les événements suivants :
   (a) C
   (b) $\overline{C}$
   (c) $A \cup C$
   (d) $A \cap C$
   (e) $A \cup \overline{C}$
   (f) $A \cap \overline{C}$

Solution

1. $A = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$.
2. (a) $A = \{(2,4)\}$.
   (b) $\overline{A} = \{(1,2),(1,3),(1,4),(2,3),(3,4)\}$.
   (c) $A \cup \overline{A} = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$.
   (d) $A \cap \overline{A} = \emptyset$.
3. (a) $C = \{(1,3),(2,4)\}$
   (b) $\overline{C} = \{(1,2),(1,4),(2,3),(3,4)\}$
   (c) $A \cup C = \{(1,3),(2,4)\}$
   (d) $A \cap C = \{(2,4)\}$
   (e) $A \cup \overline{C} = \{(1,2),(1,4),(2,3),(2,4),(3,4)\}$
   (f) $A \cap \overline{C} = \emptyset$.

Exercice 12.6. Cartes et évènements
Nous avons un jeu de cartes de 52 cartes (c'est-à-dire les cartes du 2 jusqu'à l'as). On tire deux cartes et on considère les événements suivants :
A = { les 2 cartes tirées sont des 6 }
B = { les 2 cartes tirées sont noires }
C = { les cartes tirées sont des figures (valet, dame, roi) }

1. Décrire en mots ce que vaut :
   (a) $\overline{B}$,
   (b) $A \cup B$,
   (c) $(A \cup B) \cap C$,
   (d) $(A \cap B) \cup C$.
2. Écrire à l'aide de A, B et C les événements suivants :
   (a) E = {au plus une des 2 cartes est noire}
   (b) F = {les cartes tirées sont des figures et ne sont pas toutes noires}

Solution

1. (a) Au moins 1 des cartes tirées n'est pas noire.
   (b) Les 2 cartes tirées sont noires, ou les 2 cartes tirées sont des 6.
   (c) Les 2 cartes tirées sont des figures noires, car $A \cap C = \emptyset$.
   (d) Les 2 cartes tirées sont des 6 et sont noires ($A \cap B$), ou les cartes tirées sont des figures.
2. (a) $E = \overline{B}$.
   (b) $F = C \cap \overline{B}$.

Exercice 12.7. Séquence d'enfants
On vous demande de répondre à la question suivante de façon intuitive (pas besoin de calcul).
Avec votre femme/mari vous envisagez d'avoir sept enfants. Si la probabilité d'avoir une fille est la même que celle d'avoir un garçon, lequel des événements suivants vous semble le plus probable.

1. Avoir : un garçon puis une fille puis une fille puis un garçon puis une fille puis une fille puis un garçon.
2. Avoir : un garçon puis un garçon puis un garçon puis un garçon puis un garçon puis un garçon puis un garçon.

Solution
Soient les événements F = "avoir une fille" et G = "avoir un garçon" :

$$\Pr(F) = 0.5 \text{ et } \Pr(G) = 0.5$$

Etant donné que les événements sont indépendants, pour le point (1) on obtiendra :

$$G \cap F \cap F \cap G \cap F \cap F \cap G.$$

Donc,

$$\Pr(1) = \Pr(G) \times \Pr(F) \times \Pr(F) \times \Pr(G) \times \Pr(F) \times \Pr(F) \times \Pr(G) = 0.5^7 \approx 0.007825.$$

Pour le point (2) on aura par contre :

$$G \cap G \cap G \cap G \cap G \cap G \cap G.$$

Donc,

$$\Pr(2) = \Pr(G) \times \Pr(G) \times \Pr(G) \times \Pr(G) \times \Pr(G) \times \Pr(G) \times \Pr(G) = 0.5^7 \approx 0.007825.$$

La probabilité de l'événement 1 est exactement la même que celle de l'événement 2. Cependant une étude a révélé que la plupart des gens pensent que l'événement 2 est plus improbable. Ceci est dû à une erreur de raisonnement. Nous avons tendance à assimiler l'événement 1 à l'ensemble des événements "typiques" (des mélanges de filles et de garçons qui ne sautent pas aux yeux). Au contraire, nous assimilons l'événement 2 à l'ensemble des résultats "atypiques" (comme avoir que des filles ou avoir que des garçons).
Le groupe des événements "typiques" est bien plus nombreux que le groupe des événements "atypiques". Il est donc correct de dire qu'il est plus probable d'avoir un résultat "typique" plutôt qu'un résultat "atypique".
Cependant telle n'est pas la question. On ne nous demande pas de comparer la probabilité d'avoir un événement typique et celle d'avoir un événement atypique mais de comparer la probabilité de deux événements très spécifiques appartenant à chaque groupe.

Exercice 12.8. Activités après le bachelor
Une étude sur l'activité exercée pendant l'année qui suit l'obtention d'un bachelor donne les résultats suivants :
— 50% des diplômés suivent des cours de Master
— 22% des diplômés ont une activité professionnelle
— 10% des diplômés sont en Master et ont une activité professionnelle
— le reste des diplômés ont une autre activité
On prend un individu au hasard, on aimerait connaître :

1. La probabilité que cette personne soit en Master
2. La probabilité que cette personne soit en Master ou ait une activité professionnelle
3. La probabilité que cette personne ait une activité professionnelle, mais ne soit pas en Master
4. La probabilité que cette personne ait une activité professionnelle sachant qu'elle est en Master

Solution
Soient les événements :

$$M \ : \text{“La personne est en Master”}$$

$$P \ : \text{“La personne exerce une activité professionnelle”}$$

On a :
La probabilité d'être en Master : $\Pr(M) = 0.50$
La probabilité d'être en Master et d'exercer une activité professionelle : $\Pr(M \cap P) = 0.10$
Et la probabilité d'exercer une activité professionelle : $\Pr(P) = 0.25$

1. $\Pr(M) = 0.50$
2. $\Pr(M \cup P) = \Pr(M) + \Pr(P) - \Pr(M \cap P) = 0.50 + 0.25 - 0.10 = 0.65$
3. $\Pr(P \cap \bar{M}) = \Pr(P) - \Pr(M \cap P) = 0.25 - 0.10 = 0.15$
4. $\Pr(P|M) = \dfrac{\Pr(P \cap M)}{\Pr(M)} = \dfrac{0.10}{0.50} = 0.20$

Exercice 12.9. Boîtes de chocolats
Il y a 2 boîtes de chocolats : la première contient 3 chocolats au lait, 2 chocolats blancs et un chocolat noir. La deuxième contient 2 chocolats au lait, 1 chocolat blanc et 4 chocolats noirs. On tire au hasard 1 chocolat dans la première boîte, si le chocolat est noir, on retire dans la même sans le remettre dans la boîte, sinon on tire dans la deuxième boîte.

1. Quelle est la probabilité d'obtenir 2 chocolats noirs ?
2. Calculer la probabilité d'obtenir 1 chocalat noir et 1 blanc.
3. Calculer la probabilité d'obtenir 2 chocolats au lait.

Solution

L_1	le 1er chocolat est au lait $\Pr(L_1) = 3/6$
B_1	le 1er chocolat est blanc $\Pr(B_1) = 2/6$
N_1	le 1er chocolat est noir $\Pr(N_1) = 1/6$
L_2	le 2ème chocolat est au lait
B_2	le 2ème chocolat est blanc
N_2	le 2ème chocolat est noir
L	les 2 chocolats tirés sont au lait
B	les 2 chocolats tirés sont blanc
N	les 2 chocolats tirés sont noir

1. Probabilité d'obtenir 2 chocolats noirs :

$$\Pr(N) = \Pr(N_1 \cap N_2) = \Pr(N_1) \times \Pr(N_2|N_1) = 1/6 \times 0/5 = 0.$$

2. Calculer la probabilité d'obtenir 1 chocolat noir et 1 blanc.

$$\begin{aligned}\Pr(N_1 \cap B_2) + \Pr(N_2 \cap B_1) &= \Pr(N_1) \times \Pr(B_2|N_1) + \Pr(B_1) \times \Pr(N_2|B_1) \\ &= 1/6 \times 2/5 + 2/6 \times 4/7 = 2/30 + 8/42 = 0.257.\end{aligned}$$

3. Calculer la probabilité d'obtenir 2 chocolats au lait.

$$\Pr(L) = \Pr(L_1 \cap L_2) = \Pr(L_1) \times \Pr(L_2|L_1) = 3/6 \times 2/7 = 6/42.$$

Exercice 12.10. Jeu avec des cartes
On vous propose un jeu de hasard : vous tirez une carte dans un jeu de 36 cartes. Si vous tirez un coeur vous gagnez 10.-, si vous tirez un trèfle vous gagnez 5.- sinon vous ne gagnez rien. Le prix pour participer est de 2.-. Combien pouvez-vous espérer gagner ?

Solution

Distribution des probabilités

	coeur	trèfle	autres
x	$10-2$	$5-2$	-2
$\mathrm{Pr}_X(x)$	$9/36$	$9/36$	$18/36$

$$\mathrm{E}(X) = \sum_x x\mathrm{Pr}_X(x) = 9/36 \times 8 + 9/36 \times 3 + 18/36 \times (-2) = 1.75.$$

Cela signifie qu'en moyenne, je gagne 1.75.-

Exercice 12.11. Boules de Noël

1. Une boite contient 40 boules de Noël distinctes, dont 20 sont rouges, 10 sont dorées et 10 sont bleues. Huit personnes tirent tour à tour 5 boules au hasard.
   (a) Combien de possibilté de tirage différent a chaque joueur, si on considère que l'ordre est important ?
   (b) Combien de possibilité de tirage y a-t-il indépendemment de l'ordre ?
   (c) Quel est la probabilité de tirer 5 boules rouges ?
   (d) Quel est la probabilité d'avoir 4 boules bleues si la 1ère boule tirée est bleue ?
2. On vous propose un jeu de hasard : Vous tirez une carte dans un jeu de 36 cartes. Si vous tirez un coeur vous gagnez 10.-, si vous tirez un trèfle vous gagnez 5.- sinon vous ne gagnez rien. Le prix pour participer est de 2.-, êtes vous partant ?
3. Le Tableau 12.1 donne la distribution de probabilité de X.

Table 12.1 – Distribution de probabilité de X

x	3	4	5	6
$\Pr(X = x)$	0.1	0.4	0.3	0.1

   (a) Donner les valeurs de la fonction de répartition
   (b) Calculer l'espérance de X.
   (c) Calculer la variance de X.

Solution

1. (a) Arrangement sans répétition :

$$A_n^k = A_{40}^5 = \frac{40!}{(40-5)!} = 78960960.$$

   (b) Combinaison :

$$\binom{n}{k} = \binom{40}{5} = \frac{40!}{5!(40-5)!} = 658008.$$

   (c) Nombre de couples de 5 boules qui sont toutes rouges :

$$\binom{n}{k} = \binom{20}{5} = \frac{20!}{5!(20-5)!} = 15504,$$

$$\Pr(5 \text{ boules rouges}) = \frac{\binom{20}{5}}{\binom{40}{5}} = \frac{15504}{658008} = 0.0235.$$

   (d) Quel est la probabilité d'avoir 4 boules bleues si la première est bleue ?

$$\binom{n}{k} = \binom{9}{5} = \frac{9!}{5!(9-5)!} = 126,$$

$$\binom{n}{k} = \binom{39}{5} = \frac{39!}{5!(39-5)!} = 575757,$$

$$\Pr(4 \text{ boules bleues} \mid 1\text{ère est bleue}) = \frac{\binom{9}{5}}{\binom{39}{5}} = \frac{126}{575757} = 0.0002.$$

2. Distribution de probabilité :

	coeur	trèfle	autres
x	10-2	5-2	-2
$\Pr(x)$	9/36	9/36	18/36

$$\mathrm{E}(X) = 9/36 \times 8 + 9/36 \times 3 + 18/36 \times (-2) = 1.75.$$

Cela signifie qu'en moyenne je gagne 1.75.-

3. (a) Valeurs de la fonction de répartition :

x	3	4	5	6
$\Pr(X = x)$	0.1	0.5	0.8	1

(b) L'espérance de X :

$$\mathrm{E}(X) = 3 \times 0.1 + 4 \times 0.4 + 5 \times 0.3 + 6 \times 0.1 = 4.$$

(c) La variance de X :

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = (3^2 \times 0.1 + 4^2 \times 0.4 + 5^2 \times 0.3 + 6^2 \times 0.1) - 4^2 = 18.4 - 16 = 2.4.$$

Exercice 12.12. Lancer de dés
On lance un dé deux fois. Soient les événements :

$$A = \text{Obtenir un 1 au premier lancer,}$$
$$B = \text{Obtenir un 2 au deuxième lancer.}$$

On a que $\Pr(A \cap B) = \dfrac{1}{36}$.

1. Donner la probabilité des événements A et B.
2. Donner la probabilité de l'union des événements A et B.
3. Est-ce que les événements A et B sont indépendants ? Justifier.
4. Est-ce que les événements A et B sont mutuellement exclusifs ? Justifier.

Solution

1. Donner la probabilité des événements A et B.

$$\Pr(A) = \frac{1}{6}$$
$$\Pr(B) = \frac{1}{6}$$

2. Probabilité de l'union des événements A et B :

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}.$$

3. Est-ce que les événements A et B sont indépendants ?

$$\Pr(A)\Pr(B) = \frac{1}{36},$$
$$\Pr(A \cap B) = \frac{1}{36},$$
$$\Pr(A)\Pr(B) = \Pr(A \cap B).$$

Oui, les événements sont indépendants.

4. Est-ce que les événements A et B sont mutuellement exclusifs ?

$$\Pr(A \cap B) = \frac{1}{36} \neq \emptyset.$$

Non, les événements ne sont pas mutuellement exclusifs.

Exercice 12.13. Chaussures et probabilités
Un couple entre dans une boutique de chaussures. Soient les événements :
— A = "l'homme achète une paire de chaussures",
— B = "la femme achète une paire de chaussures".
Supposons que $\Pr(A) = 0.70$, $\Pr(B) = 0.40$ et $\Pr(A \cap B) = 0.30$. Calculer les probabilités suivantes :

1. l'homme ou la femme achète(nt) une paire de chaussures,
2. l'homme achète une paire de chaussures, mais pas la femme,
3. l'homme n'achète pas de chaussures, mais la femme achète une paire,
4. l'homme achète une paire de chaussures étant donné que la femme en achète une,
5. la femme n'achète pas de chaussure sachant que l'homme en achète.

Solution

1. l'homme ou la femme achète une paire de chaussures,
$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.70 + 0.40 - 0.30 = 0.80.$$
2. l'homme achète une paire de chaussures, mais pas la femme,
$$\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B) = 0.70 - 0.30 = 0.40.$$
3. l'homme n'achète pas de chaussures, mais la femme achète une paire,
$$\Pr(\overline{A} \cap B) = \Pr(B) - \Pr(A \cap B) = 0.40 - 0.30 = 0.10.$$
4. l'homme achète une paire de chaussures étant donné que la femme en achète une,
$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{0.30}{0.40} = 0.75.$$
5. la femme n'achète pas de chaussure sachant que l'homme en achète
$$\Pr(\overline{B}|A) = \frac{\Pr(\overline{B} \cap A)}{\Pr(A)} = \frac{0.40}{0.70} = 0.57.$$

Exercice 12.14. Pratiques culturelles
Les autorités d'une ville ont recensé les pratiques culturelles de ses citoyens pendant un mois. Il ressort notamment les résultats suivants :

— 42% des citoyens sont allés (au moins une fois) au cinéma.
— 15% ont fréquenté un musée.
— 8% ont fréquenté musées et cinémas.

On prend un individu au hasard dans cette ville.

1. Calculer la probabilité que cet individu soit allé au cinéma ou au musée.
2. Calculer la probabilité qu'il ne soit pas allé au cinéma.
3. Calculer la probabilité qu'il ne soit allé ni au musée, ni au cinéma.
4. Calculer la probabilité qu'il soit allé au cinéma, mais pas au musée.
5. Calculer la probabilité qu'il soit allé au cinéma, sachant qu'il est allé au musée.
6. Est-ce que les événements "aller au cinéma" et "aller au musée" sont indépendants ? Justifier.

Solution
Evenements : A = "aller au cinéma",
B = "aller au musée".
On a : $\Pr(A) = 0.42$, $\Pr(B) = 0.15$ et $\Pr(A \cap B) = 0.08$.

1. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.42 + 0.15 - 0.08 = 0.49$.
2. $\Pr(\overline{A}) = 1 - \Pr(A) = 1 - 0.42 = 0.58$.
3. $\Pr(\overline{A} \cap \overline{B}) = \Pr(\overline{A \cup B}) = 1 - \Pr(A \cup B) = 1 - 0.49 = 0.51$.
4. $\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B) = 0.42 - 0.08 = 0.34$..
5. $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{0.08}{0.15} = 0.53$.
6. Deux événements sont indépendants si $\Pr(A|B) = \Pr(A)$.
   On a $\Pr(A|B) = 0.53 \neq 0.42 = \Pr(A)$. Donc, les événements ne sont pas indépendants.

Exercice 12.15. Qualité d'ampoules
Une entreprise fabrique des ampoules. La probabilité qu'une ampoule fabriquée soit défectueuse est de 1/50. Un inspecteur tire un échantillon de trois ampoules dans un lot d'ampoules. Calculer les probabilités suivantes :

1. les trois ampoules de l'échantillon sont défectueuses,
2. aucune ampoule de l'échantillon n'est défectueuse,
3. une seule ampoule est défectueuse dans l'échantillon,
4. au moins une ampoule dans l'échantillon est défectueuse,
5. les deux premières ampoules de l'échantillon sont défectueuses.

Solution

1. Les trois ampoules de l'échantillon sont défectueuses. Soient les événements D_i= "l'ampoule i est défectueuse" pour $i = 1, 2, 3$ et D= "une ampoule est défectueuse".
   On a que $\Pr(D_1) = \Pr(D_2) = \Pr(D_3) = \Pr(D)$.
$$\Pr(D_1 \cap D_2 \cap D_3) = \Pr(D_1)\Pr(D_2)\Pr(D_3) = \Pr(D)^3 = (1/50)^3 \simeq 0.00.$$
2. Aucune ampoule de l'échantillon n'est défectueuse.
$$\begin{aligned}\Pr(\overline{D}_1 \cap \overline{D}_2 \cap \overline{D}_3) &= \Pr(\overline{D}_1)\Pr(\overline{D}_2)\Pr(\overline{D}_3) = \Pr(\overline{D})^3 \\ &= \left[1 - \Pr(\overline{D})\right]^3 = [1 - 1/50]^3 = 0.94.\end{aligned}$$
3. Une seule ampoule est défectueuse dans l'échantillon.
$$\begin{aligned}\Pr(D_1 \cap \overline{D}_2 \cap \overline{D}_3) + \Pr(\overline{D}_1 \cap D_2 \cap \overline{D}_3) + \Pr(\overline{D}_1 \cap \overline{D}_2 \cap D_3) &= 3\Pr(D)\Pr(\overline{D})\Pr(\overline{D}) \\ = 3\Pr(D)\left[1 - \Pr(\overline{D})\right]^2 &= 3 \times 1/50 \times (49/50)^2 = 0.06.\end{aligned}$$
4. Au moins une ampoule dans l'échantillon est défectueuse.
$$1 - \Pr(\overline{D}_1 \cap \overline{D}_2 \cap \overline{D}_3) = 1 - 0.94 = 0.06.$$
5. les deux premières ampoules de l'échantillon sont défectueuses.
$$\Pr(D_1 \cap D_2 \cap \overline{D}_3) = \Pr(D)^2\Pr(\overline{D}) = \Pr(D)^2\,[1 - \Pr(D)] = (1/50)^2\,(1 - 1/50) \simeq 0.00.$$

Exercice 12.16. Urne et boules
Deux urnes contiennent respectivement 4 boules rouges et 3 boules vertes, 5 boules rouges et 3 boules vertes. On tire au hasard une boule dans la première (sans remise), puis on procède au tirage d'une deuxième boule, dans la même urne si la première boule tirée est rouge, dans l'autre urne si la première boule tirée est verte.

1. Quelle est la probabilité d'obtenir deux boules vertes ? deux boules rouges ?
2. Si les deux boules tirées sont de même couleur, quelle est la probabilité qu'elles soient rouges ?
3. Calculer la probabilité d'obtenir une boule verte et une boule rouge.

Solution
On note les deux urnes U_A et U_B. De plus, on note :
— V_1 : la première boule tirée est verte,
— R_1 : la première boule tirée est rouge,
— V_2 : la deuxième boule tirée est verte,
— R_2 : la deuxième boule tirée est rouge,
— R : les deux boules sont rouges,
— V : les deux boules sont vertes,
— $\Pr(V_1) = 3/7$ et $\Pr(R_1) = 4/7$.

1. Probabilité d'obtenir deux boules vertes :

$$\begin{aligned}\Pr(V) &= \Pr(V_2 \cap V_1) = \Pr(V_1) \times \Pr(V_2|V_1)\\ &= \frac{3}{7} \times \Pr(\text{tirer une boule verte dans } U_B)\\ &= \frac{3}{7} \times \frac{3}{8} = \frac{9}{56} \approx 0.161.\end{aligned}$$

Probabilité d'obtenir deux boules rouges :

$$\begin{aligned}\Pr(R) &= \Pr(R_2 \cap R_1) = \Pr(R_1) \times \Pr(R_2|R_1)\\ &= \frac{4}{7} \times \Pr(\text{tirer encore une boule rouge dans } U_A)\\ &= \frac{4}{7} \times \frac{3}{6} = \frac{4}{14} \approx 0.286.\end{aligned}$$

2. Si les deux boules tirées sont de même couleur, la probabilité qu'elles soient rouges est :

$$\Pr[R|(R \cup V)] = \frac{\Pr[R \cap (R \cup V)]}{\Pr(R \cup V)} = \frac{\Pr(R)}{\Pr(R \cup V)}$$

$\Pr(R) = 4/14$ (cf. question précédente)
$\Pr(R \cup V) = \Pr(R) + \Pr(V) - \Pr(R \cap V) = (4/14) + \frac{3}{7} \times (3/8) - 0 = 25/56$.
Alors,

$$\Pr[R|(R \cup V)] = \frac{\Pr(R)}{\Pr(R \cup V)} = \frac{4/14}{25/56} = \frac{16}{25} = 0.64.$$

$$\frac{\Pr(R_2 \cap R_1)}{\Pr[(R_2 \cap R_1) \cup (V_2 \cap V_1)]} = \frac{4/14}{25/56} = 16/25 = 0.64.$$

3. Probabilité d'obtenir une boule verte et une boule rouge :

$$\begin{aligned}&\Pr(V_1 \cap R_2) + \Pr(V_2 \cap R_1) = \Pr(V_1) \times \Pr(R_2|V_1) + \Pr(R_1) \times \Pr(V_2|R_1)\\ &= \frac{3}{7} \times \Pr(\text{rouge dans } U_2) + \frac{4}{7} \times \Pr(\text{verte dans } U_1 \text{modifiée})\\ &= \frac{3}{7} \times \frac{5}{8} + \frac{4}{7} \times \frac{3}{6} = \frac{31}{56} \approx 0.554\end{aligned}$$

Autre version (plus rapide) :

$$\Pr(\text{couleurs différentes}) = 1 - \Pr(\text{mêmes couleurs}) = 1 - \Pr(R \cup V) = 1 - \frac{25}{56} = \frac{31}{56} \approx 0.554$$

Exercice 12.17. Probabiliés d'accident
Une compagnie d'assurance automobile cherche à estimer la probabilité qu'un assuré ait un accident durant l'année. Pour ce faire, elle répartit les assurés en trois groupes : individus très enclins à un accident (groupe à haut risque), individus moyennement enclins à un accident (groupe à risque intermédiaire) et individus faiblement exposés à un accident (groupe à faible risque).
Les probabilités qu'un individu appartenant à l'un de ces trois groupes ait un accident dans l'année sont respectivement de 0.1 (groupe à haut risque), 0.01 (groupe à risque intermédiaire) et 0.005 (groupe à faible risque). On sait en outre que 20% des individus font partie du groupe à haut risque, que 10% font partie du groupe à faible risque et que tous les autres individus font partie du groupe à risque intermédiaire.

1. Calculer, dans ce cas de figure, la probabilité qu'un nouvel assuré ait un accident dans l'année.
2. Un individu assuré auprès de cette compagnie d'assurance a un accident durant l'année.
   (a) Quelle est la probabilité que cet individu fasse partie du groupe à risque intermédiaire ?
   (b) Quelle est la probabilité que cet individu fasse partie du groupe à faible risque ?

Solution

1. On choisit un individu au hasard. Posons :
   — G_1= l'individu appartient au groupe à haut risque,
   — G_2= l'individu appartient au groupe à risque intermédiaire,
   — G_3= l'individu appartient au groupe à faible risque et
   — A= l'individu a un accident durant l'année.
   On a alors
$$\begin{array}{ll} \Pr(A|G_1) = 0.1 & \Pr(G_1) = 0.2 \\ \Pr(A|G_2) = 0.01 & \Pr(G_2) = 0.7 \\ \Pr(A|G_3) = 0.005 & \Pr(G_3) = 0.1. \end{array}$$
   On a que G_1, G_2 et G_3 forment un système complet d'événements. Par le théorème des probabilités totales, on a alors :
$$\begin{aligned} \Pr(A) &= \sum_{i=1}^{3} \Pr(A|G_i)\Pr(G_i) \\ &= \Pr(A|G_1)\Pr(G_1) + \Pr(A|G_2)\Pr(G_2) + \Pr(A|G_3)\Pr(G_3) \\ &= 0.1 \times 0.2 + 0.01 \times 0.7 + 0.005 \times 0.1 = 0.0275. \end{aligned}$$
   Ainsi, la probabilité qu'un nouvel assuré ait un accident dans l'année est de 0.0275, c'est-à-dire non loin de 3%.
2. (a) On cherche $\Pr(G_2|A)$. Par le théorème de Bayes, on a (voir question 1.) :
$$\Pr(G_2|A) = \frac{\Pr(G_2)\Pr(A|G_2)}{\sum_{i=1}^{3} \Pr(A|G_i)\Pr(G_i)} = \frac{0.7 \times 0.01}{0.0275} = \frac{0.007}{0.0275} = 0.2545.$$
   (b) On cherche $\Pr(G_3|A)$. Par le théorème de Bayes, on a (voir question 1.) :
$$\Pr(G_3|A) = \frac{\Pr(G_3)\Pr(A|G_3)}{\sum_{i=1}^{3} \Pr(A|G_i)\Pr(G_i)} = \frac{0.1 \times 0.005}{0.0275} = \frac{0.0005}{0.0275} = 0.0182.$$

Exercice 12.18. Théorème des probabilités totales
On tire 2 cartes dans un jeu de 52 (qui comporte donc toutes les cartes du deux à l'as). Quelle est la probabilité pour que la deuxième carte soit un as ?

Solution
On peut décomposer cet événement selon la nature de la première carte. Notons A_1 l'événement "la première carte du paquet est un as" et A_2 l'événement "la deuxième carte du paquet est un as". Par application du théorème des probabilités totales, on a

$$\Pr(A_2) = \Pr(A_1)\Pr(A_2|A_1) + \Pr(\overline{A_1})\Pr(A_2|\overline{A_1}) = \frac{4}{52}\frac{3}{51} + \frac{48}{52}\frac{4}{51} = \frac{4}{52} = \frac{1}{13}.$$

Exercice 12.19. Monstres
Jacques joue a un jeu éléctronique fort instructif l'amenant à combattre un monstre choisi aléatoirement parmi les deux monstres féroces

— *Saturnin le canard* choisi avec une probabilité de $\frac{3}{4}$ et
— *Dora l'exploratrice* choisie avec une probabilité de $\frac{1}{4}$.

Lorsqu'il combat *Saturnin le canard*, Jacques gagne avec une probabilité de $\frac{1}{3}$ et lorsqu'il combat *Dora l'exploratrice*, Jacques gagne avec une probabilité de $\frac{1}{2}$. Soit alors

$$X = \begin{cases} 1, & \text{si Jacques gagne la partie et} \\ 0, & \text{sinon.} \end{cases}$$

Il s'agit d'une variable aléatoire *indicatrice* ou *bernoullienne*.
Donner son paramètre, son espérance et sa variance.

Solution

$$\begin{aligned} \Pr(X = 1|\text{Jacques combat } \textit{Saturnin}) &= \frac{1}{3} \\ \Pr(X = 1|\text{Jacques combat } \textit{Dora}) &= \frac{1}{2}. \end{aligned}$$

Par le théorème des probabilités totales (ou en faisant un arbre), on obtient

$$\Pr(X = 1) = \frac{3}{4} \times \frac{1}{3} + \frac{1}{4} \times \frac{1}{2} = \frac{3}{8} = 0.375.$$

Ainsi, le paramètre est $p = 0.375$. Comme X est une variable indicatrice ou bernoullienne, on en déduit immédiatement que

$$\mathrm{E}(X) = p = 0.375 \text{ et } \sigma^2 = \mathrm{var}(X) = p(1-p) = 0.375 \times 0.625 = 0.234.$$

Exercice 12.20. Temps de travail
Une petite entreprise de maçonnerie est formée de trois maçons. Chacun travaille un maximum de 40 heures hebdomadaires. Qu'ils aient du travail ou non, ils touchent un salaire de 50.- de l'heure. Avec l'expérience, la patron a pu constater que la demande de travail (en heures hebdomadaires) à laquelle est soumise son entreprise (notée Q) suit une loi uniforme sur l'intervalle $[100, 120]$. En outre, chaque heure travaillée lui rapporte 70.-.

1. Est-ce que le profit P (recettes totales - coût total) est une variable aléatoire ?
2. Donner ce profit P.
3. Quel serait le profit une semaine où la demande de travail serait de 120 heures ($Q = 120$) ?
4. Quel serait le profit une semaine où la demande de travail serait de 100 heures ($Q = 100$) ?
5. Quel est le profit espéré ?

Solution

1. Oui, car les recettes dépendent de la demande de travail.
2. $P = 70 \times Q - 3 \times 40 \times 50 = 70 \times Q - 6000$.
3. $P = 70 \times 120 - 3 \times 40 \times 50 = 8400 - 6000 = 2400$.
4. $P = 70 \times 100 - 3 \times 40 \times 50 = 7000 - 6000 = 1000$.
5. Profit espéré :

$$\begin{aligned} \mathrm{E}(P) &= \mathrm{E}(70 \times Q - 6000) = 70\mathrm{E}(Q) - 6000 \\ &= 70 \times \frac{100 + 120}{2} - 6000 = 70 \times 110 - 6000 = 1700. \end{aligned}$$

Exercice 12.21. Tabagisme et cancer
Supposons qu'une population d'adultes soit composée de 95 pourcent de non-fumeurs et de 5 pourcent de fumeurs. Nous savons aussi que la probabilité de mourir d'un cancer est 10 fois plus élevée chez les fumeurs que chez les non-fumeurs.

1. Est-ce que nous disposons de suffisamment d'information pour comparer la probabilité d'être fumeur sachant qu'on est mort d'un cancer et la probabilité d'être non-fumeur sachant qu'on est mort d'un cancer ?
2. Supposons maintenant que la probabilité de mourir d'un cancer est de 0.02 pour les non-fumeurs et de 0.2 pour les fumeurs. Calculez les probabilités qu'on doit comparer dans la question précédente.

Solution

1. Dans cet exercice, on suppose que les fumeurs ont 10 fois plus de chances de mourir d'un cancer, mais ils sont aussi 19 fois moins nombreux que les non-fumeurs.
2. On note C l'événement "Cancer", F l'événement "être un fumeur" et N l'événement "être un non-fumeur". Selon le théorème des probabilités totales nous avons :

$$\Pr(C) = \Pr(C|N)\Pr(N) + \Pr(C|F)\Pr(F) = 0.02 \times 0.95 + 0.2 \times 0.05 = 0.029.$$

On cherche $\Pr(N|C)$. Par le théorème de Bayes, on a :

$$\Pr(N|C) = \frac{\Pr(N)\Pr(C|N)}{\Pr(C)} = \frac{0.019}{0.029} = 0.6552,$$

et donc

$$\Pr(F|C) = 1 - \Pr(N|C) \quad = \quad 0.3448.$$

La probabilité d'être non-fumeur sachant qu'on est mort d'un cancer 1.9 fois supérieure à la probabilité d'être fumeur sachant qu'on est mort d'un cancer.

Exercice 12.22. Théorème des probabilités totales et de Bayes
La proportion réelle des électeurs votant pour le candidat X est p. Au cours d'un sondage, un électeur qui va réellement voter pour le candidat X répond honnêtement avec une probabilité de 90%. Ceux qui ne voteront pas pour X répondent honnêtement avec une probabilité de 95%.

1. Calculer, en fonction de p, la probabilité q pour qu'un électeur, pris au hasard, réponde qu'il va voter pour X.
2. En déduire, en fonction de p, la probabilité r pour qu'un électeur, pris au hasard, vote réellement pour X sachant qu'il a répondu qu'il vote pour X.
3. Application numérique. Calculer q et r lorsque $p = 5\%$ et $p = 45\%$.

Solution
Notons H l'événement "l'électeur interrogé répond honnêtement" et A l'événement "l'électeur vote pour le candidat X". Les données de l'énoncé sont donc :

$$\Pr(H|A) = 0.9 \;\; \Pr(H|\overline{A}) = 0.95 \;\; \Pr(A) = p.$$

1. L'électeur répond qu'il va voter pour X s'il va vraiment voter pour X et répond honnêtement, ou s'il ne vas voter pour X et répond malhonnêtement. Notons Q l'événement "l'électeur répond qu'il vote pour X". Donc, $\Pr(Q|A) = \Pr(H|A)$ et $\Pr(Q|\overline{A}) = \Pr(\overline{H}|\overline{A})$.
   Par le théorème des probabilités totales, on a
$$\begin{aligned} q &= \Pr(A)\Pr(Q|A) + \Pr(\overline{A})\Pr(Q|\overline{A}) \\ &= \Pr(A)\Pr(H|A) + \Pr(\overline{A})\Pr(\overline{H}|\overline{A}) \\ &= 0.9p + 0.05(1-p) \\ &= 0.05 + 0.85p. \end{aligned}$$
2. Par le théorème de Bayes, on a
$$r = \Pr(A|Q) = \frac{\Pr(A)\Pr(Q|A)}{\Pr(Q)} = \frac{\Pr(A)\Pr(H|A)}{\Pr(Q)} = \frac{90p}{5+85p}.$$
3. Si $p = 45\%$, $q = 43.25\%$ et $r = 93.64\%$,
   si $p = 5\%$, $q = 9.25\%$ et $r = 48.65\%$.

Exercice 12.23. Test de grossesse
Sur la notice du test de grossesse urinaire de la firme XYBB on peut lire *"...ce test détecte avec une fiabilité de 99% une grossesse chez une femme effectivement enceinte. Cependant, ce test produit 2% de faux positifs (résultat positif au test alors que le sujet n'est pas enceinte) pouvant être dus à la prise de certains médicaments contenant des hormones détectées par le test..."*.
On estime que 4% des femmes suisses âgées de 20 à 40 ans sont enceintes.
Madame X, qui est âgée de 32 ans et habite en Suisse, effectue un test de grossesse urinaire de la firme XYBB.

1. Le résultat du test est positif.
   Quelle est la probabilité que madame X soit effectivement enceinte ?
2. Le résultat du test est positif.
   Quelle est la probabilité que madame X ne soit pas enceinte ?
3. Le résultat est négatif.
   Quelle est la probabilité que madame X soit tout de même enceinte (on parle alors de *faux négatif)* ?
4. Quelle est le probabilité que le résultat du test reflète la réalité (c'est-à-dire que le résultat d'une femme effectivement enceinte soit positif et que celui d'une femme qui n'est pas enceinte soit négatif) ?

Solution
On note
— A_1 = Madame X est enceinte,
— A_2 = Madame X n'est pas enceinte et
— B = le résultat du test est positif.
On sait que

$$\Pr(A_1) = 0.04, \Pr(A_2) = 1 - \Pr(A_1) = 0.96, \Pr(B|A_1) = 0.99, \Pr(B|A_2) = 0.02.$$

1. Par le théorème de Bayes :

$$\Pr(A_1|B) = \frac{\Pr(A_1)\Pr(B|A_1)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.04 \times 0.99}{0.04 \times 0.99 + 0.96 \times 0.02} = 0.67.$$

2. Il y a deux méthodes applicables.
   — Méthode 1 :
   $$\Pr(A_2|B) = 1 - \Pr(A_1|B) = 1 - 0.67 = 0.33.$$
   — Méthode 2 :
   $$\Pr(A_2|B) = \frac{\Pr(A_2)\Pr(B|A_2)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.96 \times 0.02}{0.04 \times 0.99 + 0.96 \times 0.02} = 0.33.$$

3. On sait que

$$\Pr(\overline{B}|A_1) = 1 - 0.99 = 0.01 \text{ et } \Pr(\overline{B}|A_2) = 1 - 0.02 = 0.98.$$

   Ainsi,

$$\Pr(A_1|\overline{B}) = \frac{\Pr(A_1)\Pr(\overline{B}|A_1)}{\Pr(A_1)\Pr(\overline{B}|A_1) + \Pr(A_2)\Pr(\overline{B}|A_2)} = \frac{0.04 \times 0.01}{0.04 \times 0.01 + 0.96 \times 0.98} = 0.00042.$$

4. On a

$$\begin{aligned}\Pr(B \cap A_1) + \Pr(\overline{B} \cap A_2) &= \Pr(B|A_1)\Pr(A_1) + \Pr(\overline{B}|A_2)\Pr(A_2) \\ &= 0.99 \times 0.04 + 0.98 \times 0.96 = 0.98.\end{aligned}$$

Exercice 12.24. Test pharmaceutique
Un laboratoire pharmaceutique vend un test avec la notice suivante : si vous êtes malade, alors le test est positif avec probabilité $\alpha = 98\%$ (α est la sensibilité du test), si vous êtes sain, alors le test est positif avec probabilité $\beta = 3\%$ ($1 - \beta$ est la spécificité du test). Sachant qu'en moyenne il y a un malade sur 1000 personnes :

1. Calculer la probabilité pour que vous soyez un sujet sain sachant que votre test est positif.
2. Calculer la probabilite d'être malade sachant que le test est négatif.

Solution
Notons

S :	je suis sain,
M :	je suis malade,
P :	mon test est positif,
N :	mon test est négatif.

On connaît les valeurs de probabilités suivantes :

$$\Pr(P|M) = \alpha = 0.98,$$
$$\Pr(P|S) = \beta = 0.03,$$
$$\Pr(M) = \frac{1}{1000} = 0.001.$$

On peut trouver que :

$$\Pr(S) = 1 - \Pr(M) = 0.999$$

On déduit du théorème de Bayes que :

1. $$\Pr(S|P) = \frac{\Pr(S)\Pr(P|S)}{\Pr(M)\Pr(P|M) + \Pr(S)\Pr(P|S)} = \frac{0.999 \times 0.03}{0.001 \times 0.98 + 0.999 \times 0.03} \simeq 0.968.$$

   Notons que
   $$\Pr(S|P) = \frac{\Pr(S \cap P)}{\Pr(P)} = \frac{\Pr(S)\Pr(P|S)}{\Pr(M)\Pr(P|M) + \Pr(S)\Pr(P|S)}.$$

2. On peut trouver que :
   $$\Pr(N|M) = 1 - \Pr(P|M) = 0.02 \text{ et } \Pr(N|S) = 1 - \Pr(P|S) = 0.97.$$
   $$\Pr(M|N) = \frac{\Pr(M)\Pr(N|M)}{\Pr(M)\Pr(N|M) + \Pr(S)\Pr(N|S)} = \frac{0.001 \times 0.02}{0.001 \times 0.02 + 0.999 \times 0.97} \simeq 0.00002.$$

Exercice 12.25. Anagrammes
Un anagramme est un mot composé de toutes les lettres d'un autre mot, mais dans un ordre différent. Par exemple, "conjoint" est un anagramme de "jonction". Dans cet exercice, toute suite de lettres est considérée comme un mot, même si celui-ci n'existe pas dans la langue française.

1. Combien y a-t-il d'anagrammes du mot RHONE ?
2. Parmi ceux-ci, combien commencent par la lettre H ?
3. Combien y a-t-il d'anagrammes du mot MISSISSIPPI ?
4. Parmi ces anagrammes, combien commencent et se terminent par la lettre S ?

Solution
Permutations avec répétition (objets non tous distincts) :

$$\overline{P}_n = \frac{n!}{n_1!n_2!\dots n_p!}$$

1. Anagrammes du mot RHONE : $P_n = n!$

$$P_5 = 5! = 120.$$

2. Anagrammes du mot RHONE commençant par H :

$$P_4 = 4! = 24.$$

3. Les anagrammes du mot MISSISSIPPI sont les permutations avec répétition des lettres MIIIISSSSPP. Ils sont donc au nombre de

$$\frac{11!}{1! \times 4! \times 4! \times 2!} = 34\,650.$$

4. Si on ne considère que les anagrammes commençant et se terminant par S, il s'agit toujours des permutations avec répétition mais des lettres MIIIISSPP car on a déjà placé deux S. Ces anagrammes sont ainsi au nombre de

$$\frac{9!}{1! \times 4! \times 2! \times 2!} = 3\,780.$$

Exercice 12.26. Course de chevaux
Lors d'une course de chevaux comptant 16 partants, combien de tiercés, de quartés et de quintés sont possibles, dans l'ordre et dans le désordre ?

Solution
Arrangements simples(sans répétitions) :

$$\mathrm{A}_n^k = \frac{n!}{(n-k)!}$$

Les tiercés, les quartés et les quintés, dans l'ordre, sont des arrangements, on tient compte de l'ordre d'arrivée. On obtient ainsi le nombre de :

— tiercés

$$A_{16}^3 = \frac{16!}{13!} = 3\,360$$

— quartés

$$A_{16}^4 = \frac{16!}{12!} = 43\,680$$

— quintés

$$A_{16}^5 = \frac{16!}{11!} = 524\,160$$

Pour obtenir le nombre dans le désordre, il faut diviser ces résultats par le nombre de permutations possibles de respectivement 3, 4 et 5 éléments. On obtient :

— tiercés

$$\binom{16}{3} = \frac{16!}{13!3!} = 560$$

— quartés

$$\binom{16}{4} = \frac{16!}{12!4!} = 1\,820$$

— quintés

$$\binom{16}{5} = \frac{16!}{11!5!} = 4\,368$$

Exercice 12.27. Avion
Nous avons un avion avec 788 places et 788 passagers.

1. De combien de façons différentes peut-on placer les passagers dans l'avion ?
2. Maintenant nous avons le même avion mais nous savons qu'uniquement 75 pourcent des billets ont été vendus. De combien de façon différentes peut-on placer les passagers dans l'avion ?

Solution

1. Nous pouvons raisonner de la façon suivante :
   Le premier passager peut s'asseoir à 788 places différentes, le second passager peut s'asseoir à 787 places différentes, le troisième peut s'asseoir à 786 places différentes puis successivement jusqu'au dernier passager.
   Réponse : $P_{788} = 788!$
2. Le raisonnement est le même sauf que nous n'allons pas jusqu'au dernier passager. On s'arrête au 591ième passager.
   Réponse : $A_{788}^{591} = \dfrac{788!}{197!}$

Exercice 12.28. Places à table
De combien de manières différentes pourriez-vous disposer vos six convives et vous-même autour d'une table ronde ?
Même question si on ne tient compte que de la position relative des sept personnes les unes par rapport aux autres ?
Et si il y a parmi les invités un couple d'amoureux désireux d'être assis l'un à côté de l'autre (en ne tenant compte que de la position relative des sept personnes les unes par rapport aux autres) ?

Solution
Nombre de manières de disposer les 6 convives et leur hôte

— $7! = 5040$.
— Si on ne tient compte que de la position relative des sept personnes les unes par rapport aux autres, on divise par 7 car on aura compté sept fois chacune des configurations. On trouve

$$\frac{7!}{7} = 6! = 720.$$

— Si, de plus, il y a un couple d'amoureux désireux d'être assis l'un à côté de l'autre. On peut procéder de la manière suivante : on place le couple, il y a deux choix différents possibles au sein du couple (individu 1 à droite de l'individu2 ou inversément) et il reste à placer les 5 autres personnes aux 5 places restantes (ce qui est une permutation sans répétition, d'où le 5 !). On trouve

$$2 \times 5! = 240.$$

Exercice 12.29. Notation binaire
Voici quelques exemples de notation binaire.

$$\begin{aligned} 100 &= 1\times 2^2 + 0\times 2^1 + 0\times 2^0 = 1\times 4 + 0\times 1 + 0\times 1 = 4, \\ 1011 &= 1\times 2^3 + 0\times 2^2 + 1\times 2^1 + 1\times 2^0 = 8 + 2 + 1 = 11. \end{aligned}$$

Combien de nombres peut-on ainsi obtenir avec 2 fois le “0” et 3 fois le “1” en plaçant un “1” en tête. Donner un nombre ainsi obtenu (en décimal).

Solution
Il s’agit d’une permutation avec répétition. Comme on doit placer un “1” en tête, il ne reste que 2 fois le “0” et 2 fois le “1”. On a donc

$$\frac{4!}{2!2!} = 6$$

nombres possibles, qui sont

$$\begin{aligned} 10011 &= 2^4 + 2^1 + 2^0 = 19 \\ 10101 &= 2^4 + 2^2 + 2^0 = 21 \\ 10110 &= 2^4 + 2^2 + 2^1 = 22 \\ 11001 &= 2^4 + 2^3 + 2^0 = 25 \\ 11010 &= 2^4 + 2^3 + 2^1 = 26 \\ 11100 &= 2^4 + 2^3 + 2^2 = 28 \end{aligned}$$

Exercice 12.30. Loterie
De combien de manières différentes pouvez-vous remplir votre bulletin de loterie sachant que vous devez cocher 6 numéros parmi les numéros de 1 à 49, 1 et 49 compris ?

Solution
Comme on ne tient pas compte de l'ordre dans lequel le tirage se fait, on calcule ici les combinaisons.

$$\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}$$

Les combinaisons possibles sont au nombre de

$$\binom{49}{6} = C_{49}^6 = \frac{49!}{6!43!} = 13\,983\,816.$$

Exercice 12.31. Euro Millions
À l'Euro Millions, un pronostic se compose de deux grilles, une grille comportant 50 numéros et une seconde grille comportant 9 étoiles numérotées de 1 à 9. Un joueur doit choisir 5 numéros parmi les 50 et 2 étoiles parmi les 9. Un montant de 3.20 francs suisses est dû pour chaque pronostic. Pierre rêve de devenir riche, aussi met-il en place une stratégie pour gagner à coup sûr à ce jeu, il jouera chacune des combinaisons possible. Quel montant devra-il verser pour faire enregistrer tous ses pronostics ?

Solution

Nombre de combinaisons — Le nombre de combinaisons de 5 nombres parmi 50 est donné par

$$\binom{50}{5} = C_{50}^5 = \frac{50!}{45! \times 5!} = 2\,118\,760$$

— Le nombre de combinaisons de 2 étoiles parmi 9 est donné par

$$\binom{9}{2} = C_9^2 = \frac{9!}{7! \times 2!} = 36$$

Donc, Pierre devra faire $2\,118\,760 \times 36 = 76\,275\,360$ pronostics afin de tous les avoir.

Montant à verser Il devra verser un montant de 244 081 152 francs suisses. Donc, plus de 244 millions.

Exercice 12.32. Poker
Un joueur joue au Poker Fermé avec 7 autres joueurs. Le paquet de cartes contient 52 cartes et chaque joueur reçoit 5 cartes. Dans les questions qui suivent, on considère 4 couleurs (coeur, pique, trèfle et carreau) et on considère 13 symboles (As, deux, . . ., roi).

1. En considérant l'ordre dans lequel le joueur reçoit les cartes, combien de mains différentes peut avoir le joueur ?
2. En ne considérant pas l'ordre des cartes, combien de mains différentes peut avoir le joueur ?
3. Combien de mains sont composées de 5 cartes de coeur ? Quelle est la probabilité d'obtenir cinq coeur ?
4. Quelle est la probabilité d'obtenir cinq cartes de la même couleur ?
5. Combien de mains sont composées de quatre As ? Quelle est la probabilité d'obtenir quatres As ?

Solution

1. Il s'agit d'un arrangement :
$$A_n^k = A_{52}^5 = \frac{52!}{(52-5)!} = 311\,875\,200.$$
2. Il s'agit d'une combinaison :
$$\binom{n}{k} = \binom{52}{5} = \frac{52!}{5!(52-5)!} = 2\,598\,960.$$
3. Nombre de mains avec 5 coeurs :
$$\binom{13}{5} = \frac{13!}{5!(13-5)!} = 1\,287.$$
Notons B : Avoir 5 cartes de coeur :
$$\Pr(B) = \frac{\binom{13}{5}}{\binom{52}{5}} = 0.0004.$$
4. Notons C : Avoir 5 cartes de la même couleur :
$$\Pr(C) = \frac{4 \times \binom{13}{5}}{\binom{52}{5}} = 0.0020.$$
5. On a quatre As et une cinquième carte. On a 48 possibilités pour la cinquième carte. Donc, 48 mains sont composées de quatre As. Soit l'événement A= "Obtenir quatre As",
$$\Pr(A) = \frac{48}{\binom{52}{5}} \simeq 0.$$

Exercice 12.33. Choix de films
On vous offre un bon pour le cinéma. Vous devez choisir 15 films parmi une liste de 20 films.

1. Combien de choix différents pouvez-vous faire ?
2. Même question si vous devez choisir 7 films parmi les 10 premiers films et 8 films parmi les 10 derniers.
3. Même question si vous devez choisir au moins 6 films parmi les 10 premiers films et le reste parmi les 10 derniers (nous devons toujours choisir 15 films en total).

Solution

1. Il s'agit d'une combinaison :
$$\binom{20}{15} = 15504.$$
2. Il s'agit de calculer toutes les possibilités de choisir 7 films parmi 10 puis de les multiplier par toutes les possibilités de choisir 8 films parmi 10 films :
$$\binom{10}{7}\binom{10}{8} = 5400.$$
3. La petite subtilité de cette question est la présence du "au moins".
Nous devons calculer toutes les possibilités que nous avons si nous décidons de choisir 6 films parmi les 10 premiers et 9 films parmi les 10 derniers puis ajouter toutes les possibilités que nous avons si nous décidons d'en choisir 7 parmi les 10 premiers et 8 parmi les 10 derniers puis ainsi successivement jusqu'à 10 parmi les 10 premiers et 5 parmi les 10 derniers.
$$\binom{10}{6}\binom{10}{9} + \binom{10}{7}\binom{10}{8} + \binom{10}{8}\binom{10}{7} + \binom{10}{9}\binom{10}{6} + \binom{10}{10}\binom{10}{5} = 15252.$$

Exercice 12.34. Cartes et mains
On pioche successivement et sans remise 5 cartes d'un jeu de 36 cartes. Calculer la probabilité d'obtenir une main contenant :

1. 3 trèfles et 2 coeurs,
2. le roi de coeur, 2 autres rois et 2 as,
3. au moins un as.
4. Calculer la probabilité d'obtenir, dans l'ordre, 3 trèfles puis 2 coeurs.

Solution
On introduit l'évènement A "obtenir 3 trèfles et 2 coeurs".

1. Pour les cas possibles, on choisit sans remise 5 cartes parmi 36 : $\binom{36}{5} = 376992$ choix possibles.
   Pour les cas favorables on choisit 3 trèfles parmi les 9 trèfles possibles : $\binom{9}{3} = 84$ et 2 coeurs parmi les 9 coeurs possibles : $\binom{9}{2} = 36$. Donc,
$$\Pr(A) = \frac{\binom{9}{3} \times \binom{9}{2}}{\binom{36}{5}} = \frac{84 \times 36}{376992} = 0.008.$$
2. le roi de coeur, 2 autres rois et 2 as :
$$\Pr(B) = \frac{\binom{1}{1} \times \binom{3}{2} \times \binom{4}{2}}{\binom{36}{5}} = \frac{1 \times 3 \times 6}{376992} = 0.00005.$$
3. Pas d'as :
$$\Pr(C) = 1 - \Pr(\overline{C} = \text{ "pas d'as" }) = 1 - \frac{\binom{32}{5}}{\binom{36}{5}} = 1 - \frac{201376}{376992} = 1 - 0.534 = 0.466.$$
4. Dans le cas présent, on doit faire un choix ordonné (arrangement), puisque l'on considère l'ordre d'arrivée des types de cartes. Pour les cas possibles, on choisi 5 cartes ordonnées distinctes parmi 36 (A_{36}^5 choix possibles) et pour les cas favorables, on choisi en premier trois cartes de trèfles ordonnées distinctes (A_9^3 choix possibles) puis pour les deux dernières cartes, on choisit deux cartes de coeurs ordonnées distinctes (A_9^2 choix possibles). Donc,
$$\Pr(A) = \frac{A_9^2 \times A_9^3}{A_{36}^5} = \frac{72 \times 504}{45239040} = 0.0041.$$

Exercice 12.35. Espérance et écart-type
Le Tableau 12.2 donne la distribution d'une variable aléatoire X. Calculer l'espérance et l'écart-type de X.

Table 12.2 – Distribution de probabilité de X

x	0	1	2	3	4	5	6	8
$\Pr(X = x)$	0.11	0.14	0.23	0.18	0.15	0.09	0.06	0.04

Solution
Le tableau :

x_i	p_i	$p_i x_i$	$p_i x_i^2$
0	0.11	0.00	0.00
1	0.14	0.14	0.14
2	0.23	0.46	0.92
3	0.18	0.54	1.62
4	0.15	0.60	2.40
5	0.09	0.45	2.25
6	0.06	0.36	2.16
8	0.04	0.32	2.56
Total	1.00	2.87	12.05

Les paramètres :

$$\mu = \mathrm{E}(X) = \sum_{i=1}^{n} p_i x_i = 2.87, \sigma^2 = \mathrm{var}(X) = \sum_{i=1}^{n} p_i x_i^2 - \mu^2 = 12.05 - 2.87^2 = 3.813, \sigma = \sqrt{3.813} = 1.953.$$

Exercice 12.36. Espérance et variance
Le Tableau 12.3 donne la distribution de probabilité de X.

Table 12.3 – Distribution de probabilité de X

x	3	4	5	6
$\Pr(X = x)$	0.2	0.4	0.3	0.1

1. Calculer l'espérance de X.
2. Calculer la variance de X.

Solution

1. L'espérance de X :
$$\mathrm{E}(X) = 3 \times 0.2 + 4 \times 0.4 + 5 \times 0.3 + 6 \times 0.1 = 4.3.$$
2. La variance de X :
$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = (3^2 \times 0.2 + 4^2 \times 0.4 + 5^2 \times 0.3 + 6^2 \times 0.1) - 4^2 = 19.3 - 18.49 = 0.81.$$

Exercice 12.37. Vol dans les magasins
Le nombre de personnes qui volent dans les magasins pendant les fêtes suit une loi de Poisson. 15% des personnes entrant dans un magasin en ressortent sans avoir payé la totalité de leurs articles.

1. Quelle est la probabilité que 10 personnes volent, s'il y en a 150 qui passent dans la journée ?
2. Quelle est la probabilité que 20 personnes volent, s'il y en a 150 qui passent dans la journée ?
3. Quelle est la probabilité que 20 personnes volent, s'il y en a 250 qui passent dans la journée ?

Solution

1. On a :
$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \Pr(X = 10) = \frac{e^{-\lambda}\lambda^{10}}{10!},$$
où λ est un taux moyen connu de $0.15 \times 150 = 22.5$.
Ainsi, on a que $\mathrm{E}(X) = 0.15 \times 150 = 22.5 = \lambda$.
$$\Pr(X = 10) = \frac{e^{-22.5} \times 22.5^{10}}{10!} = 0.0015.$$
2. $\lambda = 0.15 \times 150 = 22.5$
$$\Pr(X = 20) = \frac{e^{-22.5} \times 22.5^{20}}{20!} = 0.076.$$
3. $\lambda = 0.15 \times 250 = 37.5 = \mathrm{E}(X)$
$$\Pr(X = 20) = \frac{e^{-37.5} \times 37.5^{20}}{20!} = 0.00064.$$

Exercice 12.38. Jeu de dés
On lance un dé.

1. Soit X la variable aléatoire *"nombre de points du dé"*.
   (a) Donner la distribution de la variable aléatoire X.
   (b) Donner l'espérance mathématique de la variable aléatoire X. Est-ce une valeur possible (modalité) de la variable ?
   (c) Donner la variance de X.
2. On vous propose le jeu de hasard suivant : vous lancez un dé et vous gagnez 10.- si le nombre du dé est 1 ou 2 et vous perdez 5.- sinon. Donner la distribution et l'espérance de la variable Y="gain".
3. On vous propose deux modalités de jeu différentes basées sur le lancer d'un dé.

   Modalité A : Vous payez 2.- pour un lancer du dé, vous gagnez 4.- si le nombre de points obtenu est pair et 0.- sinon.

   Modalité B : Vous payez 1.- pour un lancer du dé, vous gagnez 5.- si le nombre de points obtenu est 3 et 0.- sinon.

   Quelle modalité choisissez-vous et pourquoi ?

Solution

1. — Distribution de la variable aléatoire X.

x	1	2	3	4	5	6
$p_X(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

   — Espérance mathématique de X.

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5.$$

   Il ne s'agit pas d'une valeur possible de la variable aléatoire car elle prend que des valeurs entières.

   — Variance de X :

$$\begin{aligned} \text{var}(X) &= E(X^2) - E(X)^2 \\ &= 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + 9 \times \frac{1}{6} + 16 \times \frac{1}{6} + 25 \times \frac{1}{6} + 36 \times \frac{1}{6} - 3.5^2 = 2.92. \end{aligned}$$

2. Distribution de la variable Y.

y	-5	10
$p_Y(y)$	$\frac{4}{6} = \frac{2}{3}$	$\frac{2}{6} = \frac{1}{3}$

   Espérance mathématique de la variable Y :

$$E(Y) = -5 \times \frac{2}{3} + 10 \times \frac{1}{3} = 0.$$

3. Plusieurs raisonnements sont possibles, comme par exemple
   — la modalité B car un lancer coûte moins cher et peut rapporter d'avantage que la modalité A,
   — Posons Z="gain à la fin du jeu" (tenant compte de la mise initiale).
   — Modalité A :

$$E(Z) = 2 \times \frac{1}{2} + \frac{1}{2} \times (-2) = 0.$$

   Ceci signifie, qu'en moyenne je ne perds rien et ne gagne rien en choisissant cette modalité.

   — Modalité B :

$$E(Z) = 4 \times \frac{1}{6} + (-1) \times \frac{5}{6} = -0.1\overline{6}.$$

   Ceci signifie, qu'en moyenne, je perds de l'argent en choisissant cette modalité.

   Je choisis dans ce cas la modalité A.

Exercice 12.39. Répartition et espérance
Le Tableau 12.4 donne les valeurs de la fonction de répartition d'une variable aléatoire X.

Table 12.4 – Distribution de probabilité de X

x	10	12	14	16	18	20
$F(x)$	0.13	0.35	0.66	0.84	0.99	1.00

1. Donner la distribution de X.
2. Calculer l'ésperance et l'écart-type de X.

Solution

1. Distribution de X :

x	10	12	14	16	18	20
$\Pr(X = x)$	0.13	0.22	0.31	0.18	0.15	0.01

2. Espérance :

$$\mathrm{E}(X) = \sum_{x \in \mathbb{Z}} x \times p_X(x) = 14.06.$$

Variance :

$$\mathrm{var}(X) = \sum_{x \in \mathbb{Z}} p_X(x) \times (x - \mu)^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 6.436.$$

Écart-type :

$$\sigma(X) = \sqrt{\mathrm{var}(X)} = 2.537.$$

Exercice 12.40. Assurance et prime
Une compagnie d'assurance offre une assurance avec une prime annuelle de P CHF. Un client n'a pas de réclamation avec une probabilité de 7/10. Il a une réclamation de 5000 CHF avec une probabilité de 2/10 et il en a une de 10000 CHF avec une probabilité de 1/10. Quelle doit être la prime minimale demandée pour assurer un profit à la compagnie d'assurance ?

Solution

Profit	P	$P-5000$	$P-10\,000$
Probabilité	7/10	2/10	1/10

On veut que l'espérance de gain soit positive. Soit G=le gain de la compagnie. On veut donc $\mathrm{E}(G) > 0$.

$$\begin{aligned}
\mathrm{E}(G) &= \frac{7}{10}P + \frac{2}{10}(P-5000) + \frac{1}{10}(P-10000) \\
&= \frac{7}{10}P + \frac{2}{10}P + \frac{1}{10}P - \frac{2}{10}5000 - \frac{1}{10}10000 \\
&= P - 1000 - 1000 \\
&= P - 2000
\end{aligned}$$

On veut $\mathrm{E}(G) > 0$. Donc,

$$P - 2000 > 0,$$

$$P > 2000.$$

Exercice 12.41. Urne et jetons
Une urne contient 90 jetons numérotés de 2 à 6. On tire aléatoirement un jeton de l'urne. Soit X, la variable aléatoire représentant le numéro inscrit sur le jeton sélectionné, à laquelle on associe la distribution de probabilité décrite dans le Tableau 12.5.

Table 12.5 – Distribution de probabilité de X

x	2	3	4	5	6
$p_X(x)$	$\frac{6}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{20}{90}$	$\frac{20}{90}$

1. (a) Calculer l'espérance de X.
   (b) Calculer la variance de X.
2. On répète maintenant cette expérience aléatoire 3 fois en replaçant à chaque fois le jeton sélectionné dans l'urne et on s'intéresse à la somme des 3 numéros obtenus. On suppose que ces trois tirages sont indépendants.
   (a) Calculer l'espérance de cette somme.
   (b) Calculer la variance de cette somme.
   (c) On sait qu'un 4 a été obtenu au premier tirage, quelle est alors l'espérance de la somme ?
   (d) On sait qu'un 2 a été obtenu au premier tirage, quelle est alors la variance de la somme ?

Solution

1. (a) Espérance :

$$\mathrm{E}(X) = \sum_{x \in S} x p(x) = 2 \times \frac{6}{90} + 3 \times \frac{12}{90} + 4 \times \frac{32}{90} + 5 \times \frac{20}{90} + 6 \times \frac{20}{90} = 4.4.$$

   (b) Variance

$$\mathrm{var}(X) = \sigma^2 = \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2.$$

$$\mathrm{E}(X^2) = 2^2 \times \frac{6}{90} + 3^2 \times \frac{12}{90} + 4^2 \times \frac{32}{90} + 5^2 \times \frac{20}{90} + 6^2 \times \frac{20}{90} = 20.711.$$

   Donc,

$$\mathrm{var}(X) = 20.711 - 4.4^2 = 1.351.$$

2. Soit $Y = X_1 + X_2 + X_3$ la somme des trois résultats obtenus.
   (a) $\mathrm{E}(Y) = \mathrm{E}(X_1 + X_2 + X_3) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \mathrm{E}(X_3) = 3\mathrm{E}(X) = 3 \times 4.4 = 13.2$.
   (b) $\mathrm{var}(Y) = \mathrm{var}(X_1 + X_2 + X_3)$.
   Or les trois tirages sont indépendants et donc
   $\mathrm{var}(X_1 + X_2 + X_3) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \mathrm{var}(X_3) = 3\mathrm{var}(X) = 3 \times 1.35 = 4.05$.
   (c) Soit $Z = 4 + X_2 + X_3$,

$$\mathrm{E}(Z) = 4 + \mathrm{E}(X_1 + X_2) = 4 + 2\mathrm{E}(X) = 12.8.$$

   (d) Soit $T = 2 + X_2 + X_3$,

$$\mathrm{var}(T) = \mathrm{var}(2 + X_2 + X_3) = \mathrm{var}(X_2 + X_3) = 2\mathrm{var}(X) = 2.7.$$

Exercice 12.42. Classe et échantillon
Sandra est dans une classe de 30 personnes. On y tire un échantillon de 5 personnes. Toutes les personnes ont la même probabilité d'être sélectionnées, soit 1/6. Sandra s'intéresse à sa présence dans l'échantillon tiré.

1. Combien y a-t-il d'échantillons possibles ? Combien d'échantillons contiennent Sandra ? Quelle est la probabilité que Sandra soit dans l'échantillon ?
2. Soit X la variable valant 1 si Sandra est dans l'échantillon et 0 sinon. Quelle est la loi de probabilité de X ? Donner l'espérance et la variance de la variable X.

Solution

1. — $C_{30}^{5} = \binom{30}{5} = \dfrac{30!}{5!(30-5)!} = 142\,506.$
   — $C_{29}^{4} = \binom{29}{4} = \dfrac{29!}{4!(29-4)!} = 23\,751.$
   — Soit S = "Sandra est dans l'échantillon. Alors $\Pr(S) = \dfrac{23\,751}{142\,506} = 1/6.$
2. — X suit une loi de Bernoulli. En effet,

$$X = \begin{cases} 0 & \text{avec probabilité} \quad p, \\ 1 & \text{avec probabilité} \quad 1-p, \end{cases}$$

   où $p = 1/6$.
   — $\mathrm{E}(X) = p = 1/6$.
   — $\mathrm{var}(X) = p(1-p) = 1/6 \times (1 - 1/6) = 5/6$.

Exercice 12.43. Germination
Le taux de germination d'une graine est de 90% (c'est-à-dire que chaque graine a une probabilité de 90% de germer). La germination d'une graine n'influence pas la germination d'une autre graine. Soit X la variable *"Nombre de graines germées dans un lot de 20 graines"*.

1. De quelle nature est la variable aléatoire X ?

Aide : La première graine a une probabilité Quelle est la loi de probabilité de la variable aléatoire X ? Justifier et donner les paramètres de la loi de X.

2. Quelle est la probabilité qu'exactement 18 graines germent ?
3. Quelle est la probabilité que 95% des graines du lot germent ?
4. Quelle est la probabilité qu'au moins 95% des graines du lot germent ?
5. Quelle est la probabilité que moins de 95% des graines du lot germent ?
6. Quel est le nombre espéré de graines germées sur le lot (c'est-à-dire, l'espérance de la variable X) ?
7. Quelle est la variance de la variable X ?
8. Dans cette question n est inconnu.
   (a) Exprimer $\Pr(X \geq 1)$ en fonction de n.
   (b) Combien de graines faut-il étudier pour être sûr à 99% d'obtenir au moins une graine germée ?

Solution

1. La variable aléatoire X suit une loi binomiale.
$$X \sim \mathcal{B}(20, 0.9).$$
Donc, les paramètres sont $p = 0.9$ et $n = 20$.
En effet, on renouvelle 20 fois de manière indépendante une épreuve de Bernoulli de paramètre 0.9, qui est la probabilité de succès (ici la germination).
2. On rappelle que si $X \sim \mathcal{B}(n, p)$, alors on a
$$\Pr(X = x) = \binom{n}{x} p^x q^{(n-x)}.$$
Probabilité qu'exactement 18 graines germent :
$$\Pr(X = 18) = \binom{20}{18} 0.9^{18} \times 0.1^2 = \frac{20!}{18!2!} \times 0.9^{18} \times 0.1^2 = 0.285$$
3. Probabilité que 95% des graines du lot germent
   — 95% des graines du lot représente $\frac{95}{100} \times 20 = 19$ graines.
   — La probabilité cherchée est ainsi
$$\Pr(X = 19) = \binom{20}{19} 0.9^{19} \times 0.1 = 20 \times 0.9^{19} \times 0.1 = 0.270.$$
4. Probabilité qu'au moins 95% des graines du lot germent
$$\begin{aligned}\Pr(X \geq 19) &= \Pr(X = 19) + \Pr(X = 20) = 0.27 + \binom{20}{20} 0.9^{20} \times 0.1^0 \\ &= 0.27 + 0.9^{20} = 0.392.\end{aligned}$$
5. Probabilité que strictement moins de 95% des graines du lot germent
$$\Pr(X < 19) = 1 - \Pr(X \geq 19) = 1 - 0.392 = 0.608.$$
6. Nombre espéré de graines germées sur le lot (espérance de la variable X).
Comme $X \sim \mathcal{B}(20, 0.9)$, on a $\mathrm{E}(X) = np = 20 \times 0.9 = 18$.
7. Variance de la variable X.
Comme $X \sim \mathcal{B}(20, 0.9)$, on a $\mathrm{var}(X) = npq = 20 \times 0.9 \times 0.1 = 1.8$.
8. Dans cette question n est inconnu.
   (a) $\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - \binom{n}{0} \times 0.9^0 \times 0.1^n = 1 - 0.1^n$
   (b) on cherche n minimum tel que $\Pr(X \geq 1) \geq 0.99$ et on a $\Pr(X \geq 1) = 1 - 0.1^n$.
   Donc, $1 - 0.1^n \geq 0.99 \Rightarrow 0.1^n \leq 0.01 \Rightarrow n \times \log(0.1) \leq \log(0.01) \Rightarrow n \geq 2$. $n = 2$ convient.

Exercice 12.44. Pièce de monnaie
Nous lançons 10 fois une pièce de monnaie. À chaque fois il y a deux possibilités : Pile (P) et Face (F).

1. Quelle est la probabilité d'avoir dans l'ordre suivant : PPPPPPPFFF ?
2. Quelle est la probabilité d'avoir 7 piles et 3 faces ? Expliquez en quoi cette question est différente de la précédente.

Solution

1. Etant donné que les tirages sont indépendants nous avons :

$$\Pr(\text{PPPPPPPFFF}) = 0.5^{10} \approx 0.000977.$$

2. Nous devons utiliser la loi binomiale. Pour rappel :

$$\Pr(\text{X}{=}x) = \binom{n}{x} \times p^x \times q^{n-x}.$$

Dans notre cas, n=10, x=7, p=0.5 et q=0.5. Donc,

$$\Pr(\text{avoir 7 piles et 3 faces}) = \binom{10}{7} \times 0.5^7 \times 0.5^3 \approx 0.117.$$

Dans la première question on nous demande quelle est la probabilité d'avoir 7 piles et 3 faces dans un ordre spécifique. Dans la deuxième on nous demande quelle est la probabilité d'avoir 7 piles et 3 faces dans n'importe quel ordre (l'événement est plus large).

Exercice 12.45. Taux de réussite
Le taux de réussite (c'est-à-dire pour une note supérieur ou égale à 4) aux examens est de 75%. On considère que la réussite d'un étudiant ne dépend pas de celle d'un autre. Soit X la variable aléatoire *"nombre de réussites dans une classe de 30"*.

1. Quelle est la loi de la variable X ? Donner les paramètres et justifier le choix de loi.
2. Quelle est la probabilité qu'exactement 20 élèves réussissent ?
3. Quelle est l'espérance de X ?
4. Quelle est la variance de X ?

Solution

1. Il s'agit d'une variable binomiale. Les paramètres sont $p = 0.75$ et $n = 30$. Il y a 30 étudiants différents (donc 30 essais indépendants). Chaque étudiant peut réussir(1) ou échouer(0). La probabilité de réussite pour chaque étudiant est 0.75.
2. La probabilité qu'exactement 20 élèves réussissent :
$$\Pr(X = 20) = \frac{30!}{20!(30-20)!} \times 0.75^{20} \times 0.25^{30-20} = 30045015 \times 0.75^{20} \times 0.25^{10} = 0.091.$$
3. L'espérance d'une variable binomiale X est égale à np, c'est-à-dire :
$n \times p = 30 \times 0.75 = 22.5$.
4. La variance d'une variable binomiale X est égale à npq, où $q = 1 - p$, c'est-à-dire :
$npq = 30 \times 0.75 \times 0.25 = 5.625$.

Exercice 12.46. Excès de vitesse et loi de Poisson
Le nombre d'automobilistes en excès de vitesse par jour à un endroit précis suit une loi de Poisson et représente en moyenne 10% des automobilistes passant à cet endroit. Un radar y est placé.

1. Quelle est la probabilité que 12 automobilistes soient amendables si en moyenne 100 automobilistes empruntent cette route ?
2. Quelle est la probabilité que 10 automobilistes soient amendables ce même jour ?
3. Mêmes questions si en moyenne 150 automobilistes empruntent cette route.

Solution

1. Si en moyenne 100 automobilistes empruntent cette route, la variable X n= "nombre d'automobilistes amendables" suit une loi de Poisson de paramètre 10. En effet, sait que $X \sim \mathcal{P}(\lambda)$ et ainsi $\mathrm{E}(X) = \lambda$. De plus, $\mathrm{E}(X) = 10\% \times 100 = 10$, d'où $X \sim \mathcal{P}(10)$. Les probabilités cherchées sont ainsi
$$\Pr(X = 12) = \frac{e^{-\lambda} \times \lambda^{12}}{12!} = \frac{e^{-10} \times 10^{12}}{12!} = 0.095.$$
2. Et
$$\Pr(X = 10) = \frac{e^{-10} \times 10^{10}}{10!} = 0.125.$$
3. Si en moyenne 150 automobilistes empruntent cette route, la variable X = "nombre d'automobilistes amendables" suit une loi de Poisson de paramètre 15. En effet, ceci est obtenu en procédant de la même manière que précédemment à la différence près que dans ce cas
$$\mathrm{E}(X) = 10\% \times 150 = 15 = \lambda.$$
Les probabilités cherchées sont ainsi
$$\Pr(X = 12) = \frac{e^{-15} \times 15^{12}}{12!} = 0.083 \Pr(X = 10) = \frac{e^{-15} \times 15^{10}}{10!} = 0.049.$$

Exercice 12.47. Employés absents pour cause de maladie et loi de Poisson
Dans une entreprise de la région, le nombre d'employés absents chaque mois pour cause de maladie s'élève à 2 en moyenne et suit une loi de Poisson.

1. Quelle est la probabilité qu'aucun des employés ne soit absent pour cause de maladie le mois prochain ?
2. Quelle est la probabilité que le nombre d'employés absents le mois prochain soit supérieur ou égal à 2 ?

Solution
Soit la variable X= "nombre d'employés absents pour cause de maladie le mois prochain". On sait que $X \sim \mathcal{P}(2)$.

1. Probabilité qu'aucun des employés ne soit absent pour cause de maladie le mois prochain

$$\Pr(X = 0) = \frac{e^{-2} \times 2^0}{0!} = e^{-2} = 0.135.$$

2. Probabilité que le nombre d'employés absents le mois prochain soit supérieur ou égal à 2. On a

$$\Pr(X \geq 2) = 1 - \Pr(X < 2).$$

Or,

$$\Pr(X < 2) = \Pr(X = 0) + \Pr(X = 1) = 0.135 + \frac{e^{-2} \times 2}{1!} = 0.135 + 0.271 = 0.406.$$

Donc,

$$\Pr(X \geq 2) = 1 - 0.406 = 0.594.$$

Exercice 12.48. File d'attente et loi de Poisson
On sait que, en moyenne, six personnes par heure se présentent au guichet du service-client d'un grand magasin. En admettant que la variable aléatoire X, représentant le nombre de personne se présentant au guichet chaque heure, suit une loi de Poisson :

1. Donner l'espérance et la variance de la variable aléatoire X.
2. Calculer la probabilité pour que exactement 6 clients se présentent au guichet durant une heure.
3. Calculer la probabilité qu'il y au moins 2 clients en une heure.
4. Calculer la probabilité qu'il n'y ait aucun client durant une période de vingt minutes.

Solution

1. Si $X \sim \mathcal{P}(\lambda)$: $\mathrm{E}(X) = \lambda = 6$ et $\mathrm{var}(X) = \lambda = 6$.
2. Probabilité pour que exactement 6 clients se présentent :
$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ donc } \Pr(X = 6) = \frac{e^{-6}6^6}{6!} = 0.161.$$
3. probabilité qu'il y au moins 2 clients :
$$\begin{aligned}\Pr(X \geq 2) &= 1 - \Pr(X < 2) = 1 - [\Pr(X = 0) + \Pr(X = 1)] = 1 - \left(\frac{e^{-6}6^0}{0!} + \frac{e^{-6}6^1}{1!}\right)\\ &= 1 - [0.0025 + 0.0149] = 0.9826.\end{aligned}$$
4. Ici, pour une durée de 20 minutes, la moyenne $\mathrm{E}(X) = \lambda = 2$:
$$\Pr(X = 0) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ donc } \Pr(X = 0) = \frac{e^{-2}2^0}{0!} = 0.135.$$

Exercice 12.49. Table de la loi normale 1
Déterminer les valeurs j de la variable normale centrée réduite Z telles que :

1. $\Pr(Z \leq j) = 0.9968$,
2. $\Pr(j \leq |Z|) = 0.9876$,
3. $\Pr(Z \leq j) = 0.3300$,
4. $\Pr(Z \geq j) = 0.1251$,
5. $\Pr(j \leq Z \leq 3) = 0.5147$.

Solution
Lecture inverse de la table.

1. $\Pr[Z \leq j] = 0.9968 \Rightarrow \Phi(j) = 0.9968 \Rightarrow j = 2.74$
2. $\Pr(j \leq |Z|) = 0.9876$
   $\Pr(j \leq |Z|) = \Pr(j \leq Z) + \Pr(Z \leq -j) = \Pr(j \leq Z) + \Pr(j \leq Z) = 2\Pr(j \leq Z)$
   $= 2\left[1 - \Pr(j \leq Z)\right] = 2\left[1 - \Phi(j)\right] = 0.9876$
   $\Rightarrow \Phi(j) = 1 - \frac{0.9876}{2} = 0.5062$.
3. $\Pr(Z \leq j) = 0.3300$
   $\Pr(Z \leq -j) = 0.3300$
   $1 - \Pr(Z \leq -j) = 0.3300$
   $1 - \Phi(-j) = 0.3300$
   $\Phi(-j) = 0.6700$
   $-j = 0.44$
   $j = -0.44$.
4. $\Pr(Z \geq j) = 0.1251$
   $1 - \Pr(Z \leq j) = 0.1251$
   $\Phi(j) = 1 - 0.1251 = 0.8749$
   $j = 1.15$.
5. $\Pr(j \leq Z \leq 3) = 0.5147$
   $\Pr(Z \leq 3) - \Pr(Z \leq j) = 0.5147$
   $\Phi(3) - 0.5147 = \Pr(Z \leq j)$
   $0.9987 - 0.5147 = 0.4840 = \Pr(Z \leq j)$
   $\Pr(Z > -j) = 0.4840$
   $1 - \Pr(Z \leq -j) = 0.4840$
   $\Phi(-j) = 1 - 0.4840 = 0.5160$
   $-j = 0.04$
   $j = -0.04$.

Exercice 12.50. Table de la loi normale 2
Soit $Z \sim \mathcal{N}(0,1)$. Déterminer :

1. $\Pr[Z \leq 1.23]$,
2. $\Pr[Z \leq -1.23]$,
3. $\Pr[Z \in [0.36, 1.23]]$,
4. $\Pr[Z \in [-0.88, 1.23]]$,
5. $\Pr[Z > 2.65 \text{ ou } Z \leq -1.49]$.

Solution

1. $\Pr[Z \leq 1.23] = \Phi(1.23) = 0.8907$.
2. $\Pr[Z \leq -1.23] = 1 - \Phi(1.23) = 0.1093$.
3. $\Pr[Z \in [0.36, 1.23]] = \Phi(1.23) - \Phi(0.36) = 0.8907 - 0.6406 = 0.2501$.
4. $\Pr[Z \in [-0.88, 1.23] = \Phi(1.23) - \Phi(-0.88) = 0.8907 - (1 - \Phi(0.88)) = 0.8907 - 0.1894 = 0.7013$.
5. $\Pr[Z > 2.65 \text{ ou } Z \leq -1.49] = \Pr[Z > 2.65] + \Pr[Z \leq -1.49] = 1 - \Phi(2.65) + \Phi(-1.49)$
   $= 1 - \Phi(2.65) + 1 - \Phi(1.49) = 2 - 0.9960 - 0.9319 = 0.0721$.

Exercice 12.51. Table de la loi normale 3
Soit $Z \sim \mathcal{N}(0,1)$. Déterminer :

1. $\Pr(Z \leq 0.64)$,
2. $\Pr(Z > -0.64)$,
3. $\Pr(Z \leq -2.20)$,
4. $\Pr(1.06 \leq Z \leq 1.99)$,
5. $\Pr(-1.37 \leq Z \leq 0.45)$,
6. $\Pr(|Z| > 2.10)$,
7. $\Pr(4.33 \leq Z)$.

Solution

1. $\Pr(Z \leq 0.64) = \Phi(0.64) = 0.7389$, où $\Phi(\cdot)$ est la fonction de répartition d'une normale centrée réduite.
2. $\Pr(Z > -0.64) = \Pr(Z \leq 0.64) = \Phi(0.64) = 0.7389$.
3. $\Pr(Z \leq -2.20) = \Pr(Z > 2.20) = 1 - \Pr(Z \leq 2.20) = 1 - \Phi(2.29) = 1 - 0.9861 = 0.0139$.
4. $\Pr(1.06 \leq Z \leq 1.99) = \Pr(Z \leq 1.99) - \Pr(Z \leq 1.06) = \Phi(1.99) - \Phi(1.06) = 0.9767 - 0.8554 = 0.1213$.
5. $\Pr(-1.37 \leq Z \leq 0.45)\Pr(Z \leq 0.45) - \Pr(Z \leq -1.37) = \Pr(Z \leq 0.45) - \Pr(Z > 1.37)$
   $= \Pr(Z \leq 0.45) - [1 - \Pr(Z \leq 1.37)] = \Phi(0.45) + \Phi(1.37) - 1 = 0.6736 + 0.9147 - 1 = 0.5883$.
6. $\Pr(|Z| > 2.10) = \Pr(Z > 2.10) + \Pr(Z \leq -2.10) = \Pr(Z > 2.10) + \Pr(Z > 2.10) = 2\,[1 - \Pr(Z \leq 2.10)]$
   $= 2\,[1 - \Phi(2.10)] = 2(1 - 0.9821) = 0.0358$.
7. $\Pr(4.33 \leq Z) = 1$.

Exercice 12.52. Table de la loi normale 4
Soit $Z \sim \mathcal{N}(0,1)$. Déterminer :

1. $\Pr(Z \leq 2.24)$,
2. $\Pr(Z \leq -2.24)$,
3. $\Pr(-0.6 \leq Z \leq 1.52)$.
4. Déterminer ce que vaut j : $\Pr(Z \leq j) = 0.9406$.
5. Déterminer ce que vaut j : $\Pr(j \leq Z \leq 1.34) = 0.2882$.

Solution

1. $\Pr(Z \leq 2.24) = \Phi(2.24) = 0.9875$.
2. $\Pr(Z \leq -2.24) = 1 - \Phi(2.24) = 0.0125$.
3. $\Pr(-0.6 \leq Z \leq 1.52) = \Phi(1.52) - (1 - \Phi(0.6)) = 0.9357 - (1 - 0.7257) = 0.9357 - 0.2743 = 0.6614$.
4. $\Pr(Z \leq j) = 0.9406 \Rightarrow \Phi(j) = 0.9406 \Rightarrow j = 1.56$.
5. $\Pr(j \leq Z \leq 1.34) = 0.2882 \Rightarrow \Phi(1.34) - \Phi(j) = 0.2882 \Rightarrow 0.9099 - \phi(j) = 0.2882 \Rightarrow j = 0.31$.

Exercice 12.53. Lecture inverse de la table de la loi normale
Déterminer les valeurs j de la variable normale centrée réduite Z telles que :

1. $\Pr[Z \leq j] = 0.9332$,
2. $\Pr[-j \leq Z \leq j] = 0.3438$,
3. $\Pr[Z \leq j] = 0.0125$,
4. $\Pr[Z \geq j] = 0.0125$,
5. $\Pr[j \leq Z \leq 3] = 0.7907$.

Solution
Lecture inverse de la table.

1. $\Pr[Z \leq j] = 0.9332 \Rightarrow F(j) = 0.9332 \Rightarrow j = 1.5$.
2. $\Pr[-j \leq Z \leq j] = 0.3438 \Rightarrow F(j) - F(-j) = F(j) - 1 + F(j) = 2F(j) - 1 = 0.3438$
   $F(j) = 0.6719 \Rightarrow j = 0.45$.
3. $\Pr[Z \leq j] = 0.0125 \Rightarrow F(j) = 0.0125$ (j est négatif)
   $1 - F(-j) = 0.0125 \Rightarrow F(-j) = 0.9875 \Rightarrow j = -2.24$.
4. $\Pr[Z \geq j] = 0.0125 = 1 - F(j) \Rightarrow F(j) = 0.9875 \Rightarrow j = 2.24$.
5. $\Pr[j \leq Z \leq 3] = 0.7907 = F(3) - F(j) \Rightarrow 0.7907 = 0.9987 - F(j) \Rightarrow F(j) = 0.2080$ (j est négatif)
   $F(-j) = 0.7920 \Rightarrow -j = 0,81 \Rightarrow j = -0.81$.

Exercice 12.54. Résultat et loi normale
Soit une variable aléatoire $X \sim \mathcal{N}(53, \sigma^2 = 100)$ représentant le résultat d'un examen pour un étudiant d'une section. Déterminez la probabilité pour que le résultat soit compris entre 33.4 et 72.6.

Solution

$$\text{Soit } X \sim \mathcal{N}(53, , \sigma^2 = 100) \Rightarrow Z = \frac{X - 53}{10} \sim \mathcal{N}(0, 1).$$

$$\begin{aligned}
\Pr[33.4 \leq X \leq 72.6] &= \Pr\left[\frac{33.4 - 53}{10} \leq \frac{X - 53}{10} \leq \frac{72.6 - 53}{10}\right] \\
&= \Pr[-1.96 \leq Z \leq 1.96] = 2\mathrm{F}(1.96) - 1 = 2 \;\; 0.975 - 1. = 0.95.
\end{aligned}$$

Exercice 12.55. Variable normale standardisation
Soit une variable aléatoire X $\mathcal{N}(50; \sigma^2 = 100)$. Déterminez le premier quartile de cette distribution.

Solution
Si $X \sim \mathcal{N}(50, 10)$, alors $Z = (X - 50)/10 \sim \mathcal{N}(0, 1)$. Par définition le premier quartile $x_{1/4}$ est tel que

$$\Pr\left[X \leq x_{1/4}\right] = 1/4.$$

Donc,

$$\Pr\left[X \leq x_{1/4}\right] = P\left[\frac{X - 50}{10} \leq \frac{x_{1/4} - 50}{10}\right] = P[Z \leq z_{1/4}] = 0.25,$$

où $z_{1/4}$ est le premier quartile d'une variable aléatoire normale centrée réduite. Si $\Phi(.)$ est la fonction de répartition d'une variable aléatoire normale centrée réduite, on a par la définition du quartile que

$$\Phi(z_{1/4}) = 0.25.$$

Le premier quartile $z_{1/4}$ est donc négatif. On a cependant. par la symétrie de la distribution, que

$$\Phi(z_{1/4}) = 1 - \Phi(-z_{1/4}) = 0.25,$$

ce qui donne

$$\Phi(-z_{1/4}) = 0.75.$$

La table nous donne que $-z_{1/4} = 0.67$ et donc $z_{1/4} = -0.67$. Enfin, comme

$$\frac{x_{1/4} - 50}{10} = z_{1/4} = -0.67,$$

on a une équation en $x_{1/4}$ qu'il suffit de résoudre

$$x_{1/4} = 50 - 0.67 \times 10 = 43.3.$$

Exercice 12.56. Tailles et loi normale
En supposant que les tailles en cm des étudiants d'un pays admettent la distribution normale $\mathcal{N}(172, 9)$. On demande de déterminer le pourcentage théorique :

1. d'étudiants mesurant au moins 180 cm.
2. d'étudiants dont la taille est comprise entre 168 et 180.

Solution
Si $X \sim \mathcal{N}(172, 9)$, alors $Z = (X - 172)/3 \sim \mathcal{N}(0, 1)$.

1. $\Pr[X > 180] = \Pr[Z > 2.67] = 1 - \Pr[Z \leq 2.67] = 1 - \Phi(2.67) = 1 - 0.9962 = 0.0038$.
2. $\Pr[X \in [168, 180]] = \Pr[Z \in [-1.33, 2.67]] = \Phi(2.67) - \Phi(-1.33)$
$= \Phi(2.67) - (1 - \Phi(1.33)) = 0.9962 - 1 + 0.9082 = 0.9044$.

Exercice 12.57. Temps et loi normale
On suppose que le temps en minutes pour se rendre à l'école d'un étudiant suit une loi normale $\mathcal{N}(45, 9)$.

1. Déterminer la proportion des trajets d'une durée de plus de 60 minutes.
2. Déterminer la proportion des trajets d'une durée entre 40 et 50 minutes.
3. Déterminer la durée telle que 2.5% des trajets sont plus courts que cette durée.

Solution
Soit X la durée d'un trajet, avec $X \sim \mathcal{N}(45, 9)$.

1. $\Pr(X > 60) = \Pr\left(\frac{X-45}{3} > \frac{60-45}{3}\right) = \Pr(Z > 5) = 0$.
2. $\Pr(40 \leq X \leq 50) = \Pr\left(\frac{40-45}{3} \leq \frac{X-45}{3} \leq \frac{50-45}{3}\right) = \Pr(-1.67 \leq Z \leq 1.67)$
$= \Pr(Z \leq 1.67) - \Pr(Z \leq -1.67) = \Pr(Z \leq 1.67) - \Pr(Z > 1.67) = \Pr(Z \leq 1.67) - [1 - \Pr(Z \leq 1.67)]$
$= 2\Pr(Z \leq 1.67) - 1 = 2(0.9525) - 1 = 0.9050$.
3. $\Pr(X \leq j) = 0.0250$.
$\Pr\left(Z \leq \frac{j-45}{3}\right) = 0.0250$.
$\Pr\left(Z > -\frac{j-45}{3}\right) = 0.0250$.
$\Pr\left(Z \leq -\frac{j-45}{3}\right) = 1 - 0.0250 = 0.9750$.
$-\dfrac{j-45}{3} = 1.96$.
$j = -1.96 \times 3 + 45 = 39.12$.

Exercice 12.58. Vitesse et loi normale
Sur une route principale où la vitesse est limitée à 80 km/h, un radar a mesuré la vitesse de toutes les automobiles pendant une journée. En supposant que les vitesses recueillies soient distribuées normalement avec une moyenne de 72 km/h et un écart-type de 8 km/h, quelle est approximativement la proportion d'automobiles ayant commis un excès de vitesse ?

Solution
La proportion d'automobiles ayant commis un excès de vitesse vaut

$$P[X > 80] = 1 - P[X \leq 80] = 1 - P\left[\frac{X - \bar{x}}{s} \leq \frac{80 - 72}{8}\right] = 1 - P[Z \leq 1] = 0.159,$$

où X représente la vitesse.

Exercice 12.59. Consommation d'eau et loi normale
En supposant que la consommation d'eau journalière en m^3, par habitant, en Suisse, suit une loi normale $\mathcal{N}(100, 16)$.

1. Quel est le pourcentage théorique d'habitants qui consomment plus de 110 m^3 ?
2. Quel est le pourcentage théorique d'habitants qui consomment entre 90 et 110 m^3 ?
3. Quel est la probabilité qu'en choisissant une personne au hasard, elle consomme 100 m^3 ou moins ?
4. Une consommation est considéré comme "hors norme" si elle s'écarte de plus de 13 m^3 de la moyenne. Calculer le pourcentage d'habitants "hors norme".

Solution
Si X$\sim \mathcal{N}(100, 16)$, alors $Z = \dfrac{(X-\mu)}{\sigma} = \dfrac{(X-100)}{\sqrt{16}} \sim \mathcal{N}(0,1)$:

1. Pourcentage théorique d'habitants qui consomment plus de 110 m^3 ?

$$\begin{aligned}\Pr(X > 110) &= \Pr\left(Z > \frac{110-100}{\sqrt{16}}\right)\\ &= \Pr(Z > 2.5) = 1 - \Pr(Z \leq 2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062.\end{aligned}$$

2. Pourcentage théorique d'habitants qui consomment entre 90 et 110 m^3 :

$$\begin{aligned}\Pr(90 \leq X \leq 110) &= \Pr\left(\frac{90-100}{\sqrt{16}} \leq Z \leq \frac{110-100}{\sqrt{16}}\right)\\ &= \Pr(-2.5 \leq Z \leq 2.5) = \Phi(2.5) - \Phi(-2.5) = \Phi(2.5)(1-\Phi(2.5)).\\ &= 0.9938 - 0.0062 = 0.9876.\end{aligned}$$

3. Probabilité qu'en choisissant une personne au hasard, elle consomme 100 m^3 ou moins :

$$\Pr(X \leq 100) \Rightarrow \Pr\left(Z \leq \frac{100-100}{\sqrt{16}}\right) = \Pr(Z \leq 0) = \Phi(0) = 0.5.$$

4. Pourcentage d'habitants "hors norme".

$$\begin{aligned}&\Pr(X \leq 87) + \Pr(X > 113) = \Pr(X \leq 87) + (1 - \Pr(X \leq 113))\\ &= \Pr\left(Z \leq \frac{87-100}{\sqrt{16}}\right) + \left(1 - \Pr\left(Z \leq \frac{113-100}{\sqrt{16}}\right)\right)\\ &= \Pr(Z \leq -3.25) + (1 - \Pr(Z \leq 3.25)) = (1 - \Phi(3.25) + (1 - \Phi(3.25))\\ &= (1 - 0.9994) + (1 - 0.9994) = 2 \times 0.0006 = 0.0012.\end{aligned}$$

Exercice 12.60. Boîte de conserve et loi normale
Une entreprise fabrique en série des boîtes de conserve qui devraient toutes mesurer 15 cm de haut. On note X la variable aléatoire prenant pour valeur la hauteur d'une boîte de conserve. On suppose que celle-ci est une normale de moyenne 15 cm et d'écart-type 0.2 cm.

1. Calculer la probabilité qu'une boîte choisie au hasard dans la production ait une hauteur inférieure à 14.8 cm.
2. Calculer la probabilité qu'une boîte choisie au hasard dans la production ait une hauteur qui s'écarte de moins de 0.2 cm de la hauteur désirée (15cm).
3. On considère une boîte comme défectueuse si sa hauteur s'écarte de plus de 3% par rapport à la hauteur désirée. Dans ce cas la boîte doit être jetée. Calculer le pourcentage de pièces défectueuses dans la production.
4. Pour rentabiliser cette production, un gestionnaire considère qu'il ne faudrait jeter que 1% de la production. Quelle serait, dans ce cas, la hauteur minimale et maximale des boites qui ne sont pas jetées.

Solution

1.
$$\begin{aligned}\Pr(X \leq 14.8) &= \Phi\left(\frac{14.8-15}{0.2}\right) = \Phi\left(\frac{-0.2}{0.2}\right)\\ &= \Phi(-1) = 1-\Phi(1)\\ &= 1-0.8413 = 0.1587\end{aligned}$$

2.
$$\begin{aligned}\Pr(15-0.2 \leq X \leq 15+0.2) &= \Pr(14.8 \leq X \leq 15.2)\\ &= \Pr(X \leq 15.2) - \Pr(X \leq 14.8)\\ &= \Phi(1) - (1-\Phi(1)) = 2\Phi(1)-1\\ &= 2\times 0.8413 - 1 = 0.6826\end{aligned}$$

3. Une boîte est défectueuse si sa hauteur diffère de plus de $3\% \times 15 = 0.45cm$. La probabilité qu'une boîte choisie au hasard soit défectueuse est donnée par
$$\begin{aligned}\Pr(X \leq 14.55) + \Pr(X \geq 15.45) &= \Pr(X \leq 14.55) + 1 - \Pr(X \leq 15.45)\\ &= \Phi\left(\frac{-0.45}{0.2}\right) + 1 - \Phi\left(\frac{0.45}{0.2}\right)\\ &= \Phi(-2.25) + 1 - \Phi(2.25)\\ &= 1-\Phi(2.25) + 1 - \Phi(2.25)\\ &= 2 - 2\Phi(2.25)\\ &= 2 - 2\times 0.9878 = 0.0244\end{aligned}$$
Dans la production, 2.44% des boîtes sont défectueuses.
4. Soit y un écartement à la valeur de 15cm. On cherche une condition sur y pour que que la probabilité qu'une pièce choisie au hasard dans la production s'écarte de y vaille moins de 1%. On cherche donc une condition sur y pour avoir
$$\begin{aligned}\Pr(X \leq 15-y) + \Pr(X \geq 15+y) &< 0.01\\ \Phi\left(\frac{15-y-15}{0.2}\right) + 1 - \Phi\left(\frac{15+y-15}{0.2}\right) &< 0.01\\ \Phi\left(\frac{-y}{0.2}\right) + 1 - \Phi\left(\frac{y}{0.2}\right) &< 0.01\\ 1-\Phi(5y) + 1 - \Phi(5y) &< 0.01\\ 2 - 2\Phi(5y) &< 0.01\\ \Phi(5y) &> 0.995\end{aligned}$$
Il suffit d'imposer la condition $5y > 2.58$, c'est-à-dire $y > 0.516$, pour que moins de 1% des pièces de la production soient défectueuses, soit, en pourcentage, une tolérance de 3.44%.

Exercice 12.61. Cylindres et loi normale
Pour l'assemblage d'une machine, on produit des cylindres dont le diamètre varie d'après une loi normale de moyenne 10 cm et d'écart-type 0.2 cm. On groupe les cylindres en 3 catégories :
- A : défectueux et inutilisable si le diamètre est ≤ 9.95, le cylindre est alors détruit.
- B : utilisable et vendu au prix réduit de Fr. 5.-, si 9.95 le diamètre ≤ 9.99.
- C : correspond aux normes et est vendu Fr. 15.-, si le diamètre est > 9.99.

1. Calculer les proportions de cylindres produits de chaque type A, B et C.
2. La production d'un cylindre coûte Fr. 7.-. Quel est le profit moyen par cylindre produit ?

Solution

1. Soit X le diamètre, ainsi $X \sim \mathcal{N}(10, 0.2^2)$,

$$\Pr[X \leq 9.95] = P\left[\frac{X-10}{0.2} \leq 0.25\right] = 0.401,$$

$$\Pr[9.95 < X \leq 9.99] = P\left[-0.25 < \frac{X-10}{0.2} \leq -0.05\right] = 0.079,$$

$$\Pr[X > 9.99] = 1 - (\Pr[X \leq 9.95] + \Pr[9.95 < X \leq 9.99] = 0.52.$$

2. Profit moyen $= 5 \times 0.079 + 15 \times 0.52 - 7 = 1.195$ fr.

Exercice 12.62. Théorème des probabilités totales
La compagnie d'assurance pour laquelle vous travaillez décide de se lancer dans l'assurance automobile. Dans cette optique, on vous charge notamment d'estimer la probabilité qu'un assuré ait un accident durant une année.
Pour ce faire, vous commencez par répartir les assurés dans trois groupes : individus très enclins à un accident (groupe à haut risque), individus moyennement enclins à un accident (groupe à risque intermédiaire) et individus faiblement exposés à un accident (groupe à faible risque).
Les probabilités qu'un individu du groupe à haut risque ait un accident dans l'année, qu'un individu du groupe à risque intermédiaire ait un accident dans l'année et qu'un individu du groupe à faible risque ait un accident dans l'année sont respectivement 0.1, 0.01 et 0.005.
Vous estimez que 20% des individus font partie du groupe à haut risque, que 10% font partie du groupe à faible risque et que les autres individus font partie du groupe à risque intermédiaire.
Quelle est, dans ce cas de figure, la probabilité qu'un nouvel assuré ait un accident dans l'année ?

Solution
On choisit un individu au hasard. Posons
- G_1= l'individu appartient au groupe à haut risque,
- G_2= l'individu appartient au groupe à risque intermédiaire,
- G_3= l'individu appartient au groupe à faible risque et
- A= l'individu a un accident durant l'année.

On a alors

$$\begin{array}{ll} \Pr(A|G_1) = 0.1 & \Pr(G_1) = 0.2 \\ \Pr(A|G_2) = 0.01 & \Pr(G_2) = 0.7 \\ \Pr(A|G_3) = 0.005 & \Pr(G_3) = 0.1. \end{array}$$

Par le théorème des probabilités totales, on a

$$\begin{aligned} \Pr(A) &= \sum_{i=1}^{3} \Pr(A|G_i)\Pr(G_i) \\ &= \Pr(A|G_1)\Pr(G_1) + \Pr(A|G_2)\Pr(G_2) + \Pr(A|G_3)\Pr(G_3) \\ &= 0.1 \times 0.2 + 0.01 \times 0.7 + 0.005 \times 0.1 \\ &= 0.0275 \end{aligned}$$

Ainsi, la probabilité qu'un nouvel assuré ait un accident dans l'année est de 0.0275, c'est-à-dire non loin de 3%.

Exercice 12.63. Théorème de Bayes
On se remet dans le contexte de l'Exercice 12.62. Un assuré a un accident durant l'année suivant la souscription d'une assurance automobile auprès de la compagnie pour laquelle vous travaillez. Quelle est la probabilité que cet individu fasse partie du groupe à risque intermédiaire ? Quelle est la probabilité que cet individu fasse partie du groupe à faible risque ?

Solution
On cherche $\Pr(G_2|A)$ et $\Pr(G_3|A)$. Par le théorème de Bayes et par l'exercice 12.62, on a

$$\Pr(G_2|A) = \frac{\Pr(G_2)\Pr(A|G_2)}{\sum_{i=1}^{3}\Pr(A|G_i)\Pr(G_i)} = \frac{0.7 \times 0.01}{0.0275} = \frac{0.007}{0.0275} = 0.2545,$$

$$\Pr(G_3|A) = \frac{\Pr(G_3)\Pr(A|G_3)}{\sum_{i=1}^{3}\Pr(A|G_i)\Pr(G_i)} = \frac{0.1 \times 0.005}{0.0275} = \frac{0.0005}{0.0275} = 0.0182.$$

Exercice 12.64. Lecture des tables statistiques
Donner les quantiles d'ordre 99%, 97.5% et 95% :

1. d'une variable normale centrée réduite,
2. d'une variable Khi-carrée à 17 degrés de liberté,
3. d'une variable de Student à 8 degrés de liberté,
4. d'une variable de Fisher(uniquement d'ordre 95%) à 5 et 7 degrés de liberté.

Solution

1. à 99% : 2.3263, à 97.5% : 1.9600, à 95% : 1.6449,
2. à 99% : 33.41, à 97.5% : 30.19, à 95% : 27.59,
3. à 99% : 2.896, à 97.5% : 2.306, à 95% : 1.860,
4. à 95% : 3.972.

Troisième partie

Matériel pour l'évaluation

Chapitre 13

Questions à choix multiples

Question 1.

On considère la série donnée par $x_i = i$ pour $i = 1, 2, 3, 4, 5$. Que vaut $\sum_{i=1}^{5} (x_i - 1)^2$?

A) 0 B) 10 C) 30 D) 90 E) 100 F) 100
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned}\sum_{i=1}^{5} (x_i - 1)^2 &= \sum_{i=1}^{5} \left(x_i^2 - 2 \times x_i + 1\right) = \sum_{i=1}^{5} \left(i^2 - 2 \times i + 1\right) \\ &= \underbrace{1 - 2 + 1}_{i=1} + \underbrace{4 - 4 + 1}_{i=2} + \underbrace{9 - 6 + 1}_{i=3} + \underbrace{16 - 8 + 1}_{i=4} + \underbrace{25 - 10 + 1}_{i=5} \\ &= 0 + 1 + 4 + 9 + 16 = 30\end{aligned}$$

Réponse : C

Question 2.
Soit la série $1, 2, 4, 2, 1$
La distribution de cette série est :
A) asymétrique à gauche et mésokurtique
B) asymétrique à gauche et leptokurtique
C) asymétrique à gauche et platykurtique
D) asymétrique à droite et mésokurtique
E) asymétrique à droite et leptokurtique
F) asymétrique à droite et platykurtique
G) symétrique et mésokurtique
H) symétrique et leptokurtique
I) symétrique et platykurtique

Solution

$$\bar{x} = \frac{1}{5}(1 + 1 + 2 + 2 + 4) = \frac{1}{5} \times 10 = 2$$

$$m_3 = \frac{1}{5} \sum_{i=1}^{5} (x_i - \bar{x})^3 = \frac{1}{5} \left(2 \times (-1)^3 + 2^3\right) = 1.2$$

$$s_x^2 = \frac{1}{5} \sum_{i=1}^{5} x_i^2 - \bar{x}^2 = \frac{1}{5} \left(2 \times 1^2 + 2 \times 2^2 + 4^2\right) - \bar{x}^2 = 1.2$$

$$s_x^3 = \left(\sqrt{s_x^2}\right)^3 = 2.9859$$

$$g_1 = \frac{m_3}{s_x^3} = \frac{1.2}{2.9859} = 0.402$$

Ainsi, la distribution est asymétrique à droite.

$$m_4 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^4 = \frac{1}{5}\left(2\times(1-2)^4 + 2\times(2-2)^4 + (4-2)^4\right) = \frac{1}{5}(2+16) = 3.6$$

$$s_x^4 = 1.2^2 = 1.44$$

$$g_2 = \frac{m_4}{s_x^4} - 3 = \frac{3.6}{1.44} - 3 = 2.5 - 3 = -0.5$$

Donc, la distribution est platykurtique.
Réponse : F

Question 3.
On a pesé 16 personnes, dont 10 hommes et 6 femmes.
On connaît :
- la moyenne du poids des femmes : $\bar{x}_A = 62$,
- la moyenne du poids des hommes : $\bar{x}_B = 78$,
- la variance du poids des femmes : $s_A^2 = 4$,
- la variance du poids des hommes : $s_B^2 = 6$.

Que vaut la variance totale $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$?

A) 5 B) 5.25 C) 10 D) 26.25 E) 65.25
F) Il est impossible de répondre à cette question G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\bar{x} = \frac{1}{n}\left(n_A \times s_A^2 + n_B \times s_B^2\right) = \frac{1}{16}(6\times 62 + 10\times 78) = 72$$

Ainsi, par le théorème de la variance totale :

$$\begin{aligned} s_x^2 &= \frac{n_A \times s_A^2 + n_B \times s_B^2}{n} + \frac{n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2}{n} \\ &= \frac{6\times 10 + 10\times 6}{16} + \frac{6\times(62-72)^2 + 10\times(78-72)^2}{16} = 62.25. \end{aligned}$$

Réponse : E

Question 4.
On considère la série statistique suivante : 3, 1, 3, 4, 4, 3, 2, 2, 4, 4.
Alors la médiane et la moyenne sont respectivement :
A) 3, 3.5 B) 3, 3 C) 3.5, 3 D) 3.5, 3.5 E) 2.5, 3.5 F) 2.5, 2.5 G) 3.5, 2.5 H) 1, 3
I) Aucune des réponses ci-dessus n'est correcte

Solution

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(1\times 1 + 2\times 2 + 3\times 3 + 4\times 4) = 3.$$

La série ordonnée est $1, 2, 2, 3, \underline{3, 3}, 4, 4, 4, 4$.
La médiane est $\frac{3+3}{2} = 3$.

Réponse : B

Question 5.
Dans une population, on sait que l'espérance de vie d'un individu (en années) suit une loi normale de moyenne 80 et de variance 4.
Quelle est le probabilité que la durée de vie d'un individu soit entre 80 et 85 ans ?
A) 0.0987 B) 0.4938 C) 0.4944 D) 0.5000 E) 0.5987 F) 0.8944 G) 0.9938 H) 1
I) Aucune des réponses ci-dessus n'est correcte

Solution

$$\Pr(X \leq 80) = \Phi\left(\frac{80-80}{2}\right) = \Phi(0) = 0.5$$

$$\Pr(X \leq 85) = \Phi\left(\frac{85-80}{2}\right) = \Phi(2.5) = 0.9938$$

$$\Pr(80 \leq X \leq 85) = \Pr(X \leq 85) - \Pr(X \leq 80) = 0.9938 - 0.4 = 0.4938$$

Réponse : C

Question 6.
Le tableau suivant contient la variable *nombre de jours d'arrêt de travail pour cause de maladie durant une année* des 100 employés d'une entreprise ainsi que les effectifs correspondants.

Nombre de jour	0	1	2	3	4
Effectif	15	15	40	25	10

Considérons les trois affirmations :
I. La moyenne est supérieure au troisième quartile,
II. La moyenne est inférieure au troisième quartile,
III. Le mode est égal à la moyenne.
Alors :
A) L'affirmation I est correcte mais les affirmations II et III sont incorrectes,
B) L'affirmation II est correcte mais les affirmations I et III sont incorrectes,
C) L'affirmation III est correcte mais les affirmations I et I sont incorrectes,
D) Les affirmations I et II sont correctes mais l'affirmation III est incorrecte,
E) Les affirmations II et III sont correctes mais l'affirmation I est incorrecte,
F) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\bar{x} = \frac{1}{10}(10 \times 0 + 1 \times 15 + 2 \times 40 + 3 \times 25 + 4 \times 10+) = 2.1$$

$$x_M = 2$$

$$x_{3/4} = \frac{1}{2}(x_{75} + x_{76}) = \frac{1}{2}(3+3) = 3$$

Donc, la moyenne est inférieure au troisième quartile mais le mode n'est pas égal à la moyenne.
L'affirmation II est correcte mais les affirmations I et III sont incorrectes.
Réponse : C

Question 7.
Toto va avoir 7 examens à la session de Janvier 2008. On sait que Toto réussit chacun de ses examens avec une probabilité de 0.6 (et donc échoue avec une probabilité de 0.4) et ce indépendamment des autres examens. Quelle est la probabilité que Toto réussisse exactement 4 de ses examens ?
A) 0.008 B) 0.058 C) 0.064 D) 0.129 E) 0.193 F) 0.290 G) 0.406 H) 0.822
I) Aucune des réponses ci-dessus n'est correcte

Solution
La variable *Nombre de réussite* suit une loi binomiale de paramètres $n = 7$ et $p = 0.6$. Donc,

$$\Pr(X = 4) = \binom{7}{4} \times 0.6^4 \times 0.4^3.$$

$$\frac{7!}{4! \times 3!} \times 0.6^4 \times 0.4^3 = \frac{5040}{24 \times 6} \times 0.6^4 \times 0.4^3 = 0.290.$$

Réponse : F

Question 8.
On considère la variable *Dernier diplôme obtenu* ayant pour domaine :
secondaire, primaire, supérieur non-universitaire, sans diplôme, universitaire.
Alors cette variable est :
A) qualitative nominale
B) qualitative ordinale
C) quantitative discrète
D) quantitative continue
E) Aucune des réponses ci-dessus n'est correcte

Solution
Réponse : B

Question 9.
Toto a un bulletin catastrophique, il décide d'ajouter 10 à toutes ses notes (les notes sont sur 20). L'écart-type de sa note sera :
A) augmenté de 10 B) multiplié par 10 C) réduit de 10 D) divisé par 10 E) multiplié par 100
F) divisé par 100 G) Aucune des réponses ci-dessus n'est correcte.

Solution
Soit, $y_i = x_i + 10, i = 1, \ldots, n$ alors $\bar{y} = \bar{x} + 10$ et

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i + 10 - \bar{x} - 10)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2.$$

La variance est donc insensible à un changement d'unité et l'écart-type l'est aussi.
Réponse : G

Question 10.
Quelle est la probabilité d'avoir au mois une fois face en jetant 5 fois une pièce de monnaie ?
A) 1/2 B) 1/4 C) 1/8 D) 1/16 E) 1/32 F) 3/4 G) 4/8 H) 15/16 I) 31/32 J) 63/64
K) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\Pr(\text{au moins une fois Face}) = 1 - \Pr(\text{pas de face}) = 1 - \left(\frac{1}{2}\right)^5 = 1 - \frac{1}{32} = \frac{31}{32} = 0.96875.$$

Réponse : I

Question 11.
Que vaut $\sum_{i=1}^{4} 2i^2$?
A) 1 B) 4 C) 10 D) 20 E) 25 F) 40 G) 50 H) 55 I) 60 J) 80
K) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\sum_{i=1}^{4} 2i^2 = 2\sum_{i=1}^{4} i^2 = 2(1+4+9+16) = 60.$$

Réponse : I

Question 12.
La taille en centimètre de 5 étudiants sont les suivantes : 165, 175, 176, 159, 170.
La médiane et la moyenne sont respectivement :
A) 170, 169 B) 170, 170 C) 169, 170 D) 176, 169 E) 176, 176 F) 159, 176
G) Aucune des réponses ci-dessus n'est correcte.

Solution
La série ordonnée est 159, 165, 170, 175, 176. La médiane est :

$$x_{1/2} = x_{(3)} = 170$$

et la moyenne

$$\bar{x} = \frac{1}{5}(159+165+170+175+176) = 169.$$

Réponse : A

Question 13.
Les notes obtenues à un examen partiel par 300 étudiants ont une distribution normale avec une moyenne de 11.269 (sur 20) et un écart-type de 3.6. Quelle est la probabilité approximative qu'une note soit supérieure à 15 ?
A) 0.01 B) 0.05 C) 0.08 D) 0.10 E) 0.15 F) 0.32 G) 0.5 H) 0.84 I) 0.95 J) 1
F) Aucune des réponses ci-dessus n'est correcte.

Solution
On note $Y \sim \mathcal{N}(11.269, \sigma = 3.6)$ et $X \sim \mathcal{N}(0, \sigma = 1)$

$$\begin{aligned}\Pr(Y > 15) &= 1 - \Pr(Y \leq 15) = 1 - \Pr\left(X \leq \frac{15 - 11.269}{3.6}\right) \\ &= 1 - \Pr(X \leq 1.036389) = 1 - \Phi(1.036389) = 1 - 0.8453441 = 0.1500104.\end{aligned}$$

Réponse : E

Question 14.
La table suivant contient la variable "nombre d'accidents par jour dans une ville pour une période de 100 jours" et les effectifs correspondant.

Accidents	0	1	2	3	4
Effectifs	55	20	10	15	0

Considérons les trois affirmations suivantes :
I. La moyenne et le mode sont égaux.
II. La moyenne et la médiane sont égales.
III. La médiane et le mode sont égaux.
A) seul I est vrai
B) seul II est vrai
C) seul III est vrai
D) I et II sont vrais mais III est faux
E) I et III sont vrais mais II est faux
F) II et III sont vrais mais I est faux
G) I, II et III sont faux

Solution
Le mode est $x_M = 0$, la médiane est

$$x_{1/2} = \frac{x_{(50)} + x_{(51)}}{2} = \frac{0+0}{2} = 0.$$

et la moyenne

$$\bar{x} = \frac{1}{100}(55 \times 0 + 20 \times 1 + 10 \times 2 + 15 \times 3 + 0 \times 4) = 0.85.$$

Réponse : C

Question 15.
Parmi les expressions ci-dessous laquelle n'est pas égale à la variance de la série $x_1, \ldots, x_i, \ldots, x_n$?

A) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ B) $\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$ C) $\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$

D) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})x_i$ E) $\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$

Solution
On a

$$\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\bar{x}^2.$$

Donc, C est n'est pas égal à la variance.
Réponse : C

Question 16.
Toto a un bulletin catastrophique, il décide de multiplier par 2 toutes ses notes (les notes sont sur 20). L'écart-type de sa note sera :

A) augmenté de 2 B) réduit de 2 C) multiplié par 2
D) augmenté de 4 E) réduit de 4 F) multiplié par 4
G) Aucune des réponses ci-dessus n'est correcte.

Solution
Comme $y_i = 2x_i, i = 1, \ldots,$
$bary = 2\bar{x}$. Donc,

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{2}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{2}(2x_i - 2\bar{x})^2 = 4\frac{1}{n}\sum_{i=1}^{2}(x_i - \bar{x})^2 = 4s_x^2.$$

La variance est multipliée par 4. Donc, l'écart-type est multiplié par 2.
Réponse : C

Question 17.
On tire au hasard 4 cartes dans un jeu de 32 cartes. Si le tirage se fait sans remise, alors la probabilité d'avoir 2 as vaut approximativement :

A) 0.013 B) 0.063 C) 0.053 D) 0.033 E) 0.023 F) 0.093 G) 0.113 H) 0.253
I) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\begin{aligned}\frac{\binom{4}{2}\binom{28}{2}}{\binom{32}{4}} &= \frac{4! \times 28! \times 4! \times 28!}{2! \times 2! \times 2! \times 26! \times 32!} = \frac{3 \times 28! \times 4! \times 28!}{26! \times 32!} \\ &= \frac{3 \times 28 \times 27 \times 24}{32 \times 31 \times 30 \times 29} = \frac{7 \times 27 \times 3}{2 \times 31 \times 5 \times 29} = \frac{567}{8990} \approx 0.06307008\end{aligned}$$

Réponse : B

Question 18.

Soient les cinq valeurs $x_1, x_2, \ldots, x_5$ dont la moyenne est 2. Que vaut $\sum_{i=1}^{5}(x_i + 2^i)$?

A) 62 B) 92 C) 32 D) 52 E) 72 F) 82 G) 102 H) 112
I) Aucune des réponses ci-dessus n'est correcte

Solution

$$\sum_{i=1}^{5}(x_i + 2^i) = \sum_{i=1}^{5} x_i + \sum_{i=1}^{5} 2^i = n\bar{x} + \sum_{i=1}^{5} 2^i = 5 \times 2 + 2 + 4 + 8 + 16 + 32 = 72$$

Réponse : E

Question 19.
Les tailles en centimètres de 6 étudiants sont les suivantes : 184, 175, 182, 186, 177, 188.
La médiane et la moyenne sont respectivement :
A) 182, 180 B) 180, 183 C) 182, 183 D) 183, 182 E) 180, 182 F) 182, 182
G) Aucune des réponses ci-dessus n'est correcte.

Solution
La série ordonnée est : 175, 177, 182, 184, 186, 188. La médiane est donc $x_{1/2} = \frac{182+184}{2} = 183$. La moyenne vaut

$$\bar{x} = \frac{1}{5}(184 + 175 + 182 + 186 + 177 + 188) = 182.$$

Réponse : D

Question 20.
Une entreprise fabrique des ampoules électriques dont la durée de vie (en heures) suit une distribution normale de moyenne 1000h et d'écart-type 200h. Quelle est la probabilité que la durée de vie d'une ampoule soit supérieure à 1234h ?
A) 0.01 B) 0.021 C) 0.121 D) 0.521 E) 0.721 F) 0.921 G) 1 H) 0.421
I) Aucune des réponses ci-dessus n'est correcte.

Solution
On pose $Y \sim \mathcal{N}(100, \sigma = 200)$ et $Y \sim \mathcal{N}(0, 1)$.

$$\begin{aligned} \Pr(Y > 1234) &= 1 - \Pr(Y \leq 1234) = 1 - \Pr\left(X \leq \frac{1234 - 1000}{200}\right) \\ &= 1 - \Pr(X \leq 1.17) = 1 - \Phi(X \leq 1.17) = 1 - 0.8789995 = 0.1210005 \end{aligned}$$

Réponse : C

Question 21.
Une série statistique de 99 distances a une moyenne de 25 mètres et une médiane de 25.5 mètres. Malheureusement, on vient de découvrir qu'une observation avait été encodée par erreur "28" à la place de la valeur correcte "30". Si on corrige cette valeur alors :
A) la moyenne reste la même mais la médiane est augmentée,
B) la moyenne et la médiane restent inchangées,
C) la médiane reste la même, mais la moyenne est augmentée,
D) la médiane et la moyenne sont augmentées,
E) il est impossible de répondre à cette question.

Solution
La médiane reste inchangée si on modifie une valeur supérieure à la médiane par une valeur supérieure à la médiane. La moyenne est augmentée.
Réponse à compléter : C

Question 22.
Parmi les expressions ci-dessous laquelle n'est pas égale à la covariance de la série $x_1, \ldots, x_i, \ldots, x_n$?

A) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ B) $\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}$ C) $\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right)$

D) $\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i\right)$ E) $\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j)$

Solution
$\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\bar{x}\bar{y}$ n'est pas égal à la covariance.
Réponse : C

Question 23.
Le propriétaire d'un café décide d'augmenter les prix de toutes ses boissons de 10%. L'écart-type des prix de ses boissons sera :
A) augmenté 21% B) diminué de 21% C) augmenté de 10%
D) diminué de 10% E) multiplié par 10 F) divisé par 10
G) Aucune des réponses ci-dessus n'est correcte.

Solution

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}, s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}}$$

$$\begin{aligned} {s_{x+10\%}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i + 0.1x_i)^2 - \left[\frac{1}{n}\sum_{i=1}^{n}(x_i + 0.1x_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n} 1.1^2 x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} 1.1x_i\right)^2 \\ &= 1.1^2\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1.1^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = 1.1^2\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] = 1.1^2 s_x^2 \end{aligned}$$

$$s_{x+10\%} = \sqrt{{s_{x+10\%}}^2} = 1.1 s_x$$

On voit donc que l'écart-type a augmenté de 10%.
Réponse : C

Question 24.
On tire au hasard 3 boules dans un sac contenant 4 boules jaunes, 7 boules rouges et 5 boules blanches.
Quelle est la probabilité approximative d'obtenir au moins 1 boule blanche si le tirage se fait sans remise ?
A) 0.125 B) 0.195 C) 0.325 D) 0.495 E) 0.505 F) 0.675 G) 0.705 H) 0.875
I) Aucune des réponses ci-dessus n'est correcte.

Solution
Pr(au moins une boule blanche)= 1 - Pr(0 boule blanche)
Pr(0 boule blanche)=$\frac{11}{16} \times \frac{10}{15} \times \frac{9}{14}$ = 0.295
Pr(au moins une boule blanche) = 1 - 0.295 = 0.705
Réponse : G

Question 25.

Soient les valeurs $x_1, x_2, \ldots, x_{25}$ dont la moyenne est 5. On sait que $\sum_{i=1}^{25}(x_i)^2 = 500$. Que vaut $\sum_{i=1}^{25}(x_i+2)^2$?

A) 600 B) 725 C) 850 D) 975 E) 1100 F) 1225
I) G) aucune des réponses ci-dessus n'est correcte

Solution

$$\sum_{i=1}^{25}(x_i+2)^2 = \sum_{i=1}^{25} x_i^2 + 4\sum_{i=1}^{25} x_i + 4\sum_{i=1}^{25} 1 = 500 + 4\times 25\times 5 + 4\times 25 = 500 + 500 + 100 = 1100$$

Réponse : E

Question 26.
Les poids (en kg) de 5 étudiants sont les $61, 68, 62, 79, 85$.
La médiane et la moyenne sont respectivement :
A) 71, 68 B) 62, 73 C) 68, 73 D) 62, 71 E) 71, 62 F) 68, 71
G) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\text{Moyenne} = \frac{1}{5}\times(61+62+68+79+85) = 71$$

$$\text{Médiane} = 68$$

Réponse : F

Question 27.
La taille des individus d'une population suit un loi normale de moyenne 170cm et d'écart-type 10cm. Quelle est la probabilité qu'un individu mesure plus de 180cm ?
A) 0.1587 B) 0.4602 C) 0.4960 D) 0.5040 E) 0.5398 F) 0.8413
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\Pr(X \geq 180) = 1 - \Pr(X \leq 180) = 1 - \Phi(\frac{180-170}{10}) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$
Réponse : A

Question 28.
On s'intéresse à la variable *nombre de personnes par ménages*. Le tableau suivant représente cette variable ainsi que les effectifs correspondants.

Nombre de personnes par ménage	1	2	3	4	5
Effectifs	2	10	40	42	6

Si on note $\bar{x}$ la moyenne, $x_{1/2}$ la médiane et x_M le mode de cette série statistique alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
Moyenne $= \dfrac{1}{100}\times(1\times 10 + 2\times 10 + 3\times 40 + 4\times 42 + 5\times 6) = 1/100\times 340 = 3.4.$
Médiane $= \dfrac{1}{2}\times(x_{50}+x_{51}) = \dfrac{1}{2}\times(3+3) = 3.$
Mode = 4
$x_{1/2} < \bar{x} < x_M$
Réponse : B

Question 29.
49 étudiants ont rendu leur examen de statistique. La moyenne des notes est de 5 alors que la médiane est de 4.5. On se rend compte qu'un "4" a été mis par erreur à la place d'un 4.5. Si on corrige cette valeur alors :
A) la moyenne reste la même mais la médiane est augmentée,
B) la moyenne et la médiane restent inchangées,
C) la médiane reste la même, mais la moyenne est augmentée,
D) la médiane et la moyenne sont augmentées,
E) il est impossible de répondre à cette question.

Solution
Réponse : C

Question 30.
Soit m_r le moment centré d'ordre $r \in \mathbb{N}$ et m'_r le moment à l'origine d'ordre $r \in \mathbb{N}$. Parmi les assertions suivantes, laquelle n'est pas égale au moment centré d'ordre trois m_3 de la série $x_1, \ldots, x_i, \ldots, x_n$?

A) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$ B) $\frac{1}{n}\sum_{i=1}^{n} x_i(x_i - \bar{x})^2 - \bar{x}m_2$ C) $m'_3 - \bar{x}^3$

D) $m'_3 - 2\bar{x}m'_2 + \bar{x}^3 - \bar{x}m_2$

Solution

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \neq \frac{1}{n}\sum_{i=1}^{n} x_i^3 - \frac{1}{n}\sum_{i=1}^{n}\bar{x}^3 = m'_3 - \bar{x}^3$$

Réponse : C

Question 31.
On admet que la longueur du pied d'un homme adulte suit une loi normale de moyenne 25 cm et de variance égale à 9. Quelle est la probabilité qu'un individu choisi au hasard ait un pied d'une longueur supérieure à 26.5 cm ?
A) 0.3085 B) 0.5675 C) 0.6915 D) 0.4325 E) 0.5000 F) 0.9025
G) Aucune des réponses ci-dessus n'est correcte

Solution
On note $Y \sim \mathcal{N}(25, 9)$ et $X \sim \mathcal{N}(0, 1)$.

$$\begin{aligned} \Pr(X \leq 26.5) &= \Phi\left(\frac{26.5 - 25}{3}\right) = \Phi(0.5) = 0.6915 \\ \Pr(X \geq 26.5) &= 1 - 0.6915 = 0.3085 \end{aligned}$$

Réponse : A

Question 32.
On considère la série statistique $x_1, \ldots, x_5$.
On sait que $\sum_{j=1}^{5} x_j = 8$ et que $\sum_{j=1}^{5} x_j^2 = 16$. Que vaut $\sum_{i=1}^{2}\sum_{j=1}^{5}(x_j - i)^2$?
A) 0 B) 8 C) 9 D) 11 E) 16 F) 25
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{5}(x_j-i)^2 &= \sum_{i=1}^{2}\sum_{j=1}^{5}(x_j^2-2ix_j+i^2)\\ &= \sum_{i=1}^{2}\sum_{j=1}^{5}x_j^2-2\sum_{i=1}^{2}i\sum_{j=1}^{5}x_j+\sum_{i=1}^{2}i^2\sum_{j=1}^{5}1\\ &= \sum_{i=1}^{2}16-2\sum_{i=1}^{2}i\times 8+5\sum_{j=1}^{5}1 = 2\times 16-2\times 3\times 8+5\times 5 = 9\end{aligned}$$

Réponse : C

Question 33.
Au sein d'un échantillon de taille n, on remarque que tous les individus ont le même revenu, noté $\bar{x}$. Que valent respectivement l'indice de Gini (G) et le Quintile Share Ratio (QSR) pour cet échantillon ?
A) G=0.5 et QSR=1 B) G=1 et QSR=$\bar{x}$ C) G=1 et QSR=$\bar{x}$
D) G=0 et QSR=0.5 E) G=0 et QSR=1 F) G=0 et QSR=$\bar{x}$
G) Aucune des réponses ci-dessus n'est correcte

Solution
Si tous les $x_i = \bar{x}$, alors $G = 0$ et $QSR = 1$.
Réponse : E

Question 34.
Combien existe-t-il d'anagrammes du mot "VARIABLE" commençant par la lettre A ?
A) 40320 B) 56 C) 5040 D) 2900 E) 5760 F) 20160
G) Aucune des réponses ci-dessus n'est correcte

Solution
Permutation sans répétition de 7! = 5040
Réponse : C

Question 35.
On considère la série temporelle suivante :

t	1	2	3	4	5
y_t	2	11	6	4	3

Que valent respectivement L^3y_4 et ∇y_2 ?
A) 11 et 2 B) 2 et 6 C) 12 et -10 D) 12 et 12 E) 11 et 9 F) 2 et 9
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned}L^3y_4 &= y_{4-3} = y_1 = 2\\ \nabla y_3 &= Iy_2 - Ly_2 = y_2 - y_1 = 11-2 = 9\end{aligned}$$

Réponse : F

Question 36.
Soit la série statistique suivante :

17	17	24	33	36	40	44	53

Si on note $\bar{x}$ la moyenne, $x_{1/2}$ la médiane et x_M le mode de cette série, alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 33$, $x_M = 17$, $x_{1/2} = 34.5$
Réponse : D

Question 37.
Dans un pays, les familles nombreuses recoivent en 2010 une allocation familiale de l'État. Dans un échantillon de familles nombreuses, le montant de l'aide alloué à la famille i est noté x_i. Le montant alloué moyen en francs dans l'échantillon est $\bar{x} = 500$ et la variance est $s_x^2 = 10000$. Pour l'année 2011, l'État a prévu d'augmenter de 100 francs toutes les allocations. La variance des x_i pour l'année 2011 sera égale à
A) 10 000 B) 20 000 C) 600 D) 60 000 E) 10100 F) 12000
G) Aucune des réponses ci-dessus n'est correcte

Solution
La variance n'est pas affectée par un changement d'origine.
Réponse : A

Question 38.
Parmi les expressions suivantes, laquelle n'est pas égale à g_1, le coefficient d'asymétrie de Fisher de la série $x_1, \ldots, x_n$? (Rappel : m_r désigne le moment centré d'ordre r)

A) $\dfrac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3}$ B) $\dfrac{\sqrt{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3}$ C) $\dfrac{m_3}{s_x^3}$

D) $\dfrac{\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3}$ E) Elles sont toutes égales à g_1 pour la série $x_1, \ldots, x_n$.

Solution
Le coefficient d'asymétrie de Fisher de la série $x_1, \ldots, x_n$ est donné par :

$$\begin{aligned} g_1 &= \underbrace{\frac{m_3}{s_x^3}}_{C} = \underbrace{\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3}}_{A} = \frac{\frac{1}{n}}{\left(\sqrt{\frac{1}{n}}\right)^3} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3} \\ &= \frac{1}{\sqrt{n}^2}\sqrt{n}^3 \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3} = \underbrace{\sqrt{n}\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3}}_{B} \neq \underbrace{\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^3}}_{D}. \end{aligned}$$

Réponse : D

Question 39.
Une chaîne de magasins d'éléctro-ménager étudie le nombre de machines à laver vendues en un jour dans ses 50 succursales. Le tableau suivant contient la variable *nombre de machines vendues* pour les 50 succursales, ainsi que les effectifs correspondants.

Nombre de machines vendues	0	1	2	3	4	5
Nombre de succursales (Effectifs)	7	16	13	8	2	4

Si on note $\bar{x}$ - la moyenne, $x_{1/2}$ - la médiane et x_M - le mode de cette série statistique alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 1.88$, $x_M = 1$, $x_{1/2} = 2$
Réponse : D

Question 40.
On pioche 4 cartes dans un jeu de 36 cartes. Quelle est la probabilité d'avoir une main contenant exactement un pique et trois coeurs (les résultats sont arrondis à la quatrième décimale) ?
A) 0.1389 B) 0.0128 C) 0.8685 D) 0.9872 E) 0.7560 F) 0.2440
G) Aucune des réponses ci-dessus n'est correcte

Solution
$$\frac{\binom{9}{3}\binom{9}{1}}{\binom{36}{4}} = 0.0128$$
Réponse : B

Question 41.
Dans un pays d'Europe, on suppose que la taille des femmes (en centimètres) suit une loi normale de moyenne 165 et d'écart-type 6. Quelle est la probabilité, pour une habitante de ce pays, de mesurer entre 159 et 168 cm ?
A) 0.6826 B) 0.5328 C) 0.8502 D) 0.4672 E) 0.5000 F) 0.1498 G) 0.6915 H) 1
I) Aucune des réponses ci-dessus n'est correcte

Solution
$\Pr(159 \leq X \leq 168) = \Pr(-1 \leq Z \leq 0.5) = \Pr(Z < 0.5) - 1 + \Pr(Z < 1) = 0.5328$
Réponse : B

Question 42.
On connaît le salaire annuel des 10 employés d'une petite entreprise. Suite à une excellente année en termes de chiffre d'affaires pour l'entreprise, la direction de celle-ci décide de doubler le salaire de chacun de ses employés. En comparaison de la variance de la distribution initiale des salaires, la variance de la nouvelle distribution sera :
A) augmentée de 2 B) réduite de 2 C) multipliée par 2
D) augmentée de 4 E) réduite de 4 F) multipliée par 4
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\text{var}(2X) = 4 \times \text{var}(X)$
Réponse : F

Question 43.
La Figure 13.1 représente les courbes de Lorenz de la distribution des revenus au sein du pays A et du pays B. On note G_A et G_B l'indice de Gini du pays A et du pays B respectivement et H_A et H_B l'indice de Hoover du pays A et du pays B respectivement. Quelle information peut-on déduire à l'aide de ce graphique ?
A) $G_A > G_B$ et $H_A > H_B$ B) $G_A < G_B$ et $H_A > H_B$
C) $G_A > G_B$ et $H_A < H_B$ D) $G_A < G_B$ et $H_A < H_B$
E) Les habitants du pays A sont en moyenne plus riches que ceux du pays B.
F) Les habitants du pays B sont en moyenne plus riches que ceux du pays A.

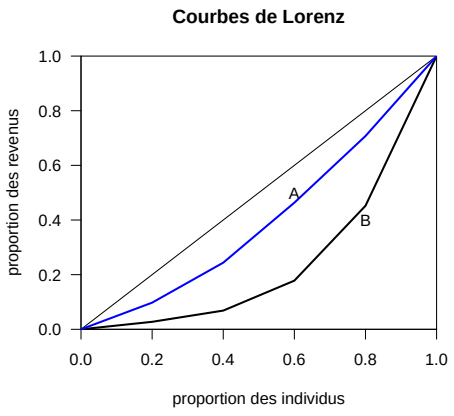


FIGURE 13.1 – Courbes de Lorenz des pays A et B

Solution
Aire entre la diagonale est la courbe (Gini) plus petite pour A. Distance verticale maximale entre la droite et la courbe (Hoover) plus petite pour A également.
Réponse : D

Question 44.
25 étudiants sont réunis dans une salle de cours. Chacun donne son âge afin de calculer la moyenne, qui est de 23.5 ans, ainsi que la médiane, qui est de 22 ans. Malheureusement, un étudiant âgé de 30 ans n'a pas dit la vérité : il a déclaré n'avoir que 26 ans. Sachant désormais qu'il a 30 ans, on recalcule la moyenne et la médiane. Que remarque-t-on ?
A) la moyenne reste la même mais la médiane augmente.
B) la moyenne et la médiane restent inchangées.
C) la médiane reste la même, mais la moyenne augmente.
D) la médiane et la moyenne sont augmentées.
E) il est impossible de répondre à cette question.

Solution
Réponse : médiane ne bouge pas (l'étudiant reste au dessus de la médiane), la moyenne augmente.
Réponse : C

Question 45.
Soit la distribution statistique suivante : $x_i = 36, 16, 4, 25, 9$. Que vaut $\sum_{i=1}^{5} |x_i - x_{1/2}|$?

A) 10 B) 70 C) 67 D) 0 E) 50 F) 100 G) 48 H) 73
I) Aucune des réponses ci-dessus n'est correcte

Solution
La médiane $x_{1/2} = 16 \Rightarrow \sum_{i=1}^{5} |x_i - x_{1/2}| = 20 + 0 + |-12| + 9 + |-7| = 48$

Réponse : G

Question 46.
Le Tableau 13.1 contient le nombre d'années d'ancienneté de 20 professeurs d'une école publique : La distance interquartile (IQ) de cette distribution vaut :
A) 8 B) 10.5 C) 30 D) 9 E) 15.5 F) 7 G) 11 H) 7.5
I) Aucune des réponses ci-dessus n'est correcte

Table 13.1 – Ancienneté de 20 professeurs

2	3	3	5	7	8	8	8	8	10
12	13	13	14	14	17	22	26	31	32

Solution
$x_{np} = x_{20\times\frac{1}{4}} = x_5 \Rightarrow x_{1/4} = \frac{x_5+x_6}{2} = 7.5$
$x_{np} = x_{20\times\frac{3}{4}} = x_{15} \Rightarrow x_{3/4} = \frac{x_{15}+x_{16}}{2} = 15.5$
$IQ = 15.5 - 7.5 = 8$
Réponse : A

Question 47.
On approxime le QI (quotient intellectuel) des habitants d'un pays par une loi normale de moyenne 102 et de variance 16. Quelle est la probabilité qu'un individu choisi au hasard ait un QI supérieur à 114 ?
A) 0.0013 6 B) 0.2266 6 C) 0.25 D) 0.75 E) 0.7734 F) 0.9987
G) Aucune des réponses ci-dessus n'est correcte

Solution
On a $X \sim \mathcal{N}(102, 13)$. Donc,

$$\begin{aligned} F(x) &= \Phi\left(\frac{x-102}{4}\right), \\ \Pr(X \leq 114) &= \Phi\left(\frac{114-102}{4}\right) = \Phi(3) = 0.9987, \\ \Pr(X \geq 114) &= 1 - 0.9987 = 0.0013 \end{aligned}$$

Réponse : A

Question 48.
Supposons que l'indice simple du café soit $I(1/0) = 114$. Sachant qu'au temps 1 j'ai dépensé 19 francs pour 2 kg de café, quelle somme aurais-je approximativement dépensé au temps 0 pour la même quantité de café ?
A) 8.33 francs B) 10.85 francs C) 16.67 francs D)18.85 francs E) 19.15 francs F) 21.67 francs
G) Aucune des réponses ci-dessus n'est correcte

Solution
On a :

$$I(1/0) = 100\frac{P_1}{P_0} = 100\frac{19}{P_0} = 114.$$

D'où

$$P_0 = \frac{100 \times 19}{114} \approx 16.67.$$

Réponse : C

Question 49.
Au sein d'un parlement, on désire créer une commission chargée d'étudier la faisabilité d'un projet. Il y a 25 députés dans ce parlement et la commission sera constituée de 4 parlementaires. Etant donné que les quatre postes sont identiques au sein de la commission, de combien de manière différente peut-on former cette commission ?
A) 303600 B) 100 C) 570024 D) 2900 E) 12650 F) 23751
G) Aucune des réponses ci-dessus n'est correcte

Solution
C'est une combinaison :

$$\binom{25}{4} = \frac{25 \times 24 \times 23 \times 21}{1 \times 2 \times 3 \times 4} = 12650.$$

Réponse : E

Question 50.
On considère la série temporelle suivante :

t	1	2	3	4	5
y_t	2	11	6	4	3

Que valent respectivement $F^2 y_3$ et $\nabla^2 y_3$?
A) 16 et 25 B) 16 et -10 C) 8 et -10 D) 8 et -15 E) 3 et -14 F) 3 et 25
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned} F^2 y_3 &= y_{3+2} = y_5 = 3 \\ \nabla^2 y_3 &= I y_3 - 2L y_3 + L^2 y_3 = y_3 - 2y_2 + y_1 = 6 - 2 \times 11 + 2 = -14 \end{aligned}$$

Réponse : E

Question 51.
Après avoir désaisonnalisé une série temporelle par la méthode additive, on a obtenu les composantes saisonnières suivantes :

S_1	=	210	S'_1	=	?
S_2	=	?	S'_2	=	?
S_3	=	-100	S'_3	=	-115
S_4	=	80	S'_4	=	65

Les composantes saisonnières S_2, S'_1 et S'_2 sont respectivement
A) -190, 195 et -205 B) -190, 195 et -145 C) -130, 195 et -145
D) -130, 225 et -205 E) -130, 225 et -145 F) -190, 225 et -205
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned} S'_3 &= S_3 - \frac{1}{4}\sum_{i=1}^{4} S_i \\ -115 &= -100 - \frac{1}{4}(210 + S_2 - 100 + 80) \\ -115 &= -147.5 - \frac{1}{4}S_2 \\ 32.5 &= -\frac{1}{4}S_2 \\ S_2 &= -130 \end{aligned}$$

$$\frac{1}{4}\sum_{i=1}^{4} S_i = \frac{1}{4}(210 - 130 - 100 + 80) = 15$$

$$S'_1 = S_1 - \frac{1}{4}\sum_{i=1}^{4} S_i = 210 - 15 = 195$$

$$S_2' = S_2 - \sum_{i=1}^{4} S_i = -130 - 15 = -145$$

Réponse : C

Question 52.
On s'intéresse à une éventuelle relation entre le sexe et le fait d'être atteint d'une maladie. Le tableau suivant reprend les effectifs des observations effectuées sur des individus.

	malade	non-malade
Hommes	33	167
Femmes	11	89

Quel pourcentage des individus malades sont des hommes ?
A) 33% B) 75% C) 19.76% D) 25% E) 16.5% F) 89%
G) Aucune des réponses ci-dessus n'est correcte

Solution
$\frac{33}{33+11} \times 100 = 75\%$ des malades sont des hommes.
Réponse : B

Question 53.
Le tableau suivant contient les notes obtenues par les étudiants à un examen ainsi que les effectifs correspondants.

Note	3	4	4.5	5	5.5	6
Effectif	1	2	6	9	5	2

Le troisième quartile est :
A) 3.75 B) 4 C) 4.5 D) 4.75 E) 5 F) 5.25 G) 5.5 H) 5.75
I) Aucune des réponses ci-dessus n'est correcte

Solution

$$x_{\frac{3}{4}} = x_{0.75} = x_{(\lceil 0.75 \times 25 \rceil)} = x_{(\lceil 18.75 \rceil)} = x_{(19)} = 5.5$$

Réponse : G

Question 54.
On a observé la note à un examen de 40 filles et 10 garçons. On a donc réparti ces observations dans deux groupes, à savoir le groupe des filles et le groupe des garçons. La moyenne et la variance des notes du groupe des filles sont respectivement 5 et 1 tandis que la moyenne et la variance des notes du groupe des garçons sont respectivement 4.5 et 0.5.
La moyenne générale et la variance totale sont respectivement :
A) 4.75 et 0.75 B) 4.75 et 0.90 C) 4.75 et 0.94 D) 4.75 et 0.96
E) 4.90 et 0.75 F) 4.90 et 0.90 G) 4.90 et 0.94 H) 4.90 et 0.96
I) Aucune des réponses ci-dessus n'est correcte

Solution
On a ici

$$\begin{array}{ll} n_A = 40 & n_B = 10 \\ \bar{x}_A = 5 & \bar{x}_B = 4.5 \\ s_A^2 = 1 & s_B^2 = 0.5 \end{array}$$

$$\bar{x} = \frac{1}{n}\left(n_A \bar{x}_A + n_B \bar{x}_B\right) = \frac{1}{50}(40 \times 5 + 10 \times 4.5) = \frac{1}{50}(200 - 45) = 4.9$$

$$s_x^2 = \frac{40 \times 1 + 10 \times 0.5}{50} + \frac{40(5-4.9)^2 + 10(4.5-4.9)^2}{50} = 0.9 + 0.04 = 0.94$$

Réponse : G

Question 55.
Le nombre d'enfants par couple d'individus d'une population suit une loi normale de moyenne 3 et d'écart-type 2. La probabilité approximative qu'un couple d'individu ait entre 1 et 3 enfants est :
A) 0.1574 B) 0.3413 C) 0.5000 D) 0.8413 E) 0.9772 F) 1.0000
G) Aucune des réponses ci-dessus n'est correcte

Solution
La fonction de répartition est ici

$$F_{3,1}(x) = \Phi\left(\frac{x-3}{2}\right)$$

$$\begin{aligned} \Pr(1 \leq X \leq 3) = F_{3,1}(3) - F_{3,1}(1) &= \Phi(0) - \Phi(-1) \\ &= \Phi(0) - (1 - \Phi(1)) = 0.5 - (1 - 0.8413) = 0.3413 \end{aligned}$$

Réponse : B

Question 56.
La variance des loyers mensuels d'un immeuble est de 100. Le propriétaire décide de les augmenter de la manière suivante : une augmentation de 10% de chacun des loyers : puis une augmentation de 100 francs suisses de chacun des loyers. Quelle est la variance des loyers après l'augmentation ?
A) 100 B) 110 C) 121 D) 210 E) 220 F) 221
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\text{var}(X) = 100$$

$$Y = aX + b = 1.1X + 100$$

$$\text{var}(Y) = \text{var}(1.1X + 100) = 1.1^2\text{var}(X) = 1.21 \times 100 = 121$$

Réponse : C

Question 57.
On jette un dé quatre fois de suite. Quelle est la probabilité (arrondie au troisième chiffre après la virgule) d'obtenir exactement deux fois un 6 ?
A) 0.019 B) 0.028 C) 0.077 D) 0.116 E) 0.333 F) 0.500
G) Aucune des réponses ci-dessus n'est correcte

Solution
On a ici

$$X \sim \mathcal{B}\left(4, \frac{1}{6}\right).$$

D'où

$$\Pr(X = 2) = \frac{4!}{2!2!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 = 0.116.$$

Réponse : D

Question 58.
On s'intéresse à une éventuelle relation entre le sexe et le fait de fumer. Le tableau suivant reprend les effectifs des observations effectuées sur des individus.

	fumeur	non-fumeur
Hommes	15	20
Femmes	35	30

Quel pourcentage des fumeurs sont des femmes ?
A) 30% B) 35% C) 50% D) 53.85% E) 65% F) 70%
G) Aucune des réponses ci-dessus n'est correcte

Solution
On remarque que l'on a 50 fumeurs (15+35). Parmi ces 50 fumeurs, 35 sont des femmes. Donc,

$$\frac{35}{50} \times 100 = 70\%$$

des fumeurs sont des femmes.
Réponse : F

Question 59.
On s'intéresse au cursus suivit par des étudiants et on considère la variable *Type de faculté* ayant les modalités suivantes : *science, lettre, droit, science économique, théologie*. Cette variable est de type :
A) qualitative nominale
B) qualitative ordinale
C) quantitative discrète
D) quantitative continue
E) Aucune des réponses ci-dessus n'est correcte

Solution
Modalité non numériques que l'on ne peut pas ordonner.
Réponse : A

Question 60.
Le tableau suivant contient l'âge d'enfants inscrits à une colonie de vacances, ainsi que les effectifs correspondants.

Âge	6	7	8	9	10
Effectif	11	6	1	1	1

L'âge moyen des enfants de cette colonie de vacances est :
A) 8.5 B) 6.75 C) 6 D) 7.5 E) 8 F) 7.75 G) 7 H) 6.25
I) Aucune des réponses ci-dessus n'est correcte

Solution
Moyenne pondérée : $\bar{x}_w = \dfrac{6 \times 11 + 7 \times 6 + 8 + 10}{11 + 6 + 1 + 1 + 1} = \dfrac{135}{20} = 6.75.$
Réponse : B

Question 61.
Dans une entreprise, chaque employé reçoit une prime de fin d'année qui dépend de la qualité de son travail. En raison de la crise, la direction de l'entreprise a décidé de diviser toutes les primes par trois par rapport à la valeur prévue initialement. L'écart-type de ces primes, quant à lui, sera :
A) augmenté de 3 B) réduit de 3 C) divisé par 3
D) augmenté de 9 E) réduit de 9 F) divisé par 9
G) Aucune des réponses ci-dessus n'est correcte.

Solution

Divisé par 3. En effet, $\sqrt{\text{var}\left(\frac{X}{3}\right)} = \sqrt{\frac{\text{var}(X)}{9}} = \frac{\sqrt{\text{var}(X)}}{3}$.

Réponse : C

Question 62.
Une entreprise fabrique des ampoules électriques dont la durée de vie (en heures) suit une distribution normale de moyenne 1000h et d'écart-type 100h. Quelle est la probabilité que la durée de vie d'une ampoule soit supérieure à 1150h ?

A) 0.0668 B) 0.5596 C) 0.1597 D) 1 E) 0.0228 F) 0.9332 G) 0.4404 H) 0.9772
I) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\Pr(X > 1150) = \Pr\left(Z > \frac{1150 - 1000}{100}\right) = 1 - \Pr(Z < 1.5) = 1 - 0.9332 = 0.0668$$

Réponse : A

Question 63.
On tire au hasard 5 boules dans une urne contenant 6 boules rouges et 4 boules blanches. Si le tirage se fait sans remise, alors la probabilité (arrondie à trois décimale) d'avoir tiré 4 boules rouges et 1 boule blanche est :

A) 0.500 B) 0.667 C) 0.250 D) 0.333 E) 0.750 F) 0.238 G) 0.017 H) 0.200
I) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\Pr(6R, 4B) = \frac{\binom{6}{4}\binom{4}{1}}{\binom{10}{5}} = 0.238$$

Réponse : F

Question 64.
Soit la série statistique suivante :

17	17	24	33	36	40	44	52

Si on note $\bar{x}$ la moyenne, $x_{1/2}$ la médiane et x_M le mode de cette série, alors :

A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 33$, $x_M = 17$, $x_{1/2} = 34.5$
Réponse : D

Question 65.

On considère la série statistique suivante : $x_i = 3, 5, 8, 14, 20$. Que vaut $\sum_{i=1}^{5} \left|x_i - x_{1/2}\right|$?

A) 8 B) 10 C) 50 D) 5.2 E) 26 F) 29
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$|3-8|+|5-8|+|8-8|+|14-8|+|20-8| = 26$
Réponse : E

Question 66.
Un concessionnaire automobile étudie le nombre de véhicules vendus en un jour dans ses 60 succursales. Le tableau suivant contient la variable *nombre de véhicules vendus* pour les 60 succursales, ainsi que les effectifs correspondants.

Nombre de véhicules vendus	0	1	2	3	4	5
Nombre de succursales (Effectifs)	12	8	14	16	6	4

Si l'on note $\bar{x}$ la moyenne de cette série statistique, $x_{1/2}$ sa médiane et x_M son mode alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 2.133$, $x_M = 3$, $x_{1/2} = 2$
Réponse : B

Question 67.
Soit la série statistique suivante :

17	17	24	33	36	40	44	53

Si on note $\bar{x}$ la moyenne, $x_{1/2}$ la médiane et x_M le mode de cette série, alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 33$, $x_M = 17$, $x_{1/2} = 34.5$
Réponse : D

Question 68.
Un concessionnaire automobile étudie le nombre de véhicules vendus en un jour dans ses 40 succursales. Le tableau suivant contient la variable *nombre de véhicules vendus* pour les 40 succursales, ainsi que les effectifs correspondants.

Nombre de véhicules vendus	0	1	2	3	4	5
Nombre de succursales (Effectifs)	9	5	11	8	6	1

Si l'on note $\bar{x}$ la moyenne de cette série statistique, $x_{1/2}$ sa médiane et x_M son mode alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 2$, $x_M = 2$, $x_{1/2} = 2$
Réponse : F

Question 69.
On suppose que le poids (en kg) d'un aigle suit une loi Normale de moyenne 4 et de variance 1. Quelle est la probabilité que le poids d'un aigle pris au hasard soit inférieur à 4.5 kg ?
A) 0.0668 B) 0.6915 C) 0.4944 D) 0.5000 E) 0.5199 F) 0.3085 G) 0.9332 H) 1
I) Aucune des réponses ci-dessus n'est correcte

Solution

$$\Pr(X \leq 4.5) = \Phi\left(\frac{4.5-4}{1}\right) = \Phi(0.5) = 0.6915$$

Réponse : B

Question 70.
Soit une série statistique représentant la température relevée (en degrés Celsius) au même moment dans 10 villes de Suisse. On a calculé la moyenne et la variance de cette série et obtenu les résultats suivants : $\bar{x} = 20$ et $s_x^2 = 100$. On souhaite maintenant construire la série y_i à l'aide de la relation $y_i = 1.8x_i + 32$ et ainsi convertir les degrés Celsius en degrés Fahrenheit. Combien valent respectivement $\bar{y}$ et s_y^2 ?

A) 68 et 324 B) 36 et 100 C) 33.8 et 324 D) 33.8 et 100 E) 36 et 180 F) 68 et 180
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\bar{y} = 1.8\bar{x} + 32 = 1.8 \times 20 + 32 = 68$$
$$s_y^2 = 1.8^2 s_y^2 = 3.24 \times 100 = 324$$

Réponse : A

Question 71.
On considère la série statistique $x_1, \ldots, x_5$. On sait que $\bar{x} = 8$ et que $x_{1/2} = 4$. Que vaut $\sum_{i=1}^{5}(x_i - x_{1/2})$?

A) 60 B) 8 C) 0 D) 20 E) 4 F) -12
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\sum_{i=1}^{5}(x_i - x_{1/2}) = \sum_{i=1}^{5} x_i - \sum_{i=1}^{5} x_{1/2} = n\bar{x} - nx_{1/2} = 5 \times 8 - 5 \times 4 = 20$$

Réponse : D

Question 72.
On tire au hasard 5 boules dans une urne contenant 6 boules rouges et 4 boules blanches. Si le tirage se fait sans remise, alors la probabilité (arrondie à trois décimale) de tirer 4 boules rouges et 1 boule blanche est :

A) 0.500 B) 0.667 C) 0.250 D) 0.333 E) 0.750 F) 0.238 G) 0.017 H) 0.200
I) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\Pr(6R, 4B) = \frac{\binom{6}{4}\binom{4}{1}}{\binom{10}{5}} = 0.238$$

Réponse : F

Question 73.
Soit une série statistique représentant la température relevée (en degrés Celsius) au même moment dans 10 villes de Suisse. On a calculé la moyenne et la variance de cette série et obtenu les résultats suivants : $\bar{x} = 10$ et $s_x^2 = 25$. On souhaite maintenant construire la série y_i à l'aide de la relation $y_i = 1.8x_i + 32$ et ainsi convertir les degrés Celsius en degrés Fahrenheit. Combien valent respectivement $\bar{y}$ et s_y^2 ?

A) 42 et 81 B) 50 et 45 C) 18 et 25 D) 18 et 45 E) 50 et 81 F) 42 et 25
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\bar{y} = 1.8\bar{x} + 32 = 1.8 \times 10 + 32 = 50$$
$$s_y^2 = 1.8^2 s_y^2 = 3.24 \times 25 = 81$$

Réponse : E

Question 74.
Un concessionnaire automobile étudie le nombre de véhicules vendus en un jour dans ses 40 succursales. Le tableau suivant contient la variable *nombre de véhicules vendus* pour les 40 succursales, ainsi que les effectifs correspondants.

Nombre de véhicules vendus	0	1	2	3	4	5
Nombre de succursales (Effectifs)	9	5	11	8	6	1

Si l'on note $\bar{x}$ la moyenne de cette série statistique, $x_{1/2}$ sa médiane et x_M son mode alors :
A) $\bar{x} < x_{1/2} < x_M$ B) $x_{1/2} < \bar{x} < x_M$ C) $\bar{x} < x_M < x_{1/2}$
D) $x_M < \bar{x} < x_{1/2}$ E) $x_M < x_{1/2} < \bar{x}$ F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\bar{x} = 2$, $x_M = 2$, $x_{1/2} = 2$
Réponse : F

Question 75.
On considère la série temporelle suivante :

t	1	2	3	4	5
y_t	5	8	3	7	11

Que valent respectivement $L^2 y_4$ et ∇y_5 ?
A) 4 et -3 B) 8 et 3 C) 4 et -4 D) 8 et 4 E) 5 et 4 F) 5 et 2
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned} L^2 y_4 &= y_{4-2} = y_2 = 8 \\ \nabla y_5 &= I y_5 - L y_5 = y_5 - y_4 \\ &= 11 - 7 = 4 \end{aligned}$$

Réponse : D

Question 76.
Dans un échantillon de 2000 individus, on s'intéresse à deux variables qualitatives, la catégorie socioprofessionnelle (8 modalités) et la catégorie d'age (5 modalités). On a calculé $\chi^2_{obs} = 1620$. Que vaut le V de Cramer représentant le degré d'association de ces deux variables ?
A) 0.45 B) 0.17 C) 0.81 D) 0.402 E) 0.9 F) 0.2025
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$V = \sqrt{\frac{1620}{2000 \times 4}} = 0.45$$

Réponse : A

Question 77.
Pour entrer dans un immeuble, il faut composer un code de 4 lettres sur un clavier qui contient les 9 lettres suivantes : A, B, C, D, E, F, G, H et I. Chaque lettre peut apparaître au maximum une fois par code. Combien existe t-il de différents codes possibles ?
A) 4000 B) 3024 C) 36 D) 126 E) 6561 F) 3600
G) Aucune des réponses ci-dessus n'est correcte

Solution
$A_4^9 = \frac{9!}{(9-4)!} = 3024$
Réponse : B

Question 78.
On considère la série temporelle suivante :

t	1	2	3	4	5
y_t	4	8	5	2	6

Que valent respectivement $F^3 y_2$ et $\nabla^2 y_4$?
A) 2 et 20 B) 6 et 20 C) 6 et 0 D) 2 et 0 E) 8 et 7 F) 5 et 7
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned} F^3 y_2 &= y_{2+3} = y_5 = 6 \\ \nabla^2 y_4 &= I y_4 - 2L y_4 + L^2 y_4 = y_4 - 2y_3 + y_2 = 2 - 2 \times 5 + 8 = 0 \end{aligned}$$

Réponse : C

Question 79.
Dans la série $x_1, \ldots, x_{30}$ dont la moyenne est égale à 2, on sait que $\sum_{i=1}^{30} x_i^2 = 280$. Combien vaut $\sum_{i=1}^{30} (x_i - 1)^2$?
A) 250 B) 190 C) 279 D) 560 E) 220 F) 161
G) Aucune des réponses ci-dessus n'est correcte

Solution
$$\sum_{i=1}^{30} (x_i - 1)^2 = \sum_{i=1}^{30} (x_i)^2 - 2 \sum_{i=1}^{30} x_i + \sum_{i=1}^{30} 1 = 280 - 2 \times 60 + 30 = 190$$
Réponse : B

Question 80.
Dans le tableau suivant, quelle est la relation entre la moyenne, la médiane et le mode du nombre d'enfants par couple ?

Nombre d'enfants par couple	0	1	2	3
Effectifs	4	5	10	2

A) $\bar{x} <$ mode $= x_{1/2}$ B) mode $< \bar{x} < x_{1/2}$ C) mode $= \bar{x} < x_{1/2}$
D) $x_{1/2} <$ mode $< \bar{x}$ E) mode $< x_{1/2} < \bar{x}$ F) $x_{1/2} < \bar{x} <$ mode
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$$\bar{x} = \frac{0 \times 4 + 5 \times 1 + 10 \times 2 + 2 \times 3}{4 + 5 + 10 + 2} = 1.48$$
mode = 2

$x_{1/2} = 2$
Donc, $\bar{x} <$ mode $= x_{1/2}$
Réponse : A

Question 81.
Soit une variable x, dont la moyenne est égale à 20 et l'écart-type 2. On définit la variable y par $y = 2x + 2$. La moyenne et la variance de y valent respectivement :
A) 40 et 16 B) 40 et 8 C) 42 et 16 D) 40 et 4 E) 42 et 8 F) 42 et 4
G) Aucune des réponses ci-dessus n'est correcte

Solution
$\bar{y} = \sum_{i=1}^{n} (2x_i + 2) = \frac{1}{n} 2 \sum_{i=1}^{n} (x_i) + \frac{1}{n} \sum_{i=1}^{n} 2 = 2\bar{x} + 2 = 20 + 2 = 42$
$\text{var}(y) = \frac{1}{n} \sum_{i=1}^{n} \left(y - \bar{y}\right)^2 = \frac{1}{n} \sum_{i=1}^{n} (2x_i + 2 - 2\bar{x} - 2)^2 = \frac{1}{n} \sum_{i=1}^{n} (2x_i - 2\bar{x})^2 = 2^2 \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 4 \times \text{var}(x) = 4 \times 4 = 16$
Réponse : C

Question 82.
Combien y a-t-il d'anagrammes du mot VOITURE qui finissent par T ? (un anagramme est une construction fondée sur une figure de style qui inverse ou permute les lettres d'un mot. Par exemple, un anagramme du mot VOITURE est EVOITUR)
A) 720 B) 5040 C) 7 D) 528 E) 360 F) 1080
G) Aucune des réponses ci-dessus n'est correcte.

Solution
Comme toutes les lettres sont différentes, c'est le nombre de permutations de 6 lettres. Donc, $6! = 720$
Réponse : A

Question 83.
Dans le tableau suivant, l'espérance mathématique de x vaut :

x	0	1	2	3
$\Pr(x)$	0.25	0.40	0.20	0.15

A) 1.5 B) 2 C) 1.75 D) 2.20 E) 1.25 F) 0.31
G) Aucune des réponses ci-dessus n'est correcte.

Solution
$\text{E}(x) = \sum_{x=0}^{3} x \times \Pr(x) = 0 \times 0.25 + 1 \times 0.40 + 2 \times 0.20 + 3 \times 0.15 = 0 + .0.40 + 0.40 + 0.45 = 1.25.$
Réponse : E

Question 84.
Une entreprise fabrique des disques dont le diamètre suit une distribution normale de moyenne 2 cm et variance 0.2 cm^2. Quelle est la probabilité (arrondie à 4 décimales) que le diamètre d'un disque soit entre 1.95 et 2.05 ?
N.B. : Des valeurs de la fonction de répartition d'une loi normale centrée réduite sont fournies dans une table à la fin de ce document.
A) 0.0438 B) 0.5438 C) 0.5040 D) 0.0722 E) 0.0548 F) 0.0876
G) Aucune des réponses ci-dessus n'est correcte.

Solution

$\Pr(1.95 < x < 2.05) = \Pr((1.95-2)/0.45 < z < (2.05-2)/0.45) = \Pr(-0.11 < z < 0.11) = 2 \times 0.0438 = 0.0876.$

Réponse : F

Question 85.

On considère la série $x_i = 4, 5, 8, 10, 12, 15$. Combien vaut $\sum_{i=1}^{6} |x_i - x_{1/2}|$?

A) 20 B)22 C) 16 D) 50 E)42 F) 0
G) Aucune des réponses ci-dessus n'est correcte

Solution
$\sum_{i=1}^{6} |\left(x_i - x_{1/2}\right)| = |4-9| + |5-9| + |8-9| + |10-9| + |12-9| + |15-9| = 20$
Réponse : A

Question 86.
Dans un sondage sur le revenu, on trouve que le QSR vaut 6. Cela veut dire que :
A) Il y a 6 fois plus de gens riches que pauvres.
B) Le revenu le plus grand est 6 fois le revenu le plus petit.
C) Le revenu moyen de 20% des plus riches est 6 fois plus grand que celui de 20% des plus pauvres.
D) La plupart de la population est riche.
E) Il y a 6 fois plus de gens pauvres que riches.
F) La plupart de la population est pauvre.
G) Aucune des réponses ci-dessus n'est correcte.

Solution
Réponse : C

Question 87.
Soit X une variable aléatoire qui suit une loi exponentielle de paramètre 2. Laquelle des assertions suivantes sont vraies ?
A) la médiane de X vaut 2 et $\Pr(1 \leq X \leq 2) = e^{-2}(1 - e^{-2})$
B) la médiane de X vaut $\log(2)/2$ et $\Pr(1 \leq X \leq 2) = e^{-2}(1 - e^{-2})$
C) la médiane de X vaut $2\log(2)$ et $\Pr(1 \leq X \leq 2) = e^{-2}$
D) la médiane de X vaut $\log(2)/2$ et $\Pr(1 \leq X \leq 2) = e^{-2}$
E) la médiane de X vaut 2 et $\Pr(1 \leq X \leq 2) = 2e^{-2}$
F) la médiane de X vaut $2\log(2)$ et $\Pr(1 \leq X \leq 2) = 2e^{-2}$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
La fonction de répartion vaut pour $x \geq 0$.

$$F(x) = 1 - \exp(-\lambda x) = 1 - \exp(-2x).$$

La médiane est telle que

$$F(x_{1/2}) = 1 - \exp(-\lambda x_{1/2}) = \frac{1}{2}$$

ou

$$\exp(-\lambda x_{1/2}) = \frac{1}{2}$$

ou encore

$$\lambda x_{1/2} = \log(2).$$

Donc,

$$x_{1/2} = \frac{1}{2}\log(2).$$

De plus,

$$\Pr(1 \leq X \leq 2) = F(2) - F(1) = 1 - \exp(-2 \times 2) - 1 + \exp(-2 \times 1) = \exp(-2) - \exp(-4) = e^{-2}(1 - e^{-2}).$$

Réponse : B

Question 88.
À partir des courbes de Lorenz en Figure 13.2. Parmi les assertions suivantes, laquelle est vraie ?

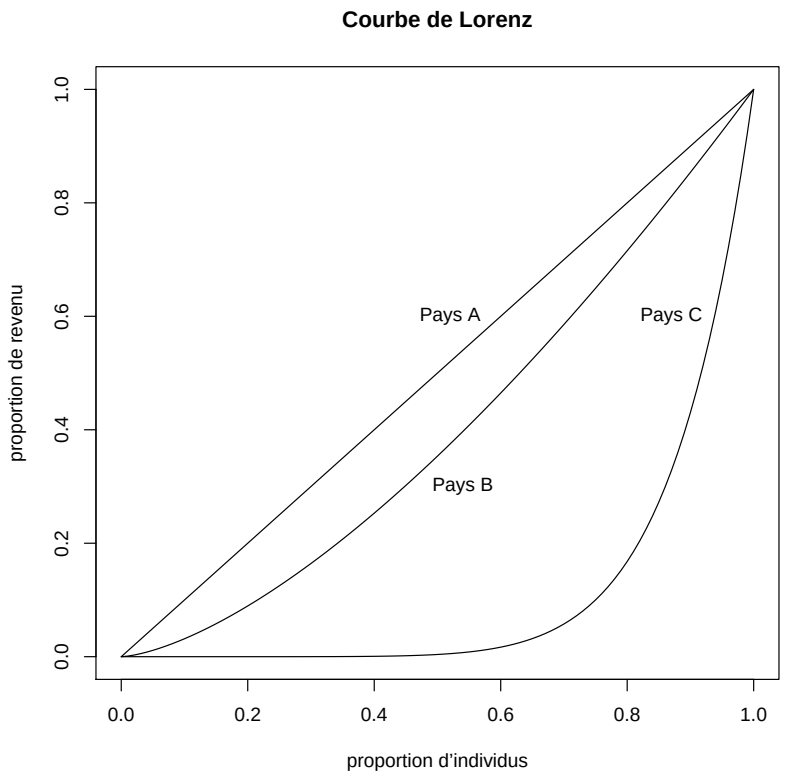


FIGURE 13.2 – Courbe de Lorenz des revenus

A) $H_A < H_B < H_C$ avec le pays C où les revenus sont le plus inégalement répartis.
B) $H_A < H_B < H_C$ avec le pays A qui est le plus riche en termes de revenu moyen.
C) $G_A < G_B < G_C$ avec le pays A où les revenus sont le plus inégalement répartis.
D) $G_A < G_B < G_C$ et $H_C < H_B < H_A$.
E) $G_C < G_B < G_A$ avec le pays A qui est le plus riche en termes de revenu moyen.
F) Aucune des réponses données n'est correcte.

Solution
Réponse : A

Question 89.
Une entreprise fabrique des pièces mécaniques qui devraient toutes mesurer 12 cm de haut. On note X la variable aléatoire prenant pour valeur la hauteur d'une boîte de conserve. On suppose que celle-ci est une normale de moyenne 12 cm et d'écart-type 0.1 cm. Pour rentabiliser la production un gestionnaire considère qu'il ne faudrait jeter que 1% de la production. Quel serait l'écartement à la valeur de 12 cm qui serait toléré ?
A) 0.995 B) 1.672 C) 0.258 D) 0.01 E) 0.516
F) Aucune des réponses ci-dessus n'est correcte

Solution
Soit y un écartement à la valeur de 12cm. On cherche une condition sur y pour que la probabilité qu'une pièce choisie au hasard dans la production s'écarte de y vaille moins de 1%. On cherche donc une condition

sur y pour avoir :

$$\begin{aligned}
\Pr(X \leq 12 - y) + \Pr(X \geq 12 + y) &< 0.01 \\
\Phi\left(\frac{12 - y - 12}{0.1}\right) + 1 - \Phi\left(\frac{12 + y - 12}{0.1}\right) &< 0.01 \\
\Phi\left(\frac{-y}{0.1}\right) + 1 - \Phi\left(\frac{y}{0.1}\right) &< 0.01 \\
1 - \Phi\left(10y\right) + 1 - \Phi\left(10y\right) &< 0.01 \\
2 - 2\Phi(10y) &< 0.01 \\
\Phi(10y) &> 0.995
\end{aligned}$$

Il suffit d'imposer la condition $10y > 2.58$, c'est-à-dire $y > 0.258$, pour que moins de 1% des pièces de la production soient défectueuses, soit, en pourcentage, une tolérance de 2.15%.
Réponse : C

Question 90.
On dispose d'un tableau statistique reportant la taille (en cm) de 3552 étudiants d'une Université parisienne. Pour analyser ces données on souhaite les diviser dans des classes. Pour cela, on a utilisé la règle de Sturge.
De plus on sait que la première observation est égale à 145 cm et la dernière à 201 cm (rangées en ordre croissant).
Quelle est la longueur de l'intervalle ? (Les calculs intermédiaires ont été arrondis au troisième chiffre après la virgule)
A) 12.717 B) 4.404 C) 56 D) 5 E) 3.333
F) Aucune des réponses ci-dessus n'est correcte

Solution
$J = 1 + (3.3 \log_{10}(3552)) = 12.717$.
Longueur de l'intervalle : $\dfrac{201 - 145}{12.717} = 4.404$.
Réponse : B

Question 91.
On connait les ventes des trois premiers semestres sur les 16 semestres consécutifs de 1991 à 1994 de cuisines d'un magasin de meubles à Aubonne : $Y_1 = 115; Y_2 = 80; Y_3 = 120$. On pense que la série est de la forme

$$Y_t = T_t + S_t + e_t,$$

où T_t est la tendance, S_t est la composante saisonnière, qui ne dépend que du numéro de trimestre dans l'année et telle que la somme des S_t sur quatre trimestres consécutifs vaut 0 et les e_t sont des résidus.
De plus, on a calculé les composantes saisonnières sur les 16 semestres consécutifs :

$$S_1 = -20.215, S_2 = 30.750, S_3 = -30.750 \text{ et } S_4 = 20.215.$$

Quelle quantité va-t-on obtenir si on désaisonnalise l'observation Y_2 ?
A) 110.215 B) 49.25 C) 59.785 D) 105 E) 0
F) Aucune des réponses ci-dessus n'est correcte

Solution
$\tilde{Y}_2 = 80 - 30.750 = 49.25$
Réponse : B

Question 92.
Les citoyens américains doivent voter pour les élections présidentielles. Ils ont le choix entre élire un président démocrate ou élire un président républicain. Une télévision locale fait un sondage dans les rues de la capitale pour déterminer quel est le président favori. On définit les évènements suivants :

1. A = "La télévision interroge une femme"

2. B = "La personne interrogée préfère le candidat républicain"
3. C = "La personne interrogée est un homme et préfère le candidat démocrate"

Comment peut-on définir l'événement $\overline{C}$?

A) $A \cup B$ B) $A \cap B$ C) $\overline{A} \cap \overline{B}$ D) $\overline{A \cup B}$ E) $\Pr(A) + \Pr(B) - \Pr(C)$
F) Aucune des réponses proposées n'est correcte

Solution
Réponse : A

Question 93.
On sait que $\mathrm{E}(X) = 0.5$ et que $\mathrm{var}(X) = 0.2$. On définit $W = 2 + 3X$. Quelle est l'espérance et la variance de W ?

A) 1.25 et 5.5 B) 0.5 et 1.8 C) 3.5 et 1.8
D) 3.5 et 0.6 E) 1.5 et 0.6 F) Aucune des réponses proposées

Solution

$$\mathrm{E}(W) = \mathrm{E}(2 + 3X) = 2 + 3\mathrm{E}(X) = 2 + 3 \times 0.5 = 3.5,$$
$$\mathrm{var}(W) = \mathrm{var}(2 + 3X) = 3^2\mathrm{var}(X) = 9 \times 0.2 = 1.8.$$

Réponse : C

Question 94.
Parmi les expressions suivantes, laquelle correspond à : $\frac{1}{n}\sum_{i=1}^{n}(2x_i - \bar{y})^2$.

A) $\frac{1}{n}\sum_{i=1}^{n} 4x_i^2 - 2x_i\bar{y} + \bar{y}^2$ B) $\frac{1}{n}\sum_{i=1}^{n} 4x_i^2 - \sum_{i=1}^{n} 2x_i\bar{y} + \sum_{i=1}^{n} \bar{y}^2$

C) $\frac{1}{n}\sum_{i=1}^{n} 4x_i^2 - 2x_i\bar{y} + n\bar{y}^2$ D) $\frac{4}{n}\sum_{i=1}^{n} x_i^2 - 4 \times \bar{x} \times \bar{y} + \bar{y}^2$

E) $\frac{1}{n}\sum_{i=1}^{n} 4x_i^2 - \frac{1}{n}\sum_{i=1}^{n} \bar{y}^2$ F) Aucune des réponses proposées

Solution

Réponse : D

Question 95.
On a récolté les données suivantes sur les poids (en kg) de deux races de chien :

1. Lévrier : 27, 30, 32, 36, 40
2. Berger Allemand : 31, 33, 34, 39

On connait $s_a^2 = 20.8$ et $s_b^2 = 7.936$ (les calculs intermédiaires ont été arrondis au troisième chiffre après la virgule).
Quelles est la variance totale de cet échantillon ?

A) 28.736 B) 15.083 C) 21.682 D) 14.975 E) 15.469
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$\bar{x}_{\text{lévrier}} = \frac{165}{5} = 33,$$

$$\bar{x}_{\text{berger}} = \frac{137}{4} = 34.250,$$
$$\bar{x}_{tot} = \frac{33 \times 5 + 34.25 \times 4}{9} = 33.556,$$
$$s^2_{tot} = \frac{5 \times 20.8 + 4 \times 7.936}{9} + \frac{5 \times (33 - 33.556)^2 + 4 \times (34.25 - 33.556)^2}{9} = 15.261.$$

Réponse : E

Question 96.
Jérôme rentre à la maison après une longue journée de travail. Il veut se cuisiner un petit plat délicieux. Il ouvre donc le frigo et y trouve 4 légumes, 2 sortes de féculant et 3 produits laitiers. Il choisit au hasard 1 légume, 1 féculant et 1 produit laitier. Combien de mets différents peut-il cuisiner ?
A) 9 B) 50 C) 3 D) 32 E) 24
G) Aucune des réponses ci-dessus n'est correcte

Solution
$4 \times 2 \times 3 = 24$ Réponse : E

Question 97.
Dans un immeuble de 3 étages sans compter le rez-de-chaussée (c.-à-d. niveau 1, 2 et 3), Mr Dubois et Mme Jeanneret s'apprêtent à monter dans l'ascenseur. Quelle est la probabilité que les deux individus descendent à des étages différents ?
A) $0.3\overline{3}$ B) $0.1\overline{6}$ C) $0.66\overline{6}$ D) 1.5 E) 0.75
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$\Omega = \{(1,1),(1,2),(2,1),(1,3),(3,1),(2,2),(2,3),(3,2),(3,3)\}$$

Soit A l'événement "descendre à des étages différents" :

$$\Pr(A) = \Pr(\{(1,2),(2,1),(1,3),(3,1),(2,3),(3,2)\}) = \frac{6}{9} = 0.66\overline{6}.$$

Réponse : C

Question 98.
En reprenant le même énoncé qu'à la question 2), quelle est la probabilité qu'au moins une personne descende au 1er étage ?
A) $0.1\overline{6}$ B) $0.55\overline{6}$ C) $0.6\overline{6}$ D) 2 E) 0.5
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$\Omega = \{(1,1),(1,2),(2,1),(1,3),(3,1),(2,2),(2,3),(3,2),(3,3)\}$$

Soit B l'événement "personne descend au 1er étage" et C "au moins une personne descend au 1er étage".

$$\Pr(C) = 1 - \Pr(B) = 1 - \Pr(\{(2,2),(2,3),(3,2),(3,3)\}) = 1 - \frac{4}{9} = 0.55\overline{6}.$$

Réponse : B

Question 99.
Le Cafignon décide d'écrire un article sur les étudiants étrangers qui séjournent à Neuchâtel. Il sait que 20% de

la population étudiante de l'Université de Neuchâtel n'a pas la nationalité suisse. Le journal a interviewé au hasard 30 étudiants à la Cafétéria. Quelle est la probabilité qu'au moins 2 soient étrangers ?
A)6 B)0.989 C) 0.2 D) $0.16\overline{6}$ E) 0.35
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$X \sim \mathcal{B}(30; 0.2)$$
$$\Pr(X > 2) = 1 - \Pr(X = 0) - \Pr(X = 1) = 1 - 0.8^{30} - 30 \times 0.2 \times 0.8^{29} = 0.989$$

Réponse : B

Question 100.
En reprenant l'énoncé de la question 4), quel est l'espérance du nombre de suisses selectionnés ?
A) 24 B) 6 C) 50 D) 22 E) 15
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$1 - \Pr(A) = 1 - 0.2 = 0.8$$
$$Y \sim \mathcal{B}(30; 0.8)$$
$$\mathrm{E}(Y) = np = 30 \times 0.8 = 24$$

Réponse : A

Question 101.
On sait que $\mathrm{E}(X) = 0.25$ et que $\mathrm{var}(X) = 0.3$. On définit $W = 4 + 5X$. Quelle est l'espérance et la variance de W ?
A) 1.25 et 5.5 B) 5.25 et 11.5 C) 5.25 et 7.5 D) 1.25 et 7.5 E) 5.25 et 5.5
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$\mathrm{E}(W) = \mathrm{E}(4 + 5X) = 4 + 5\mathrm{E}(X) = 4 + 5 \times 0.25 = 5.25.$$
$$\mathrm{var}(W) = \mathrm{var}(4 + 5X) = 5^2\mathrm{var}(X) = 25 \times 0.3 = 7.5.$$

Réponse : C

Question 102.
Quelle est l'expression simplifiée de la quantité suivante :

$$2x_{11} + 2x_{12} + 2x_{21} + 2x_{22} + 2x_{31} + 2x_{32} + 2x_{41} + 2x_{42} + 4 \times 2 \times 3.$$

A) $\sum_{i=1}^{4}\sum_{j=1}^{2} 2x_{ij} + 3$ B) $\left(2\sum_{i=1}^{4}\sum_{j=1}^{2} x_{ij}\right) + 12$ C) $2\sum_{i=1}^{n} x_{ii} + 12$
D) $\sum_{i=1}^{4}\sum_{j=1}^{2}(2x_{ij} + 3)$ E) $\sum_{i=1}^{4}(2x_{ii} + 3)$ F) Aucune des réponses proposées

Solution
Réponse : D

Question 103.
On sait que le prix d'un appartement dans la ville de Neuchâtel suit une distribution normale avec une moyenne de 1 200 francs suisses et un écart-type de 200 francs suisses. Quelle est la probabilité qu'un appartement coûte plus de 1 500 CHF ?
A) 0.7734 B)0.8531 C) 0.2266 D) 0.5557 E)0.1469
F) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned}\Pr(Z > 1500) &= \Pr\left(X > \frac{1500-1200}{200}\right) = \Pr\left(X > \frac{15-12}{2}\right)\\ &= \Pr(X > 1.5) = 1 - \Pr(X \leq 1.5) = 1 - \Phi(1.5) = 1 - 0.8531 = 0.1469.\end{aligned}$$

Réponse : E

Question 104.
Soit X une variable aléatoire qui suit une loi de Poisson de paramètre 3. Laquelle des assertions suivantes est vraie ?
A) L'espérance de X vaut 9 et $\Pr(X = 2) = 0.180$
B) La variance de X vaut 4 et $\Pr(X = 2) = 0.224$
C) L'espérance de X vaut $\frac{1}{3}$ et $\Pr(X = 2) = 0.224$
D) La variance de X vaut $\frac{1}{9}$ et $\Pr(X = 2) = 0.224$
E) La variance de X vaut 3 et $\Pr(X = 2) = 0.180$
F) Aucune des réponses ci-dessus n'est correcte

Solution
Si $X \sim \mathcal{P}(3)$ alors $\mathrm{E}(X) = 3$ et $\mathrm{var}(X) = 3$ et $\Pr(X = 2) = \dfrac{e^{\lambda} \times 3^2}{2!} = 0.224$.
Réponse : F

Question 105.
On a un échantillon de 20 personnes, comprenant 12 hommes et 8 femmes. On a mesuré la taille des individus et on a noté les résultats suivants :
— la taille moyenne (en cm) des femmes : $\bar{x}_F = 165$,
— la taille moyenne des hommes : $\bar{x}_H = 175$,
— la variance de la taille des femmes : $s_F^2 = 9$,
— la variance de la taille des hommes : $s_H^2 = 16$.
Que vaut la variance totale de la taille des individus dans cet échantillon ?
A) 37.2 B) 25 C) 5 D) 38.2 E) 12.5 F) 7.8
G) Il est impossible de répondre à cette question.
H) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\overline{x} = \frac{1}{n}\left(n_F \times \bar{x}_F + n_H \times \bar{x}_H\right) = \frac{1}{20}\left(8 \times 165 + 12 \times 175\right) = 171.$$

Ainsi, par le théorème de décomposition de variance :

$$\begin{aligned}s_x^2 &= \frac{n_F \times s_F^2 + n_H \times s_H^2}{n} + \frac{n_F\left(\bar{x}_F - \bar{x}\right)^2 + n_H\left(\bar{x}_H - \bar{x}\right)^2}{n}\\ &= \frac{8 \times 9 + 12 \times 16}{20} + \frac{8 \times (165-171)^2 + 12 \times (175-171)^2}{20} = 13.2 + 24 = 37.2.\end{aligned}$$

Réponse : A

Question 106.
Dans une population, on sait que le poids d'un individu (en kg) suit une loi normale de moyenne 78 et de

variance 16. Quelle est le probabilité que le poids d'un individu choisi au hasard soit entre 78 et 88 kg ?
A) 0.0987 B) 0.4938 C) 0.4944 D) 0.5000 E) 0.5987 F) 0.8944 G) 0.9938 H) 1
I) Aucune des réponses ci-dessus n'est correcte

Solution
$\Pr(78 \leq X \leq 88) = \Phi\left(\frac{88-78}{4}\right) - \Phi\left(\frac{78-78}{4}\right) = \Phi(2.5) - \Phi(0) = 0.9938 - 0.5 = 0.4938.$
Réponse : B

Question 107.
On considère la série statistique $x_1, \ldots, x_5$.
On sait que $\sum_{j=1}^{5} x_j = 8$ et que $\sum_{j=1}^{5} x_j^2 = 16$. Que vaut $\sum_{i=1}^{2}\sum_{j=1}^{5}(x_j - i)^2$?
A) 0 B) 8 C) 9 D) 11 E) 16 F) 25
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\sum_{i=1}^{2}\sum_{j=1}^{5}(x_j - i)^2 = \sum_{i=1}^{2}\sum_{j=1}^{5}(x_j^2 - 2ix_j + i^2) = \sum_{i=1}^{2}\sum_{j=1}^{5} x_j^2 - 2\sum_{i=1}^{2} i \sum_{j=1}^{5} x_j + \sum_{i=1}^{2} i^2 \sum_{j=1}^{5} 1$$
$$= \sum_{i=1}^{2} 16 - 2\sum_{i=1}^{2} i \times 8 + 5\sum_{j=1}^{5} 1 = 2 \times 16 - 2 \times 3 \times 8 + 5 \times 5 = 9$$

Réponse : C

Question 108.
Une compagnie d'assurance offre une assurance avec une prime annuelle de 2 000CHF. Un client n'a pas de réclamation dans une année avec une probabilité de 7/10. Il a une réclamation de 5 000CHF dans l'année avec une probabilité de 2/10 et il a une réclamation de 10 000CHF avec une probabilité de 1/10. Quelle est le profit espéré de la compagnie d'assurance ?
A) -10 000 B) -5 000 C) -2 000 D) 0 E) 2 000 F) 5 000 G) 10 000 H) 100 000
I) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\mathrm{E}(P) = 2000 \times \frac{7}{10} + (2000 - 5000) \times \frac{2}{10} + (2000 - 10000) \times \frac{1}{10} = 0.$$

Réponse : D

Question 109.
Supposons que l'indice simple du litre d'essence sans plomb soit $I(2000/1990) = 142$. Sachant qu'en l'an 2000, un individu doit payer 10 francs pour remplir un bidon d'essence contenant 5 litres, quelle somme aurait-il approximativement dépensé en 1990 pour la même quantité d'essence ?
A) 10 francs B) 14.2 francs C) 1.41 francs D) 3.52 francs E) 7.04 francs F) 7.10 francs
G) Aucune des réponses ci-dessus n'est correcte

Solution
Pour un litre : $I(1/0) = 100\frac{P_1}{P_0} = 100\frac{2}{P_0} = 142$. Donc, $P_0 = \frac{100\times 2}{142} \approx 1.408$.
Pour 5 litres = 7.04
Réponse : E

Question 110.
On considère la série temporelle suivante :

t	1	2	3	4	5
y_t	8	6	3	5	7

Que valent respectivement $F^3 y_1$ et $\nabla^2 y_5$?
A) 8 et 2 B) 5 et 4 C) 8 et 0 D) 5 et 0 E) 6 et 2 F) 6 et 4
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\begin{aligned} F^3 y_1 &= y_{1+3} = y_4 = 5 \\ \nabla^2 y_5 &= I y_5 - 2L y_5 + L^2 y_5 = y_5 - 2y_4 + y_3 = 7 - 2 \times 5 + 3 = 0 \end{aligned}$$

Réponse : D

Question 111.
La Figure 13.3 représente les courbes de Lorenz de la distribution des revenus au sein du pays A et du pays B. On note G_A et G_B l'indice de Gini du pays A et du pays B respectivement et H_A et H_B l'indice de Hoover du pays A et du pays B respectivement. Quelle information peut-on déduire à l'aide de ce graphique ?

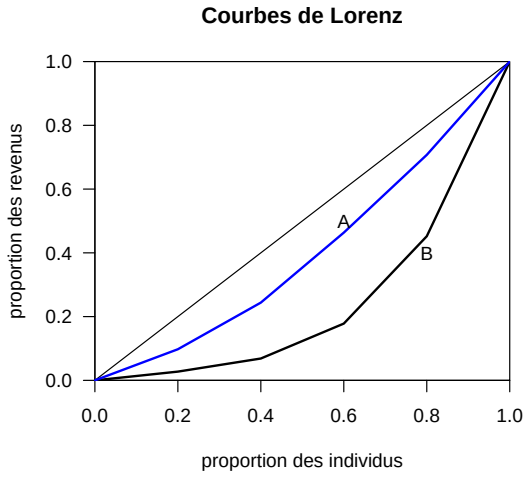


FIGURE 13.3 – Courbes de Lorenz des pays A et B

A) $G_A > G_B$ et $H_A > H_B$ B) $G_A < G_B$ et $H_A < H_B$
C) $G_A > G_B$ et $H_A < H_B$ D) $G_A < G_B$ et $H_A > H_B$
E) Les habitants du pays B sont en moyenne plus riches que ceux du pays A.
F) Les habitants du pays A sont en moyenne plus riches que ceux du pays B.

Solution
Aire entre la diagonale est la courbe (Gini) plus petite pour A. Distance verticale maximale entre la droite et la courbe (Hoover) plus petite pour A également.
Réponse : B

Question 112.
Dans une petite ville, pendant 100 jours, on a noté le nombre de naissances par jour. Le tableau suivant contient les effectifs de jours correspondant au nombre de naissances par jour :

Naissances	0	1	2	3	4
Effectifs	20	45	10	15	10

Considérons les trois affirmations suivantes :
I. La moyenne et le mode sont égaux.

II. La moyenne est supérieure à la médiane.
III. La médiane et le mode sont égaux.

A) seul I est vrai
B) seul II est vrai
C) seul III est vrai
D) I et II sont vrais mais III est faux
E) I et III sont vrais mais II est faux
F) II et III sont vrais mais I est faux
G) I, II et III sont faux

Solution
$\bar{x} = 1.5$, $x_{1/2} = 1$, $x_M = 1$.
Réponse : F

Question 113.
Soit la variable x représentant la quantité de neige tombée (en mm). La série statistique suivante est la valeur de x pour les 7 jours d'une semaine :

0	0	0	10	12	15	44

Si on note $\bar{x}$ la moyenne, $x_{1/2}$ la médiane et x_M le mode de cette série, quelle affirmation est vraie ?

A) $\bar{x} < x_{1/2} < x_M$
B) $x_{1/2} < \bar{x} < x_M$
C) $x_{1/2} < \bar{x} = x_M$
D) $x_M < \bar{x} < x_{1/2}$
E) $x_M < x_{1/2} < \bar{x}$
F) $\bar{x} = x_{1/2} = x_M$
G) Aucune des réponses ci-dessus n'est correcte.

Solution
Comme $\bar{x} = 11.57$, $x_M = 0$, $x_{1/2} = 10$, on a $x_M < x_{1/2} < \bar{x}$.
Réponse : E

Question 114.
Quelle formule ne représente pas la variance d'une série x_i ?

A) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
B) $\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$
C) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})x_i$
D) $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)^2$
E) $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i x_j$

Solution
Réponse : E

Question 115.
Alexis apprend à skier. Dans une journée, il descend 6 fois la même piste. Les conditions sont identiques à chaque descente. À la fin de chaque descente, Alexis a réussi à ne pas tomber avec une probabilité 0.45. Quelle est la probabilité qu'Alexis réussisse exactement 2 descentes sans tomber ?

A) 0.278 B) 0.919 C) 0.186 D) 0.116 E) 0.333 F) 0.500

Solution
On a ici

$$X \sim \mathcal{B}(6, 0.45).$$

Donc,

$$\Pr(X = 2) = \frac{6!}{2!4!}0.45^2 0.55^4 = 0.278.$$

Réponse : A

Question 116.
Soit une série statistique représentant la température relevée (en degrés Celsius) au même moment dans 10

stations de ski suisses. On a calculé la moyenne et la variance de cette série et on a obtenu les résultats suivants : $\bar{x} = -5$ et $s_x^2 = 100$. On souhaite maintenant construire la série y_i à l'aide de la relation $y_i = 1.8x_i + 32$ et ainsi convertir les degrés Celsius en degrés Fahrenheit. Combien valent respectivement $\bar{y}$ et s_y^2 ?

A) 23 et 356 B) 23 et 324 C) 23 et 180 D) 27 et 180
E) 27 et 324 F) 27 et 100 G) 27 et 180 H) 23 et 180

Solution

$$\bar{y} = 1.8\bar{x} + 32 = 1.8 \times (-5) + 32 = 23$$
$$s_y^2 = 1.8^2 \times s_x^2 = 3.24 \times 100 = 324$$

Réponse : B

Question 117.
Julien veut voler une paire de skis à la montagne. Il sait que la paire qu'il désire est dans un certain casier. Pour ouvrir ce casier, il faut composer un code de 4 lettres sur un clavier qui contient les 9 lettres suivantes : A, B, C, D, E, F, G, H et I. Chaque lettre peut apparaître au maximum une fois par code. Combien existe-t-il de différents codes possibles ?

A) 3600 B) 6561 C) 126 D) 36 E) 3024 F) 4000
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$A_4^9 = \frac{9!}{(9-4)!} = 3024.$$

Réponse : E

Question 118.
Dans une station de ski, le nombre de billets vendus dans une journée suit une loi normale de moyenne 1000 et d'écart-type 150. Une journée est dite "peu rentable" si moins de 700 billets sont vendus. Une journée est dite "très rentable" si plus de 1300 billets sont vendus. Quelle est la proportion de journées dites "très rentables" ou "peu rentables" ?

A) 0.9772 B) 0.0456 C) 0.0228 D) 0.9544 E) 0.5000 F) Aucune de ces réponses

Solution
$X \sim \mathcal{N}(1000, 150^2)$

$$\begin{aligned}\Pr(X > 1300) + \Pr(X < 700) &= \Pr\left(X > \frac{1300-1000}{150}\right) + \Pr\left(X < \frac{700-1000}{150}\right)\\ &= \Pr(Z > 2.00) + \Pr(Z < -2.00) = 2\left[1 - \Pr(X \leq 2.00)\right] = 2\left[1 - 0.9772\right] = 0.0456.\end{aligned}$$

Réponse : B

Question 119.
À la caisse d'une cafétéria, on peut tourner une roue de la chance suite à un achat. Cette roue est composée de 20 parties égales. La probabilité de tomber sur chacune des 20 parties est la même. Une partie permet de gagner 5CHF, trois parties permettent de gagner 2.50CHF et seize parties indiquent "meilleure chance la prochaine fois". Madeleine a exactement 2.50CHF dans ses poches. Elle utilise ce montant au complet pour acheter un chocolat chaud, puis elle tourne la roue. En sortant de la cafétéria, combien d'argent peut-elle espérer avoir dans ses poches ?

A) 1.755 B) 0.625 C) -1.750 D) -1.875 E) 2.500 F) 0.000
G) Aucune des réponses ci-dessus n'est correcte

Solution

$$\mathrm{E}(X) = 0 \times \frac{16}{20} + 2.50 \times \frac{3}{20} + 5 \times \frac{1}{20} = 0.625.$$

Réponse : B

Question 120.
On considère la série temporelle suivante :

t	1	2	3	4	5	6
y_t	22	18	16	23	20	17

Que valent respectivement Fy_4 et $\nabla^2 y_3$?
A) 18 et 0 B) 20 et 2 C) 17 et 24 D) 16 et 18 E) 20 et 18 F) 16 et 2

Solution

$$\begin{aligned} Fy_4 &= y_{4+1} = y_5 = 20 \\ \nabla^2 y_3 &= Iy_3 - 2Ly_3 + L^2 y_3 = y_3 - 2y_2 + y_1 = 16 - 2 \times 18 + 22 = 2 \end{aligned}$$

Réponse : B

Question 121.
On a une série temporelle dont l'ordre de la saisonnalité est 3. Notons S_i la composante saisonnière i et S'_i la composante saisonnière ajustée i, avec $i = 1, 2, 3$. Après avoir désaisonnalisé cette série temporelle par la méthode additive, on a obtenu les composantes saisonnières suivantes :

S_1	=	30	S'_1	=	?
S_2	=	-42	S'_2	=	-45
S_3	=	?	S'_3	=	?

Les composantes saisonnières S_3, S'_1 et S'_3 sont respectivement
A) 36, 33 et 39 B) 25, 33 et 28 C) -7, 27 et -10 D) -21, 41 et -10 E) -21, 27 et -24 F) 21, 27 et 18

Solution

$$S'_2 = S_2 - \frac{1}{3}\sum_{i=1}^{3} S_i$$

$$-45 = -42 - \frac{1}{3}(30 - 42 + S_3) = -38 - \frac{1}{3}S_3$$

$$-7 = -\frac{1}{3}S_3$$

$$S_3 = 21$$

$$\frac{1}{3}\sum_{i=1}^{3} S_i = \frac{1}{3}(30 - 42 + 21) = 3$$

$$S'_1 = S_1 - \frac{1}{3}\sum_{i=1}^{3} S_i = 30 - 3 = 27$$

$$S'_3 = S_3 - \sum_{i=1}^{3} S_i = 21 - 3 = 18$$

Réponse : F

Question 122.
On admet que la longueur du pied d'un homme adulte suit une loi normale de moyenne 25 cm et de variance égale à 9. Quelle est la probabilité qu'un individu choisi au hasard ait un pied d'une longueur supérieure à 26.5 cm ?
A) 0.3085 B) 0.5675 C) 0.6915 D) 0.4325 E) 0.5000 F) 0.9025
G) Aucune des réponses ci-dessus n'est correcte

Solution
Comme $X \sim \mathcal{N}(25, 9)$, on a

$$\begin{aligned} F(x) &= \Phi\left(\frac{x-25}{3}\right), \\ \Pr(X \leq 26.5) &= \Phi\left(\frac{26.5-25}{3}\right) = \Phi(0.5) = 0.6915, \\ \Pr(X \geq 26.5) &= 1 - 0.6915 = 0.3085. \end{aligned}$$

Réponse : A

Question 123.
Les données ci-dessous représentent le nombre d'années d'ancienneté de 20 professeurs d'une école publique :

2	3	3	5	6	8	8	8	8	10
12	13	13	14	14	18	22	26	31	32

La distance interquartile (IQ) de cette distribution vaut :
A) 8 B) 10.5 C) 30 D) 9 E) 15.5 F) 7 G) 11 H) 7.5
I) Aucune des réponses ci-dessus n'est correcte

Solution

$$x_{np} = x_{20\times\frac{1}{4}} = x_5 \Rightarrow x_{1/4} = \frac{x_5 + x_6}{2} = 7$$
$$x_{np} = x_{20\times\frac{3}{4}} = x_{15} \Rightarrow x_{3/4} = \frac{x_{15} + x_{16}}{2} = 16$$
$$IQ = 16 - 7 = 9$$

Réponse : D

Question 124.
Une entreprise fabrique des ampoules électriques dont la durée de vie (en heures) suit une distribution normale de moyenne 500h et d'écart-type 100h. Quelle est la probabilité (arrondie à 3 décimales) que la durée de vie d'une ampoule soit supérieure à 617h ?
A) 0.879 B) 0.021 C) 0.121 D) 0.521 E) 0.721 F) 0.279 G) 1 H) 0.479
I) Aucune des réponses ci-dessus n'est correcte.

Solution

$$\Pr(X > 617) = 1 - \Pr(X < 617) = 1 - \Pr(Z < (617-500)/100) = 1 - 0.879 = 0.121.$$

Réponse : C

Question 125.
Lors d'une compétition d'athlétisme, huit concurrents participent à la course de 100 mètres. De combien de manières différentes le podium (médailles d'or, d'argent et de bronze) peut-il être formé ?
A) 336 B) 24 C) 6720 D) 1680 E) 56 F) 40320
G) Aucune des réponses ci-dessus n'est correcte

Solution
Il s'agit d'un arrangement simple, car l'ordre d'arrivée compte. $A_8^3 = \frac{8!}{(8-3)!} = 336$.
Réponse : A

Chapitre 14

Questions ouvertes

Exercice 1.
Considérer les données dans le Tableau 14.1 qui nous montrent une distribution des salaires (en euros par mois) parmi les 100 employés d'une entreprise :

Table 14.1 – Distribution des salaires

Classe salariale	Effectif	Effectif cumulée	Fréquence	Fréquence cumulée
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[1000, 1500[$	30			
$[1500, 2000[$	10			
$[2000, 3000[$	60			
Total	100	–	1.00	–

1. Compléter le Tableau 14.1.

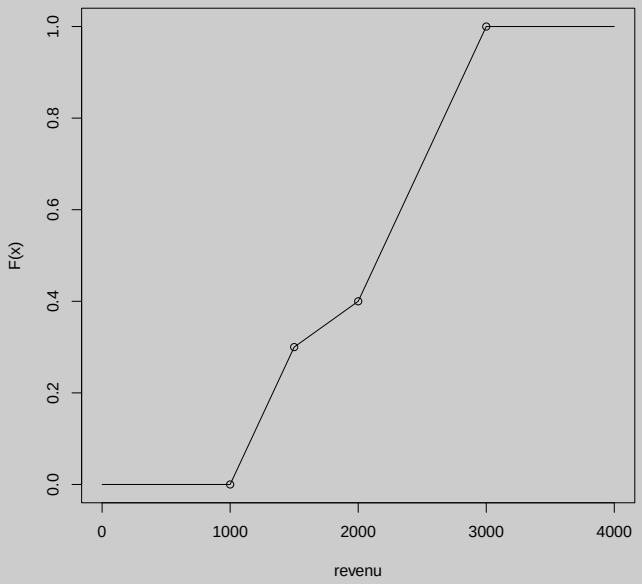
2. Dessiner la fonction de répartition.

Solution

Classe salairiale	Effectif	Effectif cumulée	Fréquence	Fréquence cumulée
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[1000, 1500[$	30	30	0.3	0.3
$[1500, 2000[$	10	40	0.1	0.4
$[2000, 3000[$	60	100	0.6	1
Total	100	–	1.00	–

En langage R

```
X=rbind(c(0,0),c(1000,0),c(1500,0.3),
c(2000,0.4),c(3000,1),c(4000,1))
plot(X,type="n",ylab="F(x)",xlab="revenu")
plot(X,type="l",ylab="F(x)",xlab="revenu")
points(X[2:5,])
```


F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
revenu

Exercice 2.
Une substance, utilisée dans la recherche en biologie et medecine, est transportée par fret aérien jusqu'aux utilisateurs dans des cartons de 1000 ampoules. Les données du Tableau 14.2, qui comprennent 10 frets, contiennent le nombre de fois qu'un carton a été transféré d'un avion à l'autre au cours de la livraison (X) et le nombre d'ampoules cassées qu'on a trouvées à la fin de la livraison (Y).

TABLE 14.2 – Données de la livraison d'ampoules

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	16			
2	0	9			
3	2	17			
4	0	12			
5	3	22			
6	1	13			
7	0	8			
8	1	15			
9	2	19			
10	0	11			
Somme					
Moyenne					

1. Compléter les cellules vides du tableau des données.
2. Calculer les variances marginales de X et Y.
3. Calculer la covariance entre X et Y.
4. Déterminer le coefficient de corrélation entre les variables X et Y.
5. Déterminer l'équation de la droite de régression de Y en fonction de X.
6. Déterminer la qualité de cet ajustement.
7. Donner la valeur ajustée et le résidu pour la première observation du tableau.

Solution

1. Tableau complété :

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	1	16	1	256	16
	0	9	0	81	0
	2	17	4	289	34
	0	12	0	144	0
	3	22	9	484	66
	1	13	1	169	13
	0	8	0	64	0
	1	15	1	225	15
	2	19	4	361	38
	0	11	0	121	0
Somme	10	142	20	2194	182
Moyenne	1	14.2	2	219.4	18.2

2. Variances marginales de X et Y :

$$\bar{x} = 1, \ \bar{y} = 14.2,$$

$$s_x^2 = 2 - 1^2 = 2 - 1 = 1,$$

$$s_y^2 = 219.4 - 14.2^2 = 219.4 - 201.64 = 17.76.$$

3. Covariance entre X et Y :

$$s_{xy} = 18.2 - 1 \times 14.1 = 18.2 - 14.2 = 4.$$

4. Coefficient de corrélation entre les variables X et Y :

$$r_{xy} = \frac{4}{\sqrt{1 \times 17.76}} = \frac{4}{4.214} = 0.949.$$

5. Equation de la droite de régression de Y en fonction de X :

$$D_{y|x} \ : y = a + b\ x, \text{ avec } b = \frac{s_{xy}}{s_x^2} = 4 \text{ et } a = \bar{y} - \frac{s_{xy}}{s_x^2} x = 14.2 - 4 \times 1 = 10.2,$$

$$D_{y|x} \ : y = 4x + 10.2.$$

6. Qualité de l'ajustement : $r^2 = 0.949^2 = 0.901 \Rightarrow$ très bon ajustement.
7. Valeur ajustée et résidu pour la première observation du Tableau 14.2 :

$$y_1^* = 4 \times 1 + 10.2 = 14.2, \ e_1 = 16 - 14.2 = 1.8.$$

Exercice 3.
Considérer la série trimestrielle donnée dans le Tableau 14.3. En l'examinant, on a remarqué que la série présente une tendance saisonnière. On estime la tendance par une moyenne mobile. On a toutes les valeurs sauf pour les neuvième et dixième trimestres. Calculer ces valeurs manquantes, en écrivant explicitement la formule pour la moyenne mobile que vous utilisez.

Table 14.3 – Données trimestrielles

Trimestre(t)	Série(Y_t)	Moyenne mobile(T_t)
1	56	
2	116	
3	128	94.375
4	64	104.125
5	83	113.375
6	167	115.625
7	151	115.625
8	59	117.875
9	88	
10	180	
11	179	
12	93	

Solution
La moyenne mobile est

$$MM(4) = \frac{L^2 + 2L + 2I + 2F + F^2}{8}.$$

Donc,

$$T_9 = \frac{151 + 2 \times 59 + 2 \times 88 + 2 \times 180 + 179}{8} = 123,$$

$$T_{10} = \frac{59 + 2 \times 88 + 2 \times 180 + 2 \times 179 + 93}{8} = 130.75.$$

Le tableau complété :

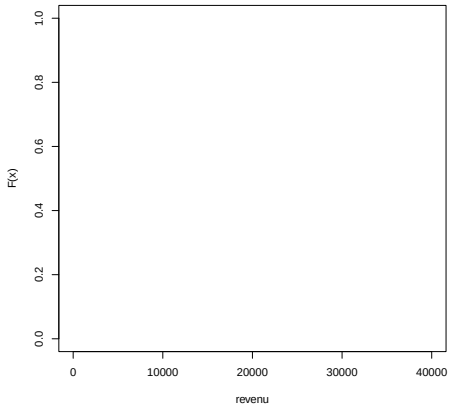
Trimestre(t)	Série(Y_t)	Moyenne mobile(T_t)
1	56	
2	116	
3	128	94.375
4	64	104.125
5	83	113.375
6	167	115.625
7	151	115.625
8	59	117.875
9	88	*123*
10	180	*130.75*
11	179	
12	93	

Exercice 4.
On dispose des résultats d'une enquête concernant les loyers annuels des appartements (en francs suisses) d'une commune. La distribution des loyers est donnée dans le Tableau 14.4.

TABLE 14.4 – Distribution des loyers

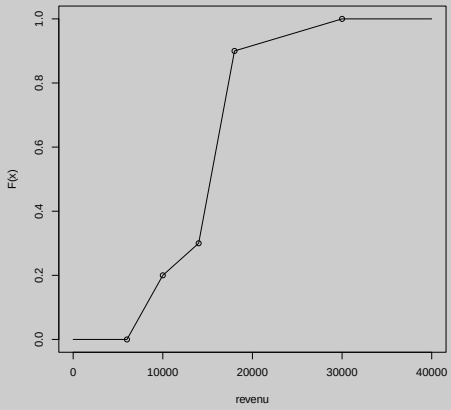
Loyers	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
[6000, 10000[	200			
[10000, 14000[	100			
[14000, 18000[	600			
[18000, 30000[	100			
Total	1000	–	1.00	–

1. Complétez le Tableau 14.4.
2. Dessinez la fonction de répartition dans le repère ci-dessous :



Solution

Loyers	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
[6000, 10000[	200	200	0.2	0.2
[10000, 14000[	100	300	0.1	0.3
[14000, 18000[	600	900	0.6	0.9
[18000, 30000[	100	1000	0.1	1
Total	1000	–	1	–



En langage R

```
X=cbind(c(0,6000,10000,14000,18000,30000,40\,000),
c(0,0,0.2,0.3,0.9,1,1))
plot(X,type="n",ylab="F(x)",xlab="revenu")
plot(X,type="l",ylab="F(x)",xlab="revenu")
points(X[2:6,])
```

Exercice 5.
Dans le but d'étudier les effets du bruit sur la productivité des travailleurs, on a relevé des données dans des différentes parties d'une entreprise. Les niveaux de bruits (X), mesurés en décibels et la productivité (Y) de 8 travailleurs sont donnés dans le Tableau 14.5.

TABLE 14.5 – Bruit et productivité

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	10	22	100	484	220
	10	24	100	576	240
	30	18	900	324	540
	30	19	900	361	570
	50	18	2500	324	900
	50	14	2500	196	700
	70	11	4900	121	770
	70	14	4900	196	980
Somme	320	140	16800	2582	4920
Moyenne	40	17.5	2100	322.75	615

1. Calculez les variances marginales de X et Y.
2. Calculez la covariance entre X et Y.
3. Déterminez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez les coefficients de corrélation et de détermination entre les variables X et Y et ensuite déterminez la qualité de l'ajustement.
5. Donnez la valeur ajustée et le résidu pour la première observation du Tableau 14.5.

Solution

1. Variances marginales de X et Y :
$$\bar{x} = 40,\ \bar{y} = 17.5,$$
$$s_x^2 = 2100 - 40^2 = 2100 - 1600 = 500,$$
$$s_y^2 = 322.75 - 17.5^2 = 322.75 - 306.25 = 16.5.$$
2. Covariance entre X et Y :
$$s_{xy} = 615 - 40 \times 17.5 = 615 - 700 = -85.$$
3. Équation de la droite de régression de Y en fonction de X. Ajustement linéaire de y en x
$$D_{y|x}\ : y = a + b\ x,\ \text{avec } b = \frac{s_{xy}}{s_x^2} = \frac{-85}{500} = -0.17 \text{ et } a = \bar{y} - b\ \bar{x} = 17.5 - (-0.17) \times 40 = 24.3,$$
$$D_{y|x}\ : y = 24.3 - 0.17x.$$
4. Coefficients de corrélation et de détermination entre les variables X et Y et qualité de l'ajustement :
$$r_{xy} = \frac{-85}{\sqrt{500 \times 16.5}} = \frac{-85}{90.83} = -0.936$$
$r^2 = (-0.936)^2 = 0.876 \Rightarrow$ très bon ajustement.
5. Valeur ajustée et le résidu pour la première observation du Tableau 14.5 :
$$y_1^* = -0.17 \times 10 + 24.3 = 22.6,\ e_1 = 22 - 22.6 = -0.6.$$

Exercice 6.
Soit le Tableau 14.6 contenant les prix et les quantités des trois biens blé, maïs et sucre, qu'on note, respectivement, A, B et C, sur les trois années 1990, 1991 et 1992, qu'on note, respectivement, années 0, 1 et 2 :

Table 14.6 – Tableau des prix et quantités

Année	0		1		2	
Bien	Prix	Quantité	Prix	Quantité	Prix	Quantité
A	10	6	16	5	20	4
B	20	6	25	8	22	8
C	5	8	15	7	10	8

1. Calculez les indices de Laspeyres des prix $L(1/0)$ et $L(2/1)$.
2. Quelle est la valeur de l'indice chaîne de Laspeyres $CL(2/0)$?

Solution

1. Indice de Laspeyre des prix :

$$\begin{aligned} L(1/0) &= 100 \times \frac{q_A(0)p_A(1) + q_B(0)p_B(1) + q_C(0)p_C(1)}{q_A(0)p_A(0) + q_B(0)p_B(0) + q_C(0)p_C(0)} \\ &= 100 \times \frac{6 \times 16 + 6 \times 25 + 8 \times 15}{6 \times 10 + 6 \times 20 + 8 \times 5} = 100 \times \frac{366}{220} = 166.364. \end{aligned}$$

$$\begin{aligned} L(2/1) &= 100 \times \frac{q_A(1)p_A(2) + q_B(1)p_B(2) + q_C(1)p_C(2)}{q_A(1)p_A(1) + q_B(1)p_B(1) + q_C(1)p_C(1)} \\ &= 100 \times \frac{5 \times 20 + 8 \times 22 + 7 \times 10}{5 \times 16 + 8 \times 25 + 7 \times 15} = 100 \times \frac{346}{385} = 89.87. \end{aligned}$$

2. Indice chaîne de Laspeyres :

$$\begin{aligned} CL(2/0) &= 100 \times \frac{L_p(2/1)}{100} \times \frac{L_p(1/0)}{100} \\ &= 100 \times \frac{166.364}{100} \times \frac{89.87}{100} = \frac{14951.132}{100} = 149.511. \end{aligned}$$

Exercice 7.
On considère la série trimestrielle donnée dans le Tableau 14.7. En l'examinant, on a remarqué que la série présente une tendance saisonnière. On a estimé la tendance par une moyenne mobile d'ordre 4 dont on a calculé toutes les valeurs. Désaisonnalisez cette série par la méthode additive. Complétez ensuite le Tableau 14.12.

Table 14.7 – Données trimestrielles

Année	Trimestre	Série	MM(4)	Différence	Composante saisonnière	Ajustement	Série désaisonnalisée
(a)	(m)	(Y_{am})	(T_{am})	$(Y_{am} - T_{am})$	(S_m)	(S'_m)	$(Y_{am} - S'_m)$
1	1	695	_	_			
1	2	706	_	_			
1	3	684	699.25	−15.25			
1	4	703	699.5	3.5			
2	1	713	701.375	11.625	_	_	
2	2	690	707.875	−17.875	_	_	
2	3	715	709.625	5.375	_	_	
2	4	724	713.25	10.75	_	_	
3	1	706	720	−14	_	_	
3	2	726	720.625	5.375	_	_	
3	3	733	_	_	_	_	
3	4	711	_	_	_	_	

Solution
Calcul des composantes saisonnières

$$S_m = \frac{1}{M}\sum_a (Y_{am} - T_{am}).$$

$$S'_m = S_m - \frac{1}{M}\sum_m S_m.$$

$$\widetilde{Y}_{am} = Y_{am} - S'_m.$$

$$S_1 = \frac{1}{2}[(Y_{21} - T_{21}) + (Y_{31} - T_{31})] = \frac{1}{2}(11.625 - 14) = -1.188.$$

$$S_2 = \frac{1}{2}[(Y_{22} - T_{22}) + (Y_{32} - T_{32})] = \frac{1}{2}(-17.875 + 5.375) = -6.25.$$

$$S_3 = \frac{1}{2}[(Y_{13} - T_{13}) + (Y_{23} - T_{23})] = \frac{1}{2}(-15.25 + 5.375) = -4.938.$$

$$S_4 = \frac{1}{2}[(Y_{14} - T_{14}) + (Y_{24} - T_{24})] = \frac{1}{2}(3.5 + 10.75) = 7.125.$$

$$\frac{1}{M}\sum_m S_m = \frac{1}{4} \times (-5.25) = -1.313.$$

$$S'_1 = S_1 - \frac{1}{M}\sum_m S_m = -1.188 + 1.313 = 0.125.$$

$$S'_2 = S_2 - \frac{1}{M}\sum_m S_m = -6.25 + 1.313 = -4.938.$$

$$S'_3 = S_3 - \frac{1}{M}\sum_m S_m = -4.938 + 1.313 = -3.625.$$

$$S'_4 = S_4 - \frac{1}{M}\sum_m S_m = 7.125 + 1.313 = 8.438.$$

Tableau complété

Année	Trimestre	Série	MM(4)	Différence	Composante saisonnière	Ajustement	Série désaisonnalisée
(a)	(m)	(Y_{am})	(T_{am})	$(Y_{am} - T_{am})$	(S_m)	(S'_m)	$(Y_{am} - S'_m)$
1	1	695	_	_	−1.188	0.125	694.875
1	2	706	_	_	−6.25	−4.938	710.938
1	3	684	699.25	−15.25	−4.938	−3.625	687.625
1	4	703	699.5	3.5	7.125	8.438	694.563
2	1	713	701.375	11.625	_	_	712.875
2	2	690	707.875	−17.875	_	_	694.938
2	3	715	709.625	5.375	_	_	718.625
2	4	724	713.25	10.75	_	_	715.563
3	1	706	720	−14	_	_	705.875
3	2	726	720.625	5.375	_	_	730.938
3	3	733	_	_	_	_	736.625
3	4	711	_	_	_	_	702.563

Exercice 8.
Une enquête sur 200 individus a été faite afin de rendre compte de la structure de la population en Suisse. Les résultats sont dans le Tableau 14.8 :

TABLE 14.8 – Structure de la population

Classe d'âge	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[0, 20[$	44			
$[20, 40[$	54			
$[40, 65[$	70			
$[65, 100[$	32			
Total	200	–	1.00	–

1. Complétez le Tableau 14.8.
2. Dessinez l'histogramme dans le repère de la page suivante :

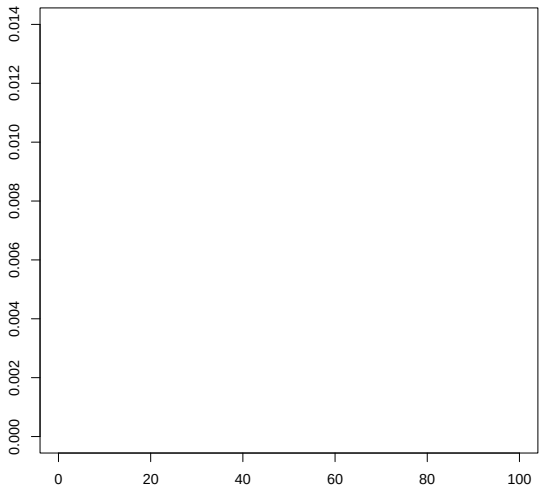


FIGURE 14.1 – Histogramme

Solution

Structure de la population

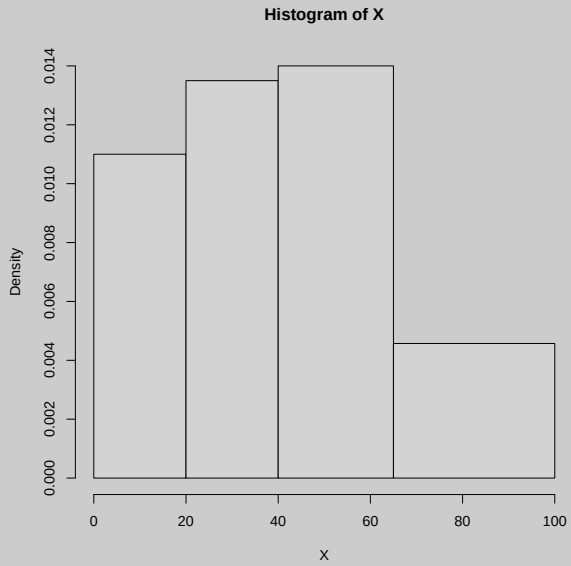
Classe d'âge	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+[$	n_j	N_j	f_j	F_j
$[0, 20[$	44	44	0.22	0.22
$[20, 40[$	54	98	0.27	0.49
$[40, 65[$	70	168	0.35	0.84
$[65, 100[$	32	100	0.16	1
Total	200	–	1	–

En ce qui concerne la hauteur de l'histogramme :

$$h_j = \frac{f_j}{c_j^+ - c_j^-}$$

$$h_1 = \frac{n_1}{c_1^+ - c_1^-} = \frac{0.22}{20} = 0.011, h_2 = \frac{n_2}{c_2^+ - c_2^-} = \frac{0.27}{20} = 0.0135,$$

$$h_3 = \frac{n_3}{c_3^+ - c_3^-} = \frac{0.35}{25} = 0.014, h_4 = \frac{n_4}{c_4^+ - c_4^-} = \frac{0.16}{35} = 0.0045.$$



Exercice 9.
On a relevé le coût de la santé en une année de 10 individus d'âge différents. Y représente le coût de la santé, en milliers de CHF et X représente l'âge des individus. Vous trouverez ces données dans le Tableau 14.9 :

TABLE 14.9 – Âge et coût de la santé

x_i	3	14	25	36	43	55	62	77	89	96
y_i	4	7	8	9	10	14	21	34	66	127

On sait en outre que :

$$\sum_{i=1}^{10} x_i = 500, \quad \sum_{i=1}^{10} y_i = 300, \quad \sum_{i=1}^{10} x_i y_i = 23820, \quad \sum_{i=1}^{10} x_i^2 = 33910, \quad \sum_{i=1}^{10} y_i^2 = 22588.$$

1. Calculez les variances marginales de X et Y.
2. Calculez la covariance entre X et Y.
3. (a) Déterminez l'équation de la droite de régression de Y en fonction de
   (b) Expliquez en trois lignes au maximum pourquoi on fait ici une régression de $Y en X$ plutôt qu'une régression de X en Y.
4. Déterminez les coefficients de corrélation et de détermination entre les variables X et Y et ensuite déterminez la qualité de l'ajustement.
5. Donnez la valeur ajustée et le résidu pour la dernière observation du Tableau 14.9.
6. Estimez à l'aide du modèle le coût de la santé d'un individu ayant 23 ans.

Solution

1. Variances marginales de X et Y :
$$\bar{x} = 50, \ \bar{y} = 30,$$
$$s_x^2 = 1/10 \times 33910 - 50^2 = 3391 - 2500 = 891,$$
$$s_y^2 = 1/10 \times 22588 - 30^2 = 2252.8 - 900 = 1352.8.$$
2. Covariance entre X et Y :
$$s_{xy} = 1/10 \times 23820 - 50 \times 30 = 2382 - 1500 = 882.$$
3. (a) Equation de la droite de régression de Y en fonction de X.
$$Dy|x \ : y = a + b\ x, \text{ avec } b = \frac{s_{xy}}{s_x^2} = \frac{882}{891} = 0.99 \text{ et } a = \bar{y} - b\ \bar{x} = 30 - \frac{882}{891} \times 50 = -19.49,$$
$$Dy|x \ : y = -19.49 + 0.99x.$$
   (b) C'est l'âge qui explique le coût et non pas le coût qui explique l'âge.
4. Coefficients de corrélation et de détermination entre les variables X et Y et qualité de l'ajustement :
$$r_{xy} = \frac{882}{\sqrt{891 \times 1352.8}} = \frac{882}{1097.88} = 0.803.$$
$r^2 = (0.803)^2 = 0.645 \Rightarrow 64\%$ de la variance est expliquée par le modèle.
5. Valeur ajustée et le résidu pour la dernière observation du Tableau 14.9 :
$$y_{10}^* = 0.99 \times 96 - 19.49 = 75.55, \ e_1 = 127 - 75.55 = 51.45.$$
6. Coût de la santé d'un individu ayant 23 ans :
$$y_{23ans}^* = 0.99 \times 23 - 19.49 = 3.28.$$

Exercice 10.
On aimerait savoir s'il y a une relation entre le sexe de 200 personnes et leur taux d'occupation professionnelle. Le Tableau 14.10 représente le tableau de contingence.

Table 14.10 – Tableau de contingence

	moins de 50%	50%	plus de 50%	Total
Homme	5	50	45	100
Femme	45	50	5	100
Total	50	100	50	200

Calculez le V de Cramer et commentez le résultat obtenu.

Solution

Tableau des effectifs théoriques

	moins de 50%	50%	plus de 50%	Total
Homme	25	50	25	100
Femme	25	50	25	100
Total	50	100	50	200

Tableau des écarts à l'indépendance

	moins de 50%	50%	plus de 50%	Total
Homme	−20	0	20	0
Femme	20	0	−20	0
Total	0	0	0	0

Tableau des $\frac{e_{jk}^2}{n_{jk}^*}$

	moins de 50%	50%	plus de 50%	Total
Homme	16	0	16	32
Femme	16	0	16	32
Total	32	0	32	64

$$\chi^2_{obs} = 16 + 16 + 16 + 16 = 64$$

$$\phi^2 = \frac{\chi^2_{obs}}{200} = 64/200 = 0.32$$

Exercice 11.
On dispose des revenus mensuels de 5 employés de deux entreprises A et B. Ces données se trouvent dans le Tableau 14.11. On aimerait savoir si les revenus sont distribués de la même manière dans ces deux entreprises.

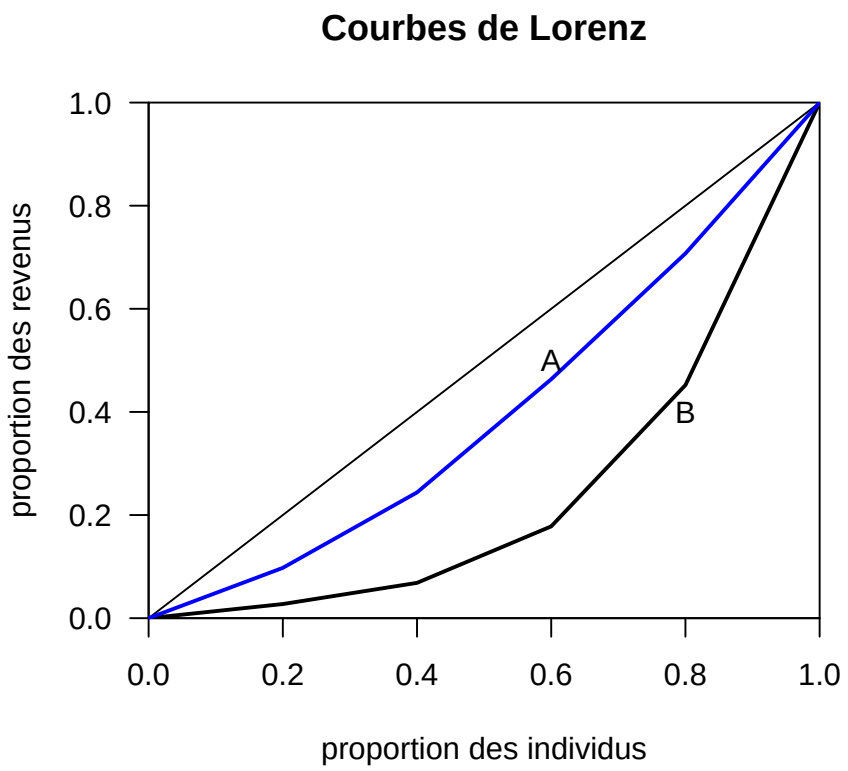
TABLE 14.11 – Salaires des employés de deux entreprises

A	B
2000	2000
3000	3000
4500	8000
5000	20000
6000	40 000

Les indices de Gini et de Hoover pour l'entreprise A sont respectivement :

$$G_A \cong 0.24, H_A \cong 0.16$$

Le graphique représente les courbes de Lorenz des deux entreprises.



1. Est-ce que la distribution des revenus est plus égalitaire au sein de l'entreprise A ou de l'entreprise B ? Répondez en vous basant sur les courbes de Lorenz.
2. Calculez les indices de Gini et de Hoover pour l'entreprise B.
3. Même question qu'en 1 mais vous basant sur les indices G_A et G_B puis H_A et H_B. Est-ce que ces résultats vont dans le même sens que ceux obtenus en se basant sur les courbes de Lorenz ?

Solution

1. La courbe de Lorenz de A est en dessus de celle de B, elle est donc toujours plus proche de la diagonale. La distribution des salaires est donc plus inégalitaire au sein de l'entreprise B qu'au sein de l'entreprise A.

2. Les indices de Gini et de Hoover sont pour l'entreprise B :

$$G_B = \frac{1}{n-1}\left(\frac{2\sum_{i=1}^{n} ix_i}{n\bar{x}} - (n+1)\right) = \frac{1}{4}\left(\frac{2\sum_{i=1}^{5} ix_i}{5\bar{x}} - 6\right),$$

$$\sum_{i=1}^{5} ix_i = 1 \times 2000 + 2 \times 3000 + 3 \times 8000 + 4 \times 20000 + 6 \times 40\,000 = 312000,$$

$$\bar{x} = \frac{1}{5} \times (2000 + 3000 + 8000 + 20000 + 40\,000 = 14600),$$

$$G_B = \frac{1}{4}\left(\frac{2 \times 312000}{5 \times 14600} - 6\right) \approx 0.64,$$

$$H_B = \frac{\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|}{2\bar{x}} = \frac{\frac{1}{5}\sum_{i=1}^{n} |x_i - 14600|}{2 \times 14600},$$

$$\begin{aligned}
&\sum_{i=1}^{n} |x_i - 14600| \\
&= |2000 - 14600| + |3000 - 14600| + |8000 - 14600| + |20000 - 14600| + |40\,000 - 14600| \\
&= 12600 + 11600 + 6600 + 5400 + 25400 = 61600,
\end{aligned}$$

$$H_B = \frac{\frac{1}{5} \times 61600}{2 \times 14600} \approx 0.42.$$

3. On a $G_A < G_B$ donc les revenus sont plus inégalement distribués dans l'entreprise B. On a $H_A < H_B$ donc les revenus sont plus inégalement distribués dans l'entreprise B. Ces constatations sont les mêmes que celles qu'on a faite en 1.

Exercice 12.
On considère la série trimestrielle donnée dans le Tableau 14.12. Elle représente le nombre d'heures de travail par trimestre d'un individu, durant les années 2001, 2002 et 2003 (correspondants respectivement aux années 1, 2 et 3 dans le Tableau 14.12). En l'examinant, on a remarqué que la série présente une tendance saisonnière. On a estimé la tendance par une moyenne mobile d'ordre 4 dont on a calculé toutes les valeurs. Désaisonnalisez cette série par la méthode additive. Complétez ensuite le Tableau 14.12.

TABLE 14.12 – Données trimestrielles

Année	Trimestre	Série	MM(4)	Différence	Composante saisonnière	Ajustement	Série désaisonnalisée
(a)	(m)	(Y_{am})	(T_{am})	$(Y_{am} - T_{am})$	(S_m)	(S'_m)	$(Y_{am} - S'_m)$
1	1	419	_	_			
1	2	417	_	_			
1	3	403	417.875	−14.875			
1	4	430	419.5	10.5			
2	1	424	421.25	2.75	_	_	
2	2	425	421.25	3.75	_	_	
2	3	409	419.325	−10.375	_	_	
2	4	424	418.625	5.375	_	_	
3	1	415	418.25	−3.25	_	_	
3	2	428	417.5	10.5	_	_	
3	3	403	_	_	_	_	
3	4	424	_	_	_	_	

Solution

$$S_m = \frac{1}{A-1}\sum_a (Y_{am} - T_{am})$$

$$S_1 = \frac{1}{2} \times (2.75 - 3.25) = -0.25$$

$$S_2 = \frac{1}{2} \times (3.75 + 10.5) = 7.125$$

$$S_3 = \frac{1}{2} \times (-10.375 - 14.875) = -12.625$$

$$S_4 = \frac{1}{2} \times (10.5 + 5.375) = 7.9375$$

$$S'_m = S_m - \frac{1}{M}\sum_m S_m$$

$$S'_1 = S_1 - \frac{1}{4} \times (-0.25 + 7.125 - 12.625 + 7.9375) = -0.25 - 0.25 \times 2.1875 = -0.796875$$

$$S'_2 = 7.125 - 0.25 \times 2.1875 = 6.578125$$

$$S'_3 = -12.625 - 0.25 \times 2.1875 = -13.171875$$

$$S'_4 = 7.9375 - 0.25 \times 2.1875 = 7.390625$$

$$\widetilde{Y}_{am} = Y_{am} - S'_m$$

Par exemple :

$$\widetilde{Y}_{11} = 419 + 0.796875 = 419.798675$$

Le tableau complété :

Année	Trimestre	Série	MM(4)	Différence	Composante saisonnière	Ajustement	Série désaisonnalisée
(a)	(m)	(Y_{am})	(T_{am})	$(Y_{am} - T_{am})$	(S_m)	(S'_m)	$(Y_{am} - S'_m)$
1	1	419	_	_	−0.25	−0.7968	419.7968
1	2	417	_	_	7.125	6.5781	410.4218
1	3	403	417.875	−14.875	−12.625	−13.1718	416.1718
1	4	430	419.5	10.5	7.9835	7.3906	422.6093
2	1	424	421.25	2.75	_	_	424.7968
2	2	425	421.25	3.75	_	_	418.4218
2	3	409	419.325	−10.375	_	_	422.1718
2	4	424	418.625	5.375	_	_	416.6093
3	1	415	418.25	−3.25	_	_	415.7968
3	2	428	417.5	10.5	_	_	421.4218
3	3	403	_	_	_	_	416.1718
3	4	424	_	_	_	_	416.6093

Exercice 13.
On considère les données du Tableau 14.13 qui nous donne la population (en milliers d'habitants) des 27 pays de l'Union Européenne :

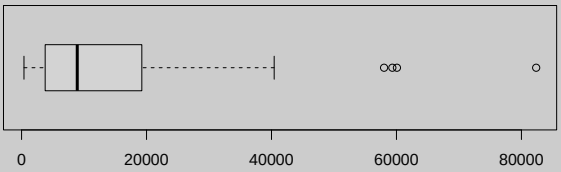
TABLE 14.13 – Population des pays de l'UE

Pays	Pop.	Pays	Pop.
Malte	384	Hongrie	9 973
Luxembourg	447	Tchéquie	10 274
Chypre	677	Belgique	10 292
Estonie	1 361	Portugal	10 303
Slovénie	1 995	Grèce	10 596
Lettonie	2 351	Pays-Bas	16 101
Lituanie	3 681	Roumanie	22 390
Irlande	3 873	Pologne	38 629
Finlande	5 195	Espagne	40 428
Danemark	5 367	Italie	58 018
Slovaquie	5 403	France	59 343
Bulgarie	8 107	Royaume-Uni	60 075
Autriche	8 140	Allemagne	82 360
Suède	8 910		

1. Calculez les premier, deuxième (la médiane) et troisième quartiles, ainsi que la distance interquartile.
2. Construisez la boite à moustaches (box-plot).

Solution

1. Le premier quartile : $np = \frac{1}{4} \times 27 = 6.75$ n'est pas un nombre entier, on arrondi vers le haut
$$x_{1/4} = x_{(7)} = 3681 \text{ (Lituanie)}.$$
La médiane : Comme $np = \frac{1}{2} \times 27 = 13.5$ n'est pas un nombre entier, on arrondi vers le haut
$$x_{1/2} = x_{(14)} = 8910 \text{ (Suède)}.$$
Le troisième quartile : Comme $np = \frac{3}{4} \times 27 = 20.25$ n'est pas un nombre entier, on arrondi vers le haut
$$x_{3/4} = x_{(21)} = 22390 \text{ (Roumanie)}.$$
IQR :
$$IQR = x_{3/4} - x_{1/4} = 22390 - 3681 = 18709.$$
2. $b^- = 3681 - 1.5 \times 18709 = -24382.5$ et $b^+ = 22390 + 1.5 \times 18709 = 50453.5$
valeurs adjacentes : 384 (Malte) et 40428 (Espagne).



En langage R

```
pop=c(384,447,677,1361,1995,2351,3681,3873,5195,5367,
5403,8107,8140,8910,9973,10274,10292,10303,10596,
16101,22390,38629,40428,58018,59343,60075,82360)
boxplot(pop1,horizontal=TRUE)
```

Exercice 14.
Pour s'approvisionner en fruits, une chaîne de grands-magasins fait acheminer des cargaisons de fruits par bateau, puis distribue, en camion, les cargaisons dans les différents magasins du pays. Pour 10 livraisons par camions de 1000 kilogrammes de fruits dans différentes villes du pays, le Tableau 14.14 contient le nombre de jours écoulés entre l'arrivée du chargement par bateau et l'arrivée des fruits dans le magasin (X) ainsi que la quantité de fruits en trop mauvais état pour être mis en vente (Y, en kilogrammes).

Table 14.14 – Données de la livraison de fruits

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	1	17	1	289	17
	0	9	0	81	0
	2	18	4	324	36
	1	15	1	225	15
	3	24	9	576	72
	0	10	0	100	0
	0	7	0	49	0
	1	14	1	196	14
	2	18	4	324	36
	0	12	0	144	0
Somme	10	144	20	2308	190
Moyenne	1	14.4	2	230.8	19

1. Calculez les variances marginales de X et Y.
2. Calculez la covariance entre X et Y.
3. Déterminez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez les coefficients de corrélation et de détermination entre les variables X et Y et ensuite déterminez la qualité de l'ajustement.
5. Donnez la valeur ajustée et le résidu pour la première observation du Tableau 14.14.

Solution

1. Variances marginales de X et Y :
$$\bar{x} = 1, \ \bar{y} = 14.4$$
$$s_x^2 = 2 - 1^2 = 2 - 1 = 1$$
$$s_y^2 = 230.8 - 14.4^2 = 230.8 - 207.36 = 23.44.$$
2. Covariance entre X et Y :
$$s_{xy} = 19 - 1 \times 14.4 = 19 - 14.4 = 4.6.$$
3. Équation de la droite de régression : Ajustement linéaire de y en x
$$D_{y|x} \ : y = a + b\ x, \text{ avec } b = \frac{s_{xy}}{s_x^2} = 4.6 \text{ et } a = \bar{y} - \frac{s_{xy}}{s_x^2} x = 14.4 - 4.6 \times 1 = 9.8$$
$$D_{y|x} \ : y = 4.6x + 9.8.$$
4. Coefficients de corrélation et de détermination et qualité de l'ajustement :
$$r_{xy} = \frac{4.6}{\sqrt{1 \times 23.44}} = 0.950.$$
$r^2 = 0.950^2 = 0.903 \Rightarrow$ très bon ajustement.
5. Donnez la valeur ajustée et le résidu pour la première observation du Tableau 14.2 :
$$y_1^* = 4.6 \times 1 + 9.8 = 14.4, \ e_1 = 17 - 14.4 = 2.6.$$

Exercice 15.
La série X_t du Tableau 14.15 représente les précipitations (en mm/an) relevées pendant 10 ans dans une région au climat aride. Complétez le tableau ci-dessus en appliquant un lissage exponentiel simple à la série X_t à l'aide

Table 14.15 – Précipitations (en mm/an) relevées pendant 10 ans

t	1	2	3	4	5	6	7	8	9	10
X_t	44	123	122	88	117	9995	124	134	97	

de la formule récursive suivante pour $\beta = 0.7$:

$$\widehat{X}_T(1) = (1-\beta)\sum_{j=0}^{T-1}\beta^j X_{T-j} = (1-\beta)X_T + \beta\widehat{X}_{T-1}(1),\ \widehat{X}_0(1) = X_1.$$

Solution

t	1	2	3	4	5	6	7	8	9	10
X_t	44	123	122	88	117	99	95	124	134	97
$\widehat{X}_t(1)$	44.000	67.700	83.990	85.193	94.735	96.015	95.710	104.197	113.138	108.297

Exercice 16.
À l'occasion de l'ouverture d'une nouvelle salle de cinéma, le propriétaire décide d'offrir des tickets de cinéma à toutes les personnes présentes le premier jour d'ouverture. Chaque spectateur reçoit une enveloppe dans laquelle il peut y avoir entre un et cinq tickets gratuits. On trouve, dans le Tableau 14.16, la fonction de répartition de X, la variable aléatoire représentant le nombre de tickets gratuits reçu par un individu présent ce jour-là.

Table 14.16 – Nombre de tickets gratuits reçu par un individu

x	1	2	3	4	5
$F(x)$	0.2	0.4	0.6	0.8	1.00

1. Trouvez la distribution de probabilité de X.
2. Calculez l'espérance de X.
3. Calculez la variance de X.
4. Un individu a réussi à tricher et à obtenir 3 enveloppes. Il ouvre la première enveloppe et obtient 2 tickets. Calculez l'espérance et la variance du nombre de ticket total obtenu par cet individu.

Solution

1. Distribution de probabilité de X :

x	1	2	3	4	5	Total
$\Pr(X = x)$	0.2	0.2	0.2	0.2	0.2	1
$x\Pr(X = x)$	0.2	0.4	0.6	0.8	0.1	3
$x^2\Pr(X = x)$	0.2	0.8	1.8	3.2	5.0	11

2. Espérance de X : $\mathrm{E}(X) = 3$.
3. Variance de X : $\mathrm{var}(X) = 11 - 3^3 = 2$.
4. Espérance et variance :
$$\mathrm{E}(2 + X_2 + X_3) = 2\mathrm{E}(X) + 2 = 2 \times 3 + 2 = 8$$
et
$$\mathrm{var}(2 + X_2 + X_3) = \mathrm{var}(X_2) + \mathrm{var}(X_3) = 4.$$

Exercice 17.
On s'intéresse à la variable statistique "Nombre de voitures par ménage". Le Tableau 14.17 donne les effectifs pour un échantillon de 60 ménages suisses.

TABLE 14.17 – Nombre de voitures par ménage

x_j	n_j	N_j	f_j	F_j
0	8			
1	22			
2	16			
3	11			
4	2			
5	1			

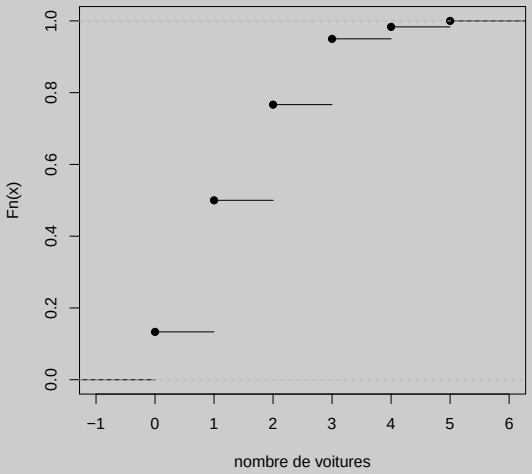
1. Compléter le tableau statistique ci-dessus.
2. Calculer le nombre moyen de voitures par ménage.
3. Calculer les quartiles.
4. La médiane et le mode sont-ils égaux ? Justifier.
5. Donner l'expression de la fonction de répartition.
6. Représenter graphiquement la fonction de répartition.

Solution

1. Tableau statistique :

x_j	n_j	N_j	f_j	F_j
0	8	8	0.133	0.133
1	22	30	0.367	0.5
2	16	46	0.267	0.767
3	11	57	0.183	0.95
4	2	59	0.033	0.983
5	1	60	0.017	1

2. Moyenne : $\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{1}{60}(8\times 0 + 22\times 1 + 2\times 16 + 3\times 11 + 4\times 2 + 5\times 1) = 1.667.$
3. Quartiles :
$$np = 60\frac{1}{4} = 15 \text{ donc } x_{\frac{1}{4}} = \frac{1}{2}\left(x_{(15)} + x_{(16)}\right) = 1,$$
$$np = 60\frac{1}{2} = 30 \text{ donc } x_{\frac{1}{2}} = \frac{1}{2}\left(x_{(30)} + x_{(31)}\right) = 1.5,$$
$$np = 60\frac{3}{4} = 45 \text{ donc } x_{\frac{3}{4}} = \frac{1}{2}\left(x_{(45)} + x_{(46)}\right) = 2.$$
4. Non car $x_{1/2} = 1.5$ et $x_M = 1$.
5. La fonction de répartition est donnée par :
$$F(x) = \begin{cases} 0 & x < 0 \\ 0.133 & 0 \leq x < 1 \\ 0.5 & 1 \leq x < 2 \\ 0.767 & 2 \leq x < 3 \\ 0.950 & 3 \leq x < 4 \\ 0.983 & 4 \leq x < 5 \\ 1 & 5 \leq x \end{cases}$$
6. On les points $(x_j; F_j)$ et segments de droites horizontales car la variable est discrète.



En langage R

```
X=c(rep(0,8),rep(1,22),rep(2,16),rep(3,11),rep(4,2),rep(5,1))
plot(ecdf(X),xlab="nombre de voitures",main="")
```

Exercice 18.
À l'aide de données sur 16 cantons suisses, on s'intéresse au lien entre le coût de la santé annuel par habitant (variable Y, en milliers de francs suisses) et la proportion d'individus agés de plus de 65 ans résidant dans le canton (X, en %). On souhaite effectuer un ajustement (régression) linéaire permettant d'expliquer le coût de la santé par la proportion de personnes agées. On a déjà les résultats suivants :

$$\sum_{i=1}^{16} x_i = 238.3, \quad \sum_{i=1}^{16} y_i = 28.6, \quad \sum_{i=1}^{16} x_i^2 = 3597.77, \quad \sum_{i=1}^{16} y_i^2 = 54.97, \quad r_{xy} = 0.653.$$

1. À l'aide de la méthode des moindres carrés, calculer les paramètres du modèle et donner la droite de régression.
2. Donner le coefficient de détermination r_{xy}^2 et interpréter la qualité d'ajustement de ce modèle.
3. Soit deux cantons, A et B. La proportion de personnes agées est élevée dans le canton B, alors qu'elle est basse dans le canton A. D'après le modèle, dans lequel des deux cantons le coût de la santé par habitant serait le plus élevé ? (Justifier)
4. Donner la valeur ajustée et le résidu pour le canton de Neuchâtel où la proportion de personnes de plus de 65 ans est de 17.2 % et où l'on a observé un coût de la santé de 2.8 millier de francs par habitant.
5. À votre avis, sur le nuage de point représentant ces deux variables, le point représentant le canton de Neuchâtel se situerait-il audessus ou en-dessous de la droite de régression ? Justifier.

Solution

1. Paramètres du modèle et droite de régression :

$$\bar{x} = \frac{238.3}{16} = 14.894,$$

$$\bar{y} = \frac{28.6}{16} = 1.788.$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{16} \times 3597.77 - 14.894^2 = 3.029,$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{16} \times 54.97 - 1.788^2 = 0.239,$$

$$s_{xy} = r_{xy}\sqrt{s_x^2 s_y^2} = 0.653\sqrt{3.029 \times 0.239} = 0.555,$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{0.555}{3.029} = 0.183,$$

$$a = \bar{y} - b\ \bar{x} 1.788 - 0.183 \times 14.894 = -0.938,$$

$$y = a + b\ x = -0.938 + 0.183\ x.$$

2. Coefficient de détermination et qualité d'ajustement de ce modèle :
$r_{xy}^2 = 0.653^2 = 0.426$ ajustement moyen (42.6% de la variance est expliquée).
3. Coût de la santé selon les cantons :
Canton B, car relation linéaire positive entre les deux variables.
4. Valeur ajustée et résidu pour le canton de Neuchâtel :
$y_{NE}^* = a + b\ x_{NE} = -0.938 + 0.183 \times 17.2 = 2.210,$
$e_{NE} = y_{NE} - y_{NE}^* = 2.8 - 2.210 = 0.59.$
5. Position du canton de Neuchâtel : au dessus car le résidu e_{NE} est positif.

Exercice 19.
Afin de sonder l'opinion publique concernant un candidat aux éléctions municipales, un institut de sondage a récolté l'avis de 300 personnes concernant ce candidat ainsi que leur zone de résidence (centre-ville ou périphérie). Le Tableau 14.18 résume ces données.

Table 14.18 – Tableau de contingence

	Avis concernant le candidat			
Zone de résidence	Positif	Indifférent	Négatif	Total
centre-ville	50	20	40	110
périphérie	10	70	110	190
Total	60	90	150	300

1. Quelle est la proportion d'individus que ce candidat laisse indifférent ?
2. Parmi les individus habitant au centre ville, quelle est la proportion de personnes qui émettent un avis positif concernant ce candidat ? Parmi les personnes émettant un avis négatif, quelle est la proportion d'habitants du centre ville ?
3. Calculer le V de Cramer et commenter le résultat obtenu.

Solution

1. Proportion d'individus que ce candidat laisse indifférent : $\frac{90}{300} = 0.3$.
2. Proportion de personnes qui émettent un avis positif : $\frac{50}{110} \approx 0.455$.
3. Proportion d'habitants du centre ville parmi les personnes émettant un avis négatif : $\frac{40}{150} \approx 0.267$.
4. Calcul du V de Cramer :

Tableau des effectifs théoriques

	Avis concernant le candidat			
Zone de résidence	Positif	Indifférent	Négatif	Total
centre-ville	22	33	55	110
périphérie	38	57	95	190
Total	60	90	150	300

Tableau des écarts à l'indépendance

	Avis concernant le candidat			
Zone de résidence	Positif	Indifférent	Négatif	Total
centre-ville	28	−13	−15	0
périphérie	−28	13	15	0
Total	0	0	0	0

$$\chi^2_{obs} = \frac{28^2}{22} + \frac{28^2}{38} + \frac{13^2}{33} + \frac{13^2}{57} + \frac{15^2}{55} + \frac{15^2}{95} = 70.813$$

$$\phi^2 = \frac{\chi^2_{obs}}{1000} = 0.109$$

$$V = \sqrt{\frac{70.813}{300 \times 1}} = 0.486$$

La dépendance entre la zone de résidence et l'avis concernant le candidat est relativement élevée.

Exercice 20.
Dans une région, les candidats au baccalauréat ont le choix entre deux filières : le baccalauréat littéraire et le baccalauréat scientifique. 40% des élèves sont inscrits dans la filière scientifique. Le taux de réussite est seulement de $\frac{2}{3}$ chez les littéraires, alors qu'il est de $\frac{3}{4}$ chez les scientifiques.

1. Après les epreuves de baccalauréat, on choisit un élève au hasard dans cette population. Calculer la probabilité que ce candidat ait réussit ses examens et qu'il vienne de la filière littéraire.
2. On choisit à nouveau un élève au hasard dans cette population. Calculer la probabilité que ce candidat ait réussit ses examens.
3. On a séléctionné un individu qui a obtenu son baccalauréat. Calculer la probabilité qu'il vienne de la filière scientifique.
4. On séléctionne successivement et avec remise 12 élèves issus de la filière littéraire. On note X, la variable aléatoire représentant le nombre d'élèves ayant obtenus leur baccalauréat. Calculer la probabilité qu'exactement 6 d'entre-eux aient réussit leurs examens de baccalauréat.
5. Calculer l'espérance et la variance de X.

Solution

1. On pose les événements suivants :
   R : l'élève a réussi les examens de baccalauréat.
   L : l'élève est inscrit dans la filière littéraire.
   S : l'élève est inscrit dans la filière scientifique.
   On a donc :
   $$\Pr(L) = 0.6, \Pr(S) = 0.4,\ \Pr(R|L) = \frac{2}{3},\ \Pr(R|S) = \frac{3}{4}.$$
   $$\Pr(R \cap L) = \Pr(L) \times \Pr(R|L) = 0.6 \times \frac{2}{3} = 0.4.$$
2. $\Pr(R) = \Pr(L) \times \Pr(R|L) + \Pr(S) \times \Pr(R|S) = 0.6 \times \dfrac{2}{3} + 0.4 \times \dfrac{3}{4} = 0.7.$
3. $\Pr(S|R) = \dfrac{\Pr(S) \times \Pr(R|S)}{\Pr(S) \times \Pr(R|S) + \Pr(L) \times \Pr(R|L)} = \dfrac{0.4 \times \frac{3}{4}}{0.7} = \dfrac{3}{7} \approx 0.429.$
4. Comme $X \sim \mathcal{B}\left(n = 12, p = \dfrac{2}{3}\right)$.
   $$\Pr(X = 6) = \binom{12}{6} \times \left(\frac{2}{3}\right)^6 \times \left(\frac{1}{3}\right)^6 = 0.111.$$
5. $\mathrm{E}(X) = np = 12 \times \frac{2}{3} = 8$ et $\mathrm{var}(X) = np(1-p) = 12 \times \dfrac{2}{3} \times \dfrac{1}{3} = \dfrac{8}{3} \approx 2.667.$

Exercice 21.
On dispose des résultats d'une enquête concernant l'âge de 2000 habitants d'une commune suisse. La distribution des classes d'âge est donnée dans le Tableau 14.19 :

TABLE 14.19 – Distribution en classes d'âge

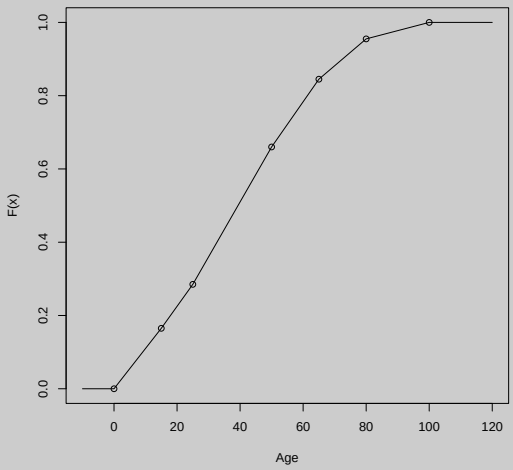
Âge	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+[$	n_j	N_j	f_j	F_j
$[0, 15[$	330			
$[15, 25[$	240			
$[25, 50[$	750			
$[50, 65[$	370			
$[65, 80[$	220			
$[80, 100]$	90			
Total	2000	–	1.000	–

1. Complétez le Tableau 14.19.
2. Dessinez la fonction de répartition dans le repère ci-dessous :

Solution
Tableau complété :

Âge	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+[$	n_j	N_j	f_j	F_j
$[0, 15[$	330	330	0.165	0.165
$[15, 25[$	240	570	0.120	0.285
$[25, 50[$	750	1320	0.375	0.660
$[50, 65[$	370	1690	0.185	0.845
$[65, 80[$	220	1910	0.110	0.955
$[80, 100]$	90	2000	0.045	1.000
Total	2000	–	1.000	–

Fonction de répartition : on représente les couples $(c_j^+; F_j)$ et on relie les points linéairement en partant de l'origine.



En langage R

```
X=rbind(c(-10,0),c(0,0),c(15,0.165),c(25,0.285),c(50,0.660),
c(65,0.845),c(80,0.955),c(100,1.000),c(120,1.000))
plot(X,type="l",xlab="Age",ylab="F(x)",main="")
points(X[2:8,])
```

Exercice 22.
Le Tableau 14.20 nous donne, pour 10 pays européens, la densité de population (variable X, en nombre d'habitants par km^2) ainsi que la part d'énergies renouvelables dans la production totale d'énergie du pays (variable Y, en pourcentage). On s'intéresse au lien éventuel entre ces deux variables. On connait également les

Table 14.20 – Données par Pays

Pays	x_i	y_i
Belgique	343.6	11.8
Danemark	125.5	9.3
Allemagne	231.2	12.4
Espagne	83.0	28.9
France	97.9	12.3
Italie	197.1	41.7
Hongrie	108.6	11.4
Royaume-Uni	246.9	1.7
Suède	21.9	44.7
Finlande	17.3	49.8
Somme	1473	224
Moyenne	147.3	22.4

informations suivantes :

$$\sum_{i=1}^{10} x_i^2 = 316118.9 \qquad \sum_{i=1}^{10} y_i^2 = 7715.86 \qquad \sum_{i=1}^{10} x_i y_i = 23408.69$$

1. Calculez les variances marginales de X et Y.
2. Calculez la covariance entre X et Y.
3. Déterminez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez les coefficients de corrélation et de détermination entre les variables X et Y et ensuite déterminez la qualité de l'ajustement.
5. Donnez la valeur ajustée et le résidu pour l'Espagne.

Solution

1. Variances marginales de X et Y : $s_x^2 = 11005.21$ et $s_y^2 = 299.507$.
2. Covariance entre X et Y : $s_{xy} = -1064.103$.
3. Équation de la droite de régression de Y en fonction de X.
   $Y = a + b\ x = 36.643 - 0.097x$,
   $b = \dfrac{s_{xy}}{s_x^2} = -0.097$,
   $a = \bar{y} - b\bar{x} = 36.643$.
4. Coefficients de corrélation et de détermination entre les variables X et Y et qualité de l'ajustement.
   $r_{xy} = \frac{s_{xy}}{s_x s_y} = -0.586$,
   $r_{xy}^2 = 0.344$.
   Qualité d'ajustement : 34,4% de la variance expliquée (médiocre).
5. Valeur ajustée et résidu pour l'Espagne.
   $y_4^* = 36.643 - 0.097 \times 83 = 28.592$,
   $e_4 = y_4 - y_4^* = 28.9 - 28.592 = 0.308$.

Exercice 23.
Lors d'une élection pour le poste de maire d'une ville, deux candidats se présentent. On aimerait savoir s'il existe une relation entre le niveau d'éducation des électeurs et le choix du candidat. Le Tableau 14.21 représente les résultats obtenus lors d'une étude réalisée sur 1000 personnes.

Table 14.21 – Tableau de contingence

	Candidat préféré		
Niveau d'éduc.	Candidat A	Candidat B	Total
primaire	220	130	350
secondaire	190	210	400
universitaire	120	130	250
Total	530	470	1000

Calculez le V de Cramer et commentez le résultat obtenu.

Solution

Tableau des effectifs théoriques

	Candidat préféré		
Niveau d'éduc.	Candidat A	Candidat B	Total
primaire	185.5	164.5	350
secondaire	212	188	400
universitaire	132.5	117.5	250
Total	530	470	1000

Tableau des écarts à l'indépendance

	Candidat préféré		
Niveau d'éduc.	Candidat A	Candidat B	Total
primaire	34.5	−34.5	0
secondaire	−22	22	0
universitaire	−12.5	12.5	0
Total	0	0	0

Tableau des $\frac{e_{jk}^2}{n_{jk}^*}$

	Candidat préféré		
Niveau d'éduc.	Candidat A	Candidat B	Total
primaire	6.416	7.236	13.652
secondaire	2.283	2.574	4.857
universitaire	1.179	1.330	2.509
Total	9.878	11.140	21.018

$$\chi^2_{obs} = 21.018,$$

$$\phi^2 = \frac{\chi^2_{obs}}{1000} = 21.018/1000 = 0.021018,$$

$$V = \frac{0.021018}{1} = 0.021018.$$

La dépendance entre le niveau d'éducation et le choix du candidat est très faible.

Exercice 24.
On trouve dans le Tableau 14.22 la distribution de probabilité d'un jeu de hasard. La variable aléatoire X représente le score (nombre de points) obtenu en une partie. À chaque partie, il est possible de gagner 2 ou 10 points (valeurs positives) ou d'en perdre 1, 2 ou 5 (valeurs négatives).

Table 14.22 – Distribution de probabilité de X

x	-5	-2	-1	2	10
$p_X(x)$	0.1	0.2	0.25	0.4	0.05

1. Calculez l'espérance de X.
2. Calculez la variance de X.
3. Quelle est la probabilité d'obtenir un score positif en jouant une fois à ce jeu ?
4. Je décide de jouer 10 parties d'affilée. Quelle est la probabilité que je réalise un score positif lors de exactement 6 de mes 10 parties ?

Solution

1. $\mathrm{E}(X) = 0.1 \times (-5) + 0.2 \times (-2) + 0.25 \times (-1) + 0.4 \times 2 + 0.05 \times 10 = 0.15$.
2. $\mathrm{var}(X) = 10.1275$.
3. $\Pr(X = 2) + \Pr(X = 10) = 0.45$.
4. Binomiale avec $n = 10$ et $p = 0.45$: $\Pr(k = 6) = \binom{10}{6} \times 0.45^6 \times 0.55^4 = 0.160$.

Exercice 25.
Le Tableau 14.23 contient la population (en milliers d'individus) des 20 régions de France de plus d'un million d'habitants (données du recensement de 1999).

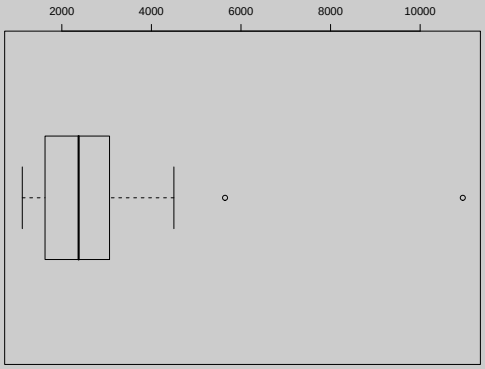
TABLE 14.23 – Population (en milliers d'individus) des 20 régions de France

Région	Population	Région	Population
Franche-Comté	1 117	Centre	2 440
Auvergne	1 309	Languedoc-Roussillon	2 496
Champagne-Ardenne	1 342	Midi-Pyrénées	2 552
Basse-Normandie	1 422	Bretagne	2 906
Bourgogne	1 610	Aquitaine	2 908
Poitou-Charentes	1 640	Pays de la Loire	3 222
Haute-Normandie	1 780	Nord-Pas-de-Calais	3 997
Alsace	1 794	Provence-Alpes-Côte d'Azur	4 502
Picardie	1 857	Rhône-Alpes	5 645
Lorraine	2 310	Île-de-France	10 952

1. Calculer les premier, deuxième (la médiane) et troisième quartiles, ainsi que la distance interquartile.
2. Donner le coefficient d'asymétrie de Yule et commenter l'asymétrie de la série.
3. Dessiner verticalement la boîte à moustache (boxplot) de cette série en prenant soin d'y annoter les éléments qui permettent de la construire.

Solution

1. Quartiles :
   $np = 20\frac{1}{4} = 5$ donc $x_{\frac{1}{4}} = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = 1625$,
   $np = 20\frac{1}{2} = 10$ donc $x_{\frac{1}{2}} = \frac{1}{2}\left(x_{(10)} + x_{(11)}\right) = 2375$,
   $np = 20\frac{3}{4} = 15$ donc $x_{\frac{3}{4}} = \frac{1}{2}\left(x_{(15)} + x_{(16)}\right) = 3065$.
   $IQ = x_{\frac{3}{4}} - x_{\frac{1}{4}} = 3065 - 1625 = 1440$
2. Asymétrie : $A_Y = \dfrac{x_{\frac{3}{4}} + x_{\frac{1}{4}} - 2x_{\frac{1}{2}}}{x_{\frac{3}{4}} - x_{\frac{1}{4}}} = \dfrac{3065 + 1625 - 2 \times 2375}{3065 - 1625} = -0.042.$
   (faible asymétrie à gauche).
3. $b^- = x_{\frac{1}{4}} - 1.5IQ = 1625 - 1.5 \times 1440 = -535$
   $b^+ = x_{\frac{3}{4}} + 1.5IQ = 3065 + 1.5 \times 1440 = 5225$
   Valeurs adjacentes : 1117 et 4502
   Valeurs extrêmes : 5645 et 10952.
   Boxplot :



Exercice 26.
On s'intéresse à la relation entre le prix d'une voiture (variable Y, en milliers de dollars) et son poids (variable X, en centaines de livres). Dans ce but, on a relevé le prix et le poids des 50 modèles de voitures les plus vendus actuellement. Un aperçu de ces données se trouve dans le Tableau 14.24.

Table 14.24 – Prix et poids des 50 modèles de voitures

voiture i	poids x_i	prix y_i	x_i^2	y_i^2	$x_i y_i$
1	28.66	26.250	821.396	689.063	752.325
2	28.88	15.969	834.054	255.009	461.185
⋮	⋮	⋮	⋮	⋮	⋮
49	32.30	15.604	1043.290	243.485	504.009
50	33.73	19.844	1137.713	393.784	669.338
Total	1440	910	45975	23721	30363
Moyenne	28.8	18.2	919.50	474.42	607.26

1. Calculer les variances marginales de X et Y.
2. Calculer la covariance entre X et Y.
3. Calculer les coefficients de régression b et a, puis déterminer l'équation de la droite de régression de Y en fonction de X.
4. Déterminer le coefficient de corrélation entre les variables X et Y et commenter la force et le sens de la relation.
5. Une des voitures considérées dans cette enquête (la voiture numérotée $i = 12$) pèse 35 centaines de livres. À l'aide de l'ajustement linéaire effectué ci-dessus, donner une estimation de son prix.
6. Le prix réel de cette voiture est 22 milliers de dollars. Expliquer en trois lignes maximum pourquoi ce prix diffère de celui calculé à la question précédente (y_{12}^*).
7. Calculer le résidu pour cette observation.

Solution

1. Variances :

$$s_x^2 = \frac{1}{50}\sum_{i=1}^{50} x_i^2 - \bar{x}^2 = 919.5 - 28.8^2 = 90.06,$$
$$s_y^2 = \frac{1}{50}\sum_{i=1}^{50} y_i^2 - \bar{y}^2 = 474.42 - 18.2^2 = 143.18.$$

2. Covariance :

$$s_{xy} = \frac{1}{50}\sum_{i=1}^{50} x_i y_i - \bar{x}\ \bar{y} = 607.26 - 28.8 \times 18.2 = 83.1.$$

3. Droite de régression :

$$b = \frac{s_{xy}}{s_x^2} = \frac{83.1}{90.06} \approx 0.923,$$
$$a = \bar{y} - b\ \bar{x} = 18.2 - 0.923 \times 28.8 \approx -8.382,$$
$$y = a + b\ x = -8.382 + 0.923x.$$

4. Corrélation :

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{83.1}{\sqrt{90.06}\sqrt{143.18}} = 0.732.$$

Relation positive relativement forte.

5. Prédiction : $y_{12}^* = -8.382 + 0.923 \times 35 = 23.923$.

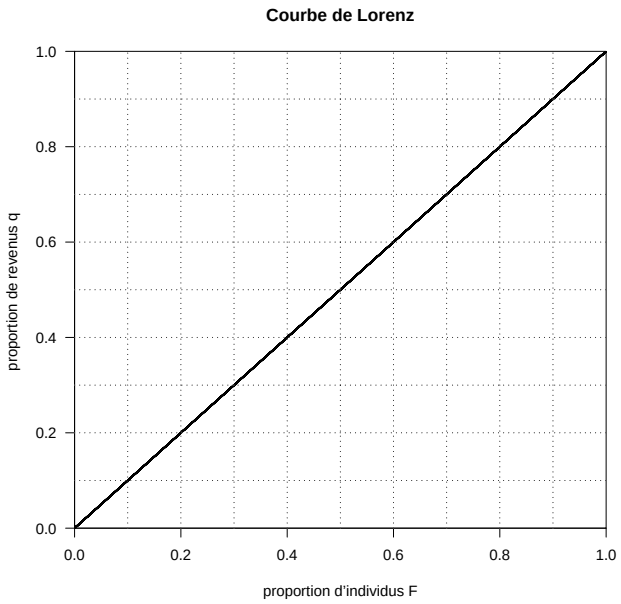
6. Le prix réel de la voiture diffère de cette estimation car cette dernière est estimée par le modèle. Elle représente le prix moyen d'une voiture pesant 3500 pounds, ce qui ne veut en aucun cas dire que toutes les voitures pesant 3500 pounds coûtent 23923 dollars.
7. Résidu : $e_{12} = y_{12} - y^*_{12} = 22 - 23.923 = -1.923$.

Exercice 27.
Dans une petite entreprise, on connait le revenu mensuel (X) de chacun des 5 salariés. Les données sont répertoriées dans le Tableau 14.25.

TABLE 14.25 – Revenu mensuel (X) de chacun des 5 salariés

salarié i	revenu x_i	effectif n_i	fréquence f_i	fréquence cumulée F_i	revenu cumulé $\sum_{j=1}^{i} x_{(j)}$	proportion de revenus q_i
1	2000	1	0.2			
2	3000	1	0.2			
3	5000	1	0.2			
4	7000	1	0.2			
5	12000	1	0.2	1	29000	1
Total	29000	5	1	—	—	—

1. Compléter toutes les cases vides du tableau.
2. Tracer la courbe de Lorenz de cette série dans le repère ci-dessous.



3. Calculer l'indice de Gini pour cette série.

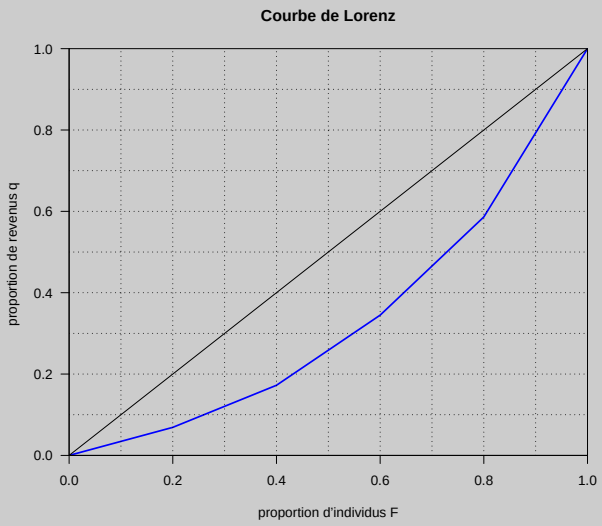
Solution

1. Tableau complété :

salarié i	revenu x_i	effectif n_i	fréquence f_i	fréquence cumulée F_i	revenu cumulé $\sum_{j=1}^{i} x_{(j)}$	proportion de revenus q_i
1	2000	1	0.2	0.2	2000	0.069
2	3000	1	0.2	0.4	5000	0.172
3	5000	1	0.2	0.6	10000	0.345
4	7000	1	0.2	0.8	17000	0.586
5	12000	1	0.2	1	29000	1
Total	29000	5	1	—	—	—

$$q_i = \frac{\sum_{j=1}^{i} x_{(j)}}{29000} \Rightarrow q_1 = \frac{2000}{29000} = 0.069$$

2. Courbe de Lorenz

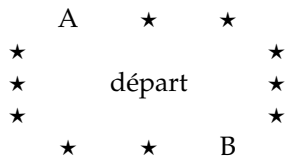


3. Indice de Gini :

$$G = \frac{1}{n-1}\left[\frac{2\sum_{i=1}^{n} i x_{(i)}}{n\bar{x}} - (n+1)\right].$$

$$G = \frac{1}{4}\left[\frac{2(1\times 2000 + 2\times 3000 + 3\times 5000 + 4\times 7000 + 5\times 12000)}{29000} - 6\right] = 0.414$$

Exercice 28.
Lors d'une expérience sur le comportement des rats, on place un rat au point central du labyrinthe représenté ci-dessous : il y avance de manière aléatoire et sans aucune préférence (chaque alternative a la même probabilité d'être choisie par le rat), mais il ne revient jamais sur ses pas. L'expérience prend fin lorsque le rat est arrivé dans un cul-de-sac (⋆) ou lorsqu'il a atteint une sortie (A ou B). S'il atteint la sortie A, on lui donne deux récompenses ; s'il atteint la sortie B, on lui donne une récompense ; s'il atteint un cul-de-sac, on ne lui donne pas de récompense. Soit alors X la variable aléatoire "Nombre de récompenses données au rat".



1. Remplir le Tableau 14.26 de la distribution de probabilité et de la fonction de répartition de la variable aléatoire X.

Table 14.26 – Distribution de probabilité de la récompense

x	0	1	2
$p_X(x)$	–	$\frac{1}{12}$	–
$F_X(x)$	–	–	–

2. Donner l'espérance et la variance de la variable aléatoire.
3. On renouvelle cette expérience 10 fois avec chaque fois un rat différent, c'est-à-dire de manière indépendante. Soit Y la variable aléatoire donnant le nombre de rats ayant atteint une sortie à la fin de l'expérience.
   (a) Quelle est la nature de la variable aléatoire Y ?
   Aide : on aimerait savoir s'il s'agit d'une variable indicatrice (ou bernoullienne), d'une variable binomiale, d'une variable de Poisson,...
   (b) Quelle est la probabilité qu'au moins un rat parvienne à une sortie ?
   (c) Quelle est la probabilité qu'au moins deux rats parviennent à une sortie ?

Solution

1. Tableau complété :

x	0	1	2
$p_X(x)$	$\frac{10}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
$F_X(x)$	$\frac{10}{12}$	$\frac{11}{12}$	1

2. Espérance et variance :

$$\begin{aligned}
\mathrm{E}(X) &= \sum_x x p_X(x) = 0 \times \frac{10}{12} + 1 \times \frac{1}{12} + 2 \times \frac{1}{12} = \frac{3}{12} = \frac{1}{4} = 0.25, \\
\mathrm{var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \sum_x x^2 p_X(x) - \left(\frac{1}{4}\right)^2 \\
&= 0^2 \times \frac{5}{6} + 1^2 \frac{1}{12} + 2^2 \frac{1}{12} - \frac{1}{16} = \frac{1}{12} + \frac{4}{12} - \frac{1}{16} = \frac{17}{48} \approx 0.354.
\end{aligned}$$

3. (a) Y est une variable aléatoire binomiale.

$$Y \sim \mathcal{B}\left(10, \frac{1}{6}\right).$$

(b) Probabilité qu'au moins un rat parvienne à une sortie :

$$\begin{aligned}
\Pr(Y \geq 1) &= 1 - \Pr(Y = 0), \\
\Pr(Y = 0) &= \binom{10}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{10} = \left(\frac{5}{6}\right)^{10} = 0.162, \\
\Pr(Y \geq 1) &= 1 - \Pr(Y = 0) = 1 - 0.162 = 0.838.
\end{aligned}$$

(c) Probabilité qu'au moins deux rats parviennent à une sortie :

$$\begin{aligned}
\Pr(Y \geq 2) &= 1 - \Pr(Y = 0) - \Pr(Y = 1), \\
\Pr(Y = 1) &= \binom{10}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^9 = 10 \times \frac{1}{6}\left(\frac{5}{6}\right)^9 = 0.323, \\
\Pr(Y \geq 2) &= 1 - \Pr(Y = 0) - \Pr(Y = 1) = 1 - 0.162 - 0.323 = 0.515.
\end{aligned}$$

Exercice 29.
On considère les données du tableau suivant contenant le loyer moyen, en francs suisses, par m^2 pour un studio dans les 26 cantons de Suisse. Les données datent de l'année 2000 (source : Office fédéral de la statistique) :

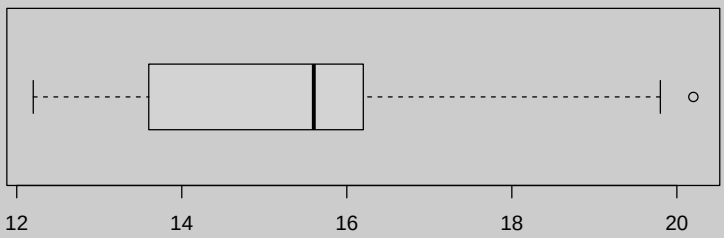
TABLE 14.27 – Loyer moyen selon le canton par m^2 pour un studio, en francs suisses, en 2000

Canton	Loyer	Canton	Loyer
Jura	12.2	Appenzell Rh.-I.	15.6
Appenzell Rh.-E.	12.9	Argovie	15.8
Neuchâtel	13.3	Schwytz	15.9
Uri	13.5	Lucerne	16
Glaris	13.5	Fribourg	16
Thurgovie	13.5	Berne	16.1
Valais	13.6	Nidwald	16.2
Soleure	14	Bâle-Ville	16.4
Schaffhouse	14.1	Vaud	16.8
Tessin	15.2	Genève	17.5
Obwald	15.3	Grisons	18.6
Saint-Gall	15.3	Zoug	19.8
Bâle-Campagne	15.6	Zurich	20.2

1. Calculez les premier, deuxième (la médiane) et troisième quartiles, ainsi que la distance interquartile.
2. Construisez la boite à moustaches (boxplot) horizontalement.

Solution

1. Quartiles :
   $np = 26\frac{1}{4} = 6.5$ donc $x_{\frac{1}{4}} = x_{(\lceil 6.5 \rceil)} = x_{(7)} = 13.6$,
   $np = 26\frac{1}{2} = 13$ donc $\frac{1}{2}\left(x_{(13)} + x_{(14)}\right) = \frac{1}{2}(15.6 + 15.6) = 15.6$,
   $np = 26\frac{3}{4} = 19.5$ donc $x_{\frac{3}{4}} = x_{(\lceil 19.5 \rceil)} = x_{(20)} = 16.2$,
   $IQ = x_{\frac{3}{4}} - x_{\frac{1}{4}} = 16.2 - 13.6 = 2.6$.
2. Calcul des valeurs adjacentes :
   $b^- = x_{\frac{1}{4}} - 1.5IQ = 13.6 - 1.58 \times 2.6 = 9.7$,
   $b^+ = x_{\frac{3}{4}} + 1.5IQ = 16.2 + 1.5 \times 2.6 = 20.1$.
   Valeurs adjacentes : 12.2 et 19.8.
   Valeur extrême : 20.2 (Zurich).



En langage R

```
X=c(12.2,12.9,13.3,13.5,13.5,13.5,13.6,14,14.1,15.2,15.3,15.3,15.6,
15.6,15.8,15.9,16,16,16.1,16.2,16.4,16.8,17.5,18.6,19.8,20.2)
boxplot(X,horizontal = TRUE)
```

Exercice 30.
On dispose des prix trimestriels d'un titre pendant les années 2006 à 2008. Dans la suite de l'exercice, X et Y représenteront respectivement le temps (en trimestre) et le prix du titre (en CHF) en question. Ces données se trouvent dans le Tableau 14.28. On souhaite effectuer un ajustement (régression) linéaire.
On a déjà les résultats suivants :

Table 14.28 – Prix trimestriels d'un titre pendant les années 2006 à 2008

X (temps)	1	2	3	4	5	6	7	8	9	10	11	12
Y (prix)	128	130	140	112	100	110	120	116	80	90	100	80

$$\sum_{i=1}^{12} x_i = 78, \qquad \sum_{i=1}^{12} y_i = 1306, \qquad \sum_{i=1}^{12} x_i^2 = 650, \qquad \sum_{i=1}^{12} y_i^2 = 146284, \qquad r_{xy} = -0.811.$$

1. Calculer les variances marginales de X et Y.
2. Calculer la covariance entre X et Y.
3. (a) Déterminer l'équation de la droite de régression de Y en fonction de X.
   (b) Expliquer en trois lignes au maximum pourquoi on fait ici une régression de Y en X plutôt qu'une régression de X en Y.
4. Déterminer le coefficients de détermination entre les variables X et Y et ensuite déterminer la qualité de l'ajustement.
5. Donner la valeur ajustée et le résidu de la dernière observation.
6. Estimer à l'aide du modèle le prix du titre le premier trimestre de 2009 (c'est-à-dire pour $t = 13$).

Solution

1. $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{12}650 - \left(\frac{78}{12}\right)^2 = 11.917,$
   $s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{12}146284 - \left(\frac{1306}{12}\right)^2 = 345.639.$
2. $s_{xy} = r_{xy}s_x s_y = -0.811\sqrt{11.917}\sqrt{345.639} = -52.049.$
3. (a) $b = \frac{s_{xy}}{s_x^2} = \frac{-52.049}{11.917} = -4.368,$
   $a = \bar{y} - b\,\bar{x} = \frac{1306}{12} + 4.368\frac{78}{12} = 137.225,$
   $Y = a + b\,x = 137.225 - 4.368\,x.$
   (b) Cela a un sens de supposer que le prix peut être une fonction du temps mais cela n'a pas de sens de supposer que le prix peut expliquer le temps.
4. $r_{xy}^2 = 0.658$ donc 65.8% de la variance est expliquée par le modèle.
5. $y_{12}^* = a + b\,x_{12} = 137.225 - 4.368 \times 12 = 84.809,$
   $e_{12} = y_{12} - y_{12}^* = 80 - 84.809 = -4.809,$
6. $y_{13}^* = a + 13b = 137.225 - 13 \times 4.368 = 80.441.$

Exercice 31.
On souhaite estimer la tendance par une moyenne mobile d'ordre 4 (dont on a déjà calculé quelques valeurs) puis désaisonnaliser la série par la méthode additive. On rappelle que :

$$MM(4)y_t = \frac{1}{8}\left(y_{t-2} + 2y_{t-1} + 2y_t + 2y_{t+1} + y_{t+2}\right)$$

Compléter toutes les cases vides du Tableau r̃efestimt.

Table 14.29 – Estimation de la tendance par une moyenne mobile

Année a	Trimestre t	Prix Y_{at}	$MM(4)$ T_{at}	S_t	S'_t	Désaisonnalisation $Y_{at} - S'_t$
1	1	128	—			
1	2	130	—			
1	3	140	124			
1	4	112	118			
2	1	100	113	—	—	
2	2	110	111	—	—	
2	3	120		—	—	
2	4	116		—	—	
3	1	80		—	—	
3	2	90		—	—	
3	3	100	—	—	—	
3	4	80	—	—	—	

Solution
On calcule les moyennes mobiles manquantes :

$$T_{23} = MM(4)Y_{23} = 1/8 \times (100 + 2 \times 110 + 2 \times 120 + 2 \times 116 + 80) = 109,$$

$$T_{24} = MM(4)Y_{24} = 1/8 \times (110 + 2 \times 120 + 2 \times 116 + 2 \times 80 + 90) = 104,$$

$$T_{11} = MM(4)Y_{11} = 1/8 \times (120 + 2 \times 116 + 2 \times 80 + 2 \times 90 + 100) = 99,$$

$$T_{12} = MM(4)Y_{12} = 1/8 \times (116 + 2 \times 80 + 2 \times 90 + 2 \times 100 + 80) = 92.$$

Ensuite, on calcule les différences. Par exemple :

$$Y_{21} - T_{21} = 100 - 113 = -13,$$

$$Y_{31} - T_{31} = 80 - 99 = -19 = -13.$$

On déduit les composantes saisonnièrem. Par exemple :

$$S_1 = \frac{1}{2}(-13 - 19) = -16$$

Les composantes saisonnières sont centrés :

$$\frac{1}{4}(S_1 + S_2 + S_3 + S_4) = -0.25,$$

$$S'_1 = -16 + 0.25 = -15.75.$$

On peut enfin désaisonnaliser :

$$Y_{11} - S'_1 = 128 + 15.75 = 143.75.$$

Année	Trimestre	Prix	$MM(4)$			Désaisonnalisation	
a	t	Y_{at}	T_{at}	$Y_{at} - T_{at}$	S_t	S'_t	$Y_{at} - S'_t$
1	1	128	—	—	*-16*	*-15.75*	*143.75*
1	2	130	—	—	*-1.5*	*-1.25*	*131.25*
1	3	140	124	*16*	*13.5*	*13.75*	*126.25*
1	4	112	118	*-6*	*3*	*3.25*	*108.75*
2	1	100	113	*-13*	—	—	*115.75*
2	2	110	111	*-1*	—	—	*111.25*
2	3	120	*109*	*1*	—	—	*106.25*
2	4	116	*104*	*2*	—	—	*112.75*
3	1	80	*99*	*-19*	—	—	*95.75*
3	2	90	*92*	*-2*	—	—	*91.25*
3	3	100	—	—	—	—	*86.25*
3	4	80	—	—	—	—	*76.75*

Exercice 32.
Attention : Les parties 1 et 2 de cet exercice sont indépendantes.
On se trouve à un péage routier entre Reims et Metz.

1. Soit X, la variable aléatoire représentant le nombre de véhicules arrivant au péage par minute. On suppose que X suit une loi de Poisson. De plus, on sait que le nombre moyen de véhicules arrivant au péage chaque minute est de 5.
   (a) Quelle est la probabilité qu'au moins 2 voitures arrivent au péage durant une minute ?
   (b) Quelle est la variance de X ?
2. Un véhicule arrivant au péage peut venir soit de Reims, soit de Metz, mais d'aucune autre direction. En moyenne, une voiture sur cinq vient de Metz.
   (a) Quelle est la probabilité qu'un véhicule arrivant au péage ne vienne pas de Metz ?
   (b) Les voitures arrivent de manière indépendante au péage. Cinq voitures arrivent successivement au péage.
      i. Quelle est la probabilité qu'aucune d'entre elles ne vienne de Metz ?
      ii. Quelle est la probabilité qu'exactement trois d'entre elles arrivent de Metz ?

Solution

1. On a ici $X \sim \mathcal{P}(5)$.
   (a) $\Pr(X \geq 2) = 1 - \Pr(X = 0) - \Pr(x = 1) = 1 - e^{-5} - 5e^{-5} = 0.960$.
   (b) $\text{var}(X) = \lambda = 5$.
2. $Y \sim \mathcal{B}(5, 1/5)$
   (a) $1 - 1/5 = 4/5$
   (b) i. $\Pr(Y = 0) = \left(\frac{4}{5}\right)^5 = 0.328$
      ii. $\Pr(Y = 3) = \frac{5!}{2!3!}\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^2 = 0.0512$

Exercice 33.
Les données suivantes représentent la longueur du lancer (en mètres) par chacun des 39 participants à un concours de lancer de javelot :

Table 14.30 – Longueur du lancer au concours de lancer de javelot

40.2	40.3	42.1	42.5	42.5	42.8	42.8	42.9	43.0	43.2	46.8	47.0	47.7
48.1	48.3	49.3	49.8	50.2	50.2	50.4	50.6	50.8	51.0	51.4	51.4	51.7
51.8	53.1	53.2	53.7	53.7	54.2	55.1	55.5	56.0	56.4	57.4	57.8	63.3

1. À quel type de variable appartient la variable *longueur du lancer* ?
2. Calculer les premier, deuxième (la médiane) et troisième quartiles de cette série statistique.
3. Calculer le coefficient de Yule et commenter l'asymmétrie de la distribution.
4. Compléter le Tableau 14.31 présentant un regroupement en classes.

Table 14.31 – Longueur du lancer au concours de lancer de javelot

Longueur	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[40, 44[$				
$[44, 48[$				
$[48, 52[$				
$[52, 56[$				
$[56, 64[$				
Total	39	–	1.00	–

5. Construire un histogramme des fréquences suite au regroupement en classes.

Solution

1. La variable est quantitative continue.
2. Calcul des quartiles :

$$np = 39\frac{1}{4} = 9.75 \text{ donc } x_{\frac{1}{4}} = x_{(\lceil 9.75 \rceil)} = x_{(10)} = 43.2,$$

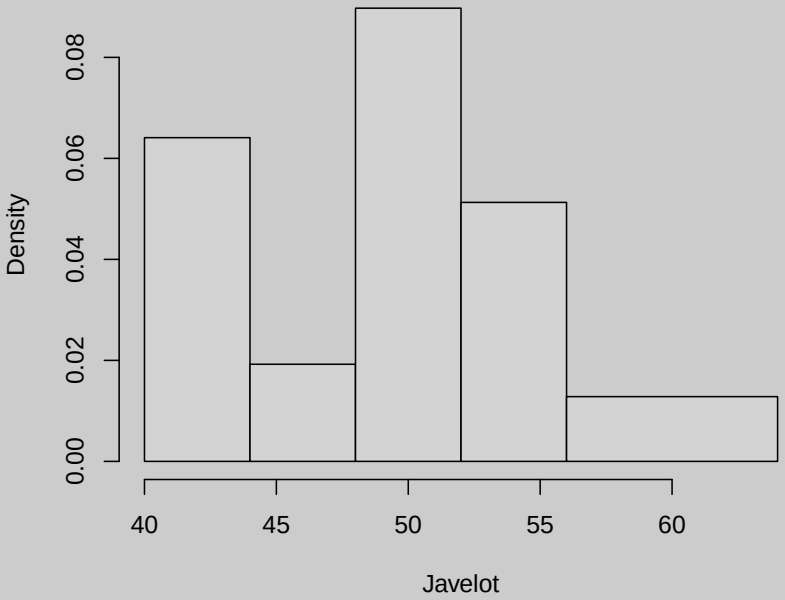
$$np = 39\frac{1}{2} = 19.5 \text{ donc } x_{(20)} = 50.4,$$

$$np = 39\frac{3}{4} = 29.25 \text{ donc } x_{\frac{3}{4}} = x_{(\lceil 29.25 \rceil)} = x_{(30)} = 53.7.$$

3. $A_y = \dfrac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}} = -0.371$ distribution allongée à gauche.
4. Tableau complété :

Longueur	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées	Hauteur
$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j	h_j
$[40, 44[$	10	10	0.256	0.256	0.0641
$[44, 48[$	3	13	0.077	0.333	0.0064
$[48, 52[$	14	27	0.359	0.692	0.0897
$[52, 56[$	7	34	0.179	0.872	0.0449
$[56, 64[$	5	39	0.128	1	0.0160
Total	39	–	1.00	–	

5. Histogramme. Calcul des $h_j = f_j/(c_j^+ - c_j^-)$.



En langage R

```
X=c(40.2,40.3,42.1,42.5,42.5,42.8,42.8,42.9,43.0,43.2,46.8,47.0,47.7,
48.1,48.3,49.3,49.8,50.2,50.2,50.4,50.6,50.8,51.0,51.4,51.4,51.7,
51.8,53.1,53.2,53.7,53.7,54.2,55.1,55.5,56.0,56.4,57.4,57.8,63.3)
bounds=c(40,44,48,52,56,64)
plot(hist(X,breaks=bounds),main="",xlab="Javelot")
```

Exercice 34.
Dans le Tableau reftauxch, on dispose de données concernant le taux de chômage (X, en %) de huit pays d'Europe ainsi que le prix moyen des transactions immobilières (Y, en milliers d'Euros) dans la capitale de chaque pays. On souhaite effectuer un ajustement (régression) linéaire permettant d'expliquer le prix de ces transactions par le taux de chômage. On a déjà les résultats suivants :

Table 14.32 – Taux de chômage (X, en %) de huit pays d'Europe

Pays	X (chômage)	Y (prix des transactions)
Irlande	4.4	427
Espagne	8.5	382
France	9.5	324
Italie	6.8	303
Suède	7.1	295
Autriche	4.7	485
Royaume Uni	5.3	473

$$\sum_{i=1}^{7} x_i = 46.3 \qquad \sum_{i=1}^{7} y_i = 2689 \quad \sum_{i=1}^{7} x_i^2 = 328.69 \qquad \sum_{i=1}^{7} y_i^2 = 1071017 \quad \sum_{i=1}^{7} x_i y_i = 17145.1.$$

1. Calculer les variances marginales de X et Y.
2. Calculer la covariance entre X et Y.
3. Déterminer l'équation de la droite de régression de Y en fonction de X.
4. Déterminer les coefficients de corrélation et de détermination entre les variables X et Y et ensuite déterminez le sens de la relation et la qualité de l'ajustement.
5. Donner la valeur ajustée et le résidu pour l'Irlande.
6. Estimer, à l'aide du modèle, le prix moyen d'une transaction immobilière dans la capitale de l'Allemagne, pays qui a un taux de chômage de 8.4 %.

Solution

1. Variances marginales de X et Y :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{7} \times 328.69 - \left(\frac{46.3}{7}\right)^2 = 3.207$$

,

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{7} \times 1071017 - \left(\frac{2689}{7}\right)^2 = 5436.694.$$

2. Covariance entre X et Y :

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{1}{7} \times 17145.1 - \left(\frac{46.3}{7} \times \frac{2689}{7}\right)^2 = -91.531.$$

3. Équation de la droite de régression de Y en fonction de X :

$$b = \frac{s_{xy}}{s_x^2} = \frac{-91.531}{3.207} = -28.541,$$

$$a = \bar{y} - b\,\bar{x} = \frac{2689}{7} + 28.541\frac{46.3}{7} = 572.921,$$

$$y = a + b\,x = 572.921 - 28.541x.$$

4. Coefficients de corrélation et de détermination entre les variables X et Y et sens de la relation et la qualité de l'ajustement :
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-91.531}{\sqrt{3.207 \times 5436.694}} = -0.693,$$
$$r_{xy}^2 = -0.693^2 = 0.481.$$
Ajustement moyen (48.1% de la variance est expliquée), relation négative entre les deux variables.
5. Valeur ajustée et le résidu pour l'Irlande :
$$y_1^* = a + b\ x_1 = 572.921 - 28.541 \times 4.4 = 447.342.$$
$$e_1 = y_1 - y_1^* = 427 - 447.342 = -20.342.$$
6. Prix moyen d'une transaction immobilière dans la capitale de l'Allemagne, pays qui a un taux de chômage de 8.4 % :
$$y_{All}^* = 572.921 - 28.541 \times 8.4 = 333.177.$$

Exercice 35.
Le Tableau 14.33 représente la répartition, par degré et par sexe, de 800 personnes appartenant au corps enseignant d'une ville. Le degré A représente la scolarité obligatoire, le degré B représente le secondaire et le degré C représente les Hautes Ecoles Universitaires et Spécialisées.

TABLE 14.33 – Répartition, par degré et par sexe, de 800 enseignants

	Hommes	Femmes
degré A	159	326
degré B	45	29
degré C	174	67

1. Quel pourcentage de ces membres du corps enseignant travaillent au sein du degré C ?
2. Quel pourcentage des hommes travaillent au sein du degré A ?
3. Quel pourcentage des personnes travaillant dans le degré B sont des femmes ?
4. Donner le tableau des effectifs théoriques.

Solution

1. Pourcentage de ces membres du corps enseignant travaillent au sein du degré C :
$$\frac{174+67}{800} = 30.125\%.$$
2. Pourcentage des hommes travaillent au sein du degré A :
$$\frac{159}{159+45+174} = 42.063\%.$$
3. Pourcentage des femmes dans le degré B :
$$\frac{29}{45+29} = 39.189\%.$$
4. Effectifs théoriques :
$$n^*_{jk} = \frac{n_{j.}n_{.k}}{n}$$

	Hommes	Femmes	Total
degré A	229.163	255.838	485
degré B	34.965	39.035	74
degré C	113.873	127.128	241
Total	378	422	800

Exercice 36.
On possède un dé pipé (truqué) à six faces. Soit X la variable aléatoire représentant le score obtenu au lancer de ce dé.

1. La fonction de répartition $F_X(x)$ est donnée dans le Tableau 14.34.

Table 14.34 – Fonction de répartition

x	1	2	3	4	5	6
$F_X(x)$	0.1	0.3	0.45	0.55	0.7	1

   (a) Quelle est la probabilité de faire un score strictement supérieur à 3 avec ce dé ?
   (b) Déterminer la distribution de probabilité $p_X(x)$ de la variable aléatoire X en remplissant le Tableau 14.35.

Table 14.35 – Distribution de probabilité $p_X(x)$

x	1	2	3	4	5	6
$p_X(x)$						

   (c) Quelle est la probabilité d'obtenir un 5 ?
2. On lance 7 fois ce même dé et on s'intéresse à la variable aléatoire Y qui représente le nombre de fois que l'on a obtenu un 6.
   (a) Quelle est la probabilité qu'au cours des 7 lancers, l'on ait obtenu aucun 6 ?
   (b) Quelle est la probabilité qu'au cours des 7 lancers, l'on ait obtenu exactement quatre fois le nombre 6 ?

Solution

1. (a) $\Pr(X > 3) = 1 - \Pr(X < 3) = 0.55$
   (b) Le tableau :

x	1	2	3	4	5	6
$p_X(x)$	0.1	0.2	0.15	0.1	0.15	0.3

   (c) $\Pr(X = 5) = 0.15$
2. (a) $Y \sim \mathcal{B}(7, 0.3)$
   $\Pr(Y = 0) = \binom{7}{0} \times 0.3^0 \times 0.7^7 = 0.082,$
   (b) $\Pr(Y = 4) = \binom{7}{4} \times 0.3^4 \times 0.7^3 = 0.097.$

Exercice 37.
La population d'une ville doit se prononcer pour ou contre la construction d'une nouvelle salle de spectacle. On aimerait savoir s'il existe une relation entre la catégorie d'âge des électeurs et leur intention de vote sur la question. Le Tableau 14.36 représente les résultats obtenus lors d'une étude réalisée sur 1000 personnes.

Table 14.36 – Pour ou contre la construction d'une nouvelle salle de spectacle

	Intention de vote		
Catégorie d'âge	Pour	Contre	Total
moins de 40 ans	130	40	170
de 40 à 60 ans	220	90	310
plus de 60 ans	210	310	520
Total	560	440	1000

1. Parmi les individus de plus de 60 ans, quelle est la proportion d'opposants à la construction de la nouvelle salle ?
2. Parmi les opposants à la construction, quelle est la proportion de jeunes de moins de 40 ans ?
3. Calculer le V de Cramer et commenter le résultat obtenu.

Solution

1. Proportion d'opposants parmi les individus de plus de 60 ans : $\frac{310}{520} = 0.596$.
2. Proportion de jeunes de moins de 40 ans Parmi les opposants $\frac{40}{440} = 0.091$.
3. Pour le V de Cramer, il faut calculer les effectifs théoriques :

Tableau des effectifs théoriques

	Intention de vote		
Catégorie d'âge	Pour	Contre	Total
moins de 40 ans	95.2	74.8	170
de 40 à 60 ans	173.6	136.4	310
plus de 60 ans	291.2	228.8	520
Total	560	440	1000

Tableau des écarts à l'indépendance

	Intention de vote		
Catégorie d'âge	Pour	Contre	Total
moins de 40 ans	34.8	−34.8	0
de 40 à 60 ans	46.4	−46.4	0
plus de 60 ans	−81.2	81.2	0
Total	0	0	0

Tableau des $\frac{e_{jk}^2}{n_{jk}^*}$

	Intention de vote		
Catégorie d'âge	Pour	Contre	Total
moins de 40 ans	12.721	16.190	
de 40 à 60 ans	12.402	15.784	
plus de 60 ans	22.642	28.817	
Total			108.557

$$\chi^2_{obs} = 108.557.$$

$$\phi^2 = \frac{\chi^2_{obs}}{1000} = 0.109.$$

$$V = \sqrt{\frac{0.021018}{1}} = 0.329.$$

La dépendance entre l'âge et l'intention de vote est relativement élevée.

Exercice 38.
On considère les données du tableau suivant contenant la superficie totale, exprimée en km^2, des 26 cantons suisses.

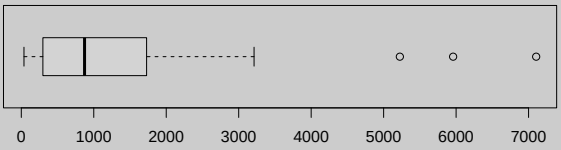
Table 14.37 – Superficie totale des cantons suisses, en km^2

Canton	Superficie	Canton	Superficie
Bâle-Ville	37	Schwytz	907
Appenzell Rh.-I.	173	Thurgovie	991
Zoug	239	Uri	1077
Appenzell Rh.-E.	243	Argovie	1404
Nidwald	276	Lucerne	1493
Genève	282	Fribourg	1671
Schaffhouse	298	Zurich	1729
Obwald	491	St. Gall	2026
Bâle-Campagne	518	Tessin	2812
Glaris	685	Vaud	3212
Soleure	790	Valais	5224
Neuchâtel	803	Berne	5959
Jura	839	Grisons	7105

1. Calculez les premier, deuxième (la médiane) et troisième quartiles, ainsi que la distance interquartile.
2. Donner le coefficient d'asymétrie de Yule l'asymétrie de la série.
3. Dessiner horizontalement la boîte à moustache (boxplot) de cette série en prenant soin d'y annoter les éléments permettant de la construire.

Solution

1. Quartiles :
$$np = 26\frac{1}{4} = 6.5 \text{ donc } x_{\frac{1}{4}} = x_{(\lceil 6.5 \rceil)} = x_{(7)} = 298$$
$$np = 26\frac{1}{2} = 13 \text{ donc } \frac{1}{2}\left(x_{(13)} + x_{(14)}\right) = \frac{1}{2}(839 + 907) = 873$$
$$np = 26\frac{3}{4} = 19.5 \text{ donc } x_{\frac{3}{4}} = x_{(\lceil 19.5 \rceil)} = x_{(20)} = 1729$$
$$IQ = x_{\frac{3}{4}} - x_{\frac{1}{4}} = 1729 - 298 = 1431$$
2. Coefficient d'asymétrie de Yule
$$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}} = \frac{1729 + 298 - 2 \times 873}{1431} = 0.196$$
La distribution est légèrement allongée à droite.
3. Boxplot
$$b^- = x_{\frac{1}{4}} - 1.5IQ = 298 - 1.5 \times 1431 = -1848.5$$
$$b^+ = x_{\frac{3}{4}} + 1.5IQ = 1729 + 1.5 * 1431 = 3875.5$$
Valeurs adjacentes (hauteur des moustaches) : 37 et 3212
Valeurs extrêmes : 5224, 5959 et 7105.



En langage R

```
Superficie=c(37,173,239,243,276,282,298,491,518,685,790,803,839,907,991,
1077,1404,1493,1671,1729,2026,2812,3212,5224,5959,7105)
boxplot(Superficie,horizontal=TRUE)
```

Exercice 39.
Une équipe de football participe à un championnat dans lequel il dispute 36 matchs. Le Tableau 14.38 représente le nombre de but marqués par match (x_j) par cette équipe et les effectifs (n_j) pour les 36 journées de championnat.

Table 14.38 – Nombre de but marqués

x_j	n_j	N_j	f_j	F_j
0	6			
1	10			
2	11			
3	7			
5	2			

1. Compléter le tableau statistique ci-dessus.
2. Calculer le nombre moyen de buts par match marqué par cette équipe.
3. Calculer le coefficient d'asymétrie de Yule et de Pearson pour la variable "Nombre de buts marqués par match".
4. Sachant que la variance est $s_x^2 = 1.576$, calculer le coefficient d'asymétrie de Pearson.
5. Commenter l'asymétrie de la variable.(3 lignes maximum)
6. Donner une représentation graphique adéquate des effectifs.

Solution

1. Tableau complété :

x_j	n_j	N_j	f_j	F_j
0	6	6	0.167	0.167
1	10	16	0.278	0.445
2	11	27	0.306	0.751
3	7	34	0.194	0.945
5	2	36	0.056	1.001

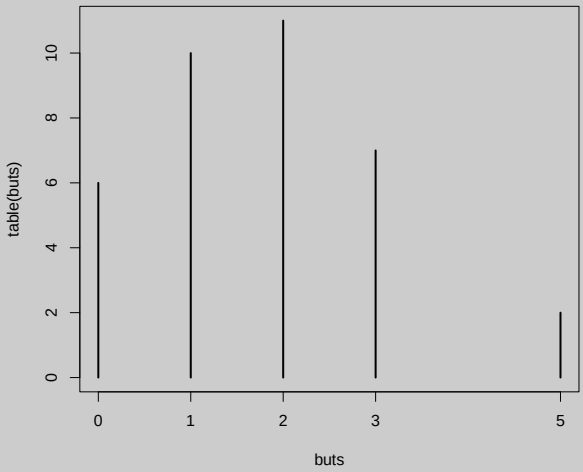
2. Moyenne :
$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{1}{36}(6\times 0 + 10\times 1 + 11\times 2 + 7\times 3 + 2\times 5) = 1.75.$$

3. Coefficient d'asymétrie de Yule :
$$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}},$$
$$np = 36\frac{1}{4} = 9 \Rightarrow x_{1/4} = \frac{1}{2}\left(x_{(9)} + x_{(10)}\right) = 1,$$
$$np = 36\frac{1}{2} = 18 \Rightarrow x_{1/2} = \frac{1}{2}\left(x_{(18)} + x_{(19)}\right) = 2,$$
$$np = 36\frac{3}{4} = 27 \Rightarrow x_{3/4} = \frac{1}{2}\left(x_{(27)} + x_{(28)}\right) = 2.5.$$
$$A_Y = \frac{2.5 + 1 - 2\times 2}{2.5 - 1} = -0.333$$

4. Coefficient d'asymétrie de Pearson :
$$A_P = \frac{\bar{x} - x_M}{s_x} = \frac{1.75 - 2}{\sqrt{1.576}} = -0.199$$

5. La variable est légèrement allongé à gauche (les deux coefficients sont négatifs).

6. Faire un diagramme en batonnets des n_j car variable quantitative discrete. Attention au fait qu'il y a un effectif nul lorsque $x = 4$.



En langage R

```
buts=c(rep(0,6),rep(1,10),rep(2,11),rep(3,7),rep(5,2))
plot(table(buts),main="")
```

Exercice 40.
À l'aide de données concernant 10 appartements en location dans le canton de Neuchâtel, on s'intéresse au lien entre le loyer annuel (variable Y, en milliers de francs suisses) et la taille du logement (X, en m^2). Une régression linéaire a été effectuée sur ces données à l'aide du principe des moindres carrés afin d'expliquer le coût du loyer par la taille du logement. Seuls une partie des résultats est disponible ci-dessous :

$$\sum_{i=1}^{10} x_i = 636, \quad \sum_{i=1}^{10} y_i = 152, \quad \sum_{i=1}^{10} x_i^2 = 48788, \quad \sum_{i=1}^{10} y_i^2 = 2776, \quad b = 0.215.$$

1. Donner l'équation de la droite de régression.
2. Calculer la covariance s_{xy}.
3. Calculer le coefficient de corrélation et commenter le sens et la force de la relation entre les deux variables.
4. Donner le coefficient de détermination r_{xy}^2 et interpréter la qualité d'ajustement.
5. Le premier appartement qui figure dans les données a une surface de 88 m^2 et son loyer annuel (en milliers de francs) est $y_1 = 24$. Donner la valeur ajustée et le résidu pour cet appartement.
6. À votre avis, sur le nuage de point représentant ces deux variables, le point représentant cet appartement se situerait-il au-dessus ou en-dessous de la droite de régression ? Justifier.
7. L'affirmation suivante est-elle vraie pour ces données ? Justifier.
$$s_{y^*}^2 > s_e^2$$
8. L'affirmation suivante est-elle vraie ? Justifier.

Si la droite de régression est croissante, alors les résidus e_i sont tous positifs ou nuls.

Solution

1. Droite de régression :
$$b = 0.215, \bar{x} = 63.6, \bar{y} = 15.2, a = \bar{y} - b\ \bar{x} = 15.2 - 0.215 \times 63.6 = 1.526.$$
$$Y = a + b\ x = 1.526 + 0.215\ x.$$
2. Covariance
$$b = \frac{s_{xy}}{s_x^2} \Rightarrow s_{xy} = b s_x^2,$$
$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{10} \times 48788 - 63.6^2 = 833.84,$$
$$s_{xy} = b s_x^2 = 0.215 \times 833.84 = 179.276.$$
3. Coefficient de corrélation :
$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}},$$
$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{10} \times 2776 - 15.2^2 = 46.56,$$
$$r_{xy} = \frac{179.276}{\sqrt{833.84 \times 46.56}} = 0.910.$$
Il y a une relation linéaire positive forte entre les deux variables.
4. Coefficient de détermination :
$$r_{xy}^2 = 0.910^2 = 0.828.$$
82.8% de la variance de Y est expliquée par X (très bon ajustement).
5. résidu, valeur ajustée :
$$y_1^* = a + b\ x_1 = 1.526 + 0.215 \times 88 = 20.446,$$
$$e_1 = y_1 - y_1^* = 24 - 20.446 = 3.554.$$

6. Au-dessus car le résidu e_1 est positif.
7. Vrai. La variance de régression est plus grande que la variance résiduelle car $r_{xy}^2 > 0.5$.
8. Faux. la somme des résidus est égal à 0. Donc, il y a forcément des résidus positifs et négatifs à partir du moment où $r_{xy} \neq 1$.

Exercice 41.
On vous offre un ticket d'un jeu de loterie. Soit X la variable aléatoire représentant le gain à ce jeu de loterie (en francs suisses).

1. La fonction de répartition $F_X(x)$ de la variable aléatoire X deécrite dans le Tableau 14.39.

Table 14.39 – Gain à ce jeu de loterie : répartition

x	0	2	4	10	20	1000
$F_X(x)$	0.6	0.8	0.9	0.95	0.99	1

   (a) Quelle est la probabilité de gagner 10 francs ou plus ?
   (b) Déterminer la distribution de probabilité $p_X(x)$ de la variable aléatoire X en remplissant le Tableau 14.40.

Table 14.40 – Gain à ce jeu de loterie : distribution

x	0	2	4	10	20	1000
$p_X(x)$						

   (c) Quelle est la probabilité de gagner exactement 20 francs à ce jeu ?
   (d) Calculer l'espérance du gain à ce jeu de loterie.
   (e) Calculer la variance de X.
2. On vous offre maintenant 7 tickets de ce même jeu de loterie. Soit Y la variable aléatoire représentant votre gain total pour ces 7 tickets. Calculer l'espérance et la variance de Y.
3. Dans ce même lot de 7 tickets, on s'intéresse maintenant à la variable aléatoire Z qui représente le nombre de tickets gagnants (un ticket gagnant est un ticket pour lequel le gain est supérieur à 0) se trouvant dans ce lot.
   (a) Quelle est la probabilité que sur les 7 tickets, aucun ne soit gagnant ?
   (b) Quelle est la probabilité que sur les 7 tickets, exactement 5 tickets soit gagnants ?
   (c) Calculer l'espérance et la variance de la variable aléatoire Z.

Solution

1. (a) $\Pr(X \geq 10) = 1 - \Pr(X < 10) = 1 - 0.9 = 0.1$.
   (b) Distribution de probabilité :

x	0	2	4	10	20	1000
$p_X(x)$	0.6	0.2	0.1	0.05	0.04	0.01

   (c) $\Pr(X = 20) = 0.04$.
   (d) $\mathrm{E}(X) = 0 \times 0.6 + 2 \times 0.2 + 4 \times 0.1 + 10 \times 0.05 + 20 \times 0.04 + 1000 \times 0.01 = 12.1$.
   (e) $\mathrm{var}(X) = 0 \times 0.6 + 2^2 \times 0.2 + 4^2 \times 0.1 + 10^2 \times 0.05 + 20^2 \times 0.04 + 1000^2 \times 0.01 - 12.1^2 = 9876.99$.
2. Comme $Y = 7X$,
$$\mathrm{E}(Y) = \mathrm{E}(7X) = 7\mathrm{E}(X) = 84.7$$
   et
$$\mathrm{var}(Y) = \mathrm{var}(7X) = 7^2\mathrm{var}(X) = 483972.51.$$
3. (a) Comme $Z \sim \mathcal{B}(7, 0.4)$, $\Pr(Z = 0) = \binom{7}{0}0.4^0 0.6^7 = 0.028$.
   (b) $\Pr(Y = 5) = \binom{7}{5}0.4^5 0.6^2 = 0.077$.
   (c) $\mathrm{E}(Y) = np = 7 \times 0.4 = 2.8$, $\mathrm{var}(Y) = np(1-p) = 7 \times 0.4 \times 0.6 = 1.68$.

Exercice 42.
Dans le Tableau 14.41, on a le revenu moyen en milliers d'Euros (y_i) et le pourcentage des gens qui travaillent les dimanches (x_i) pour 7 pays. On souhaite effectuer une régression linéaire pour expliquer le revenu moyen par le pourcentage des gens travaillant les dimanches. On connaît également les informations suivantes :

TABLE 14.41 – Données par Pays

Pays	x_i	y_i
Belgique	10.3	21.002
France	14.0	23.191
Suisse	17.5	33.991
Italie	12.6	17.963
Allemagne	13.0	21.223
Grèce	13.8	13.505
Autriche	16.4	22.106

$$\sum_{i=1}^{7} x_i = 97.6, \quad \sum_{i=1}^{7} y_i = 152.982, \quad \sum_{i=1}^{7} x_i^2 = 1395.5, \quad \sum_{i=1}^{7} y_i^2 = 3578.508, \quad r = 0.5979.$$

1. Calculez les variances marginales de x et y.
2. Calculez le coefficient de détermination r_{xy}^2 et commentez sur la qualité du résultat.
3. Calculez la covariance $\text{cov}_{x,y}$.
4. Donnez l'équation de la droite de régression.
5. Calculez la valeur ajustée et le résidu pour l'Autriche en utilisant l'équation de la droite de régression obtenue au point précédent.
6. Où se situe le point (x_i, y_i) pour l'Autriche par rapport à la droite de régression ? Justifiez.

Solution

1. Variance marginale de x : $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{7} \times 1395.5 - (13.9429)^2 = 4.95.$

   Variance marginale de y : $s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 \frac{1}{7} \times 3578.208 - 21.8546^2 = 33.574.$

2. Coefficient de détérmination : $r_{xy}^2 = r_{xy}^2 = 0.5979^2 = 0.357.$
   Le coefficient de détérmination est assez faible, car 35.7% de la variance de y est expliquée par la variance de x.
3. Covariance de x et y :
   Comme $r = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}}$, on a $s_{xy} = r_{xy}\sqrt{s_x^2}\sqrt{s_y^2}$. Donc, $s_{xy} = 0.5979 \times 2.225 \times 5.794 = 7.708.$
4. Équation de la droite de régression de y en fonction de x.

$$b = \frac{s_{xy}}{s_x^2} = \frac{7.7129}{4.9527} = 1.557,$$

$$a = \bar{y} - b\bar{x} = 21.855 - (1.557 \times 13.943) = 21.8546 - 21.7133 = 0.146,$$

$$y = a + b\ x = 0.146 + 1.557x.$$

5. Valeur ajustée et résidu pour l'Autriche :

$$y^*_{Autriche} = 0.146 + 1.557 \times 16.4 = 25.681,$$

$$e_{Autriche} = y - y^* = 22.106 - 25.681 = -3.57.5.$$

6. Le point (x_i, y_i) pour l'Autriche se trouve en-dessous de la droite de régression, parce que le résidu est négatif.

Exercice 43.
Dans le Tableau 14.42, on a relevé les préférences de 250 femmes entre 20 et 59 ans pour trois produits.

TABLE 14.42 – Préférence des femmes pour produits A, B, C

	Produit A	Produit B	Produit C	Total
Groupe $A(20 \leq$ âge $< 40)$	35	30	45	110
Groupe $B(40 \leq$ âge $< 60)$	30	50	60	140
Total	65	80	105	250

1. Quelle proportion de femmes préfère le produit B ?
2. Parmi les femmes qui préfèrent le produit C, quelle proportion a entre 40 et 60 ans ?
3. Parmi les femmes qui ont entre 20 et 40 ans, quelle proportion préfère le produit A ?
4. Dresser le tableau des effectifs théoriques.

Solution

1. Proportion de femmes qui préfère le produit B : $\frac{80}{250} = 0.32$.
2. Proportion de femmes qui a entre 40 et 60 ans, étant donné qu'elles préfèrent le produit C.

$$\Pr(40 \leq \text{ âge } < 60 | \text{produit } C) = \frac{\Pr(40 \leq \text{ âge } < 60 \text{ et produit } C)}{\text{produit } C} = \frac{\frac{60}{250}}{\frac{105}{250}} = 0.5714.$$

3. Proportion de femmes qui préfère le produit A, étant donné qu'elle a entre 20 et 40 ans

$$\Pr(\text{produit } A | 20 \leq \text{ âge } < 40) = \frac{\Pr(\text{produit } A \text{ et } 20 \leq \text{ âge } < 40)}{20 \leq \text{ âge } < 40} = \frac{\frac{35}{250}}{\frac{110}{250}} = 0.3182.$$

4. Effectifs théoriques

Effectifs théoriques des préférence des femmes pour produits A, B, C

	Produit A	Produit B	Produit C
Groupe $A(20 \leq$ âge $< 40)$	$\frac{110\times65}{250} = 28.6$	$\frac{110\times80}{250} = 35.2$	$\frac{110\times105}{250} = 46.2$
Groupe $B(40 \leq$ âge $< 60)$	$\frac{140\times65}{250} = 36.4$	$\frac{140\times80}{250} = 44.8$	$\frac{140\times105}{250} = 58.8$

Exercice 44.
Dans une urne on a 10 balles rouges, 15 balles noires et 5 balles blanches. On tire avec remise 4 balles de cette urne.

1. Quelle est la probabilité que 2 balles soient rouges parmi les 4 balles tirées ?
2. Représentez graphiquement la distribution de probabilité du nombre des boules rouges tirées (en tenant compte du fait qu'on tire 4 balles de l'urne).
3. Si l'on tire sans remise 4 balles de l'urne, quelle est la probabilité que l'on obtienne au moins 1 balle noire ?

Solution

1. Probabilité que 2 balles soient rouges : $\Pr(X = 2) = \binom{4}{2} \times \Pr(\text{rouge})^2 \times \Pr(\overline{\text{rouge}})^2 = 0.296$
2. $\Pr(X = 0) = \binom{4}{0} \times \Pr(\text{rouge})^0 \times \Pr(\overline{\text{rouge}})^4 = 0.198$

   $\Pr(X = 1) = \binom{4}{1} \times \Pr(\text{rouge})^1 \times \Pr(\overline{\text{rouge}})^3 = 0.395$

   $\Pr(X = 2) = \binom{4}{2} \times \Pr(\text{rouge})^2 \times \Pr(\overline{\text{rouge}})^2 = 0.296$

   $\Pr(X = 3) = \binom{4}{3} \times \Pr(\text{rouge})^3 \times \Pr(\overline{\text{rouge}})^1 = 0.099$

   $\Pr(X = 4) = \binom{4}{4} \times \Pr(\text{rouge})^4 \times \Pr(\overline{\text{rouge}})^0 = 0.012$
3. $1 - \Pr(\text{noire}) = 1 - \frac{15}{30}\frac{14}{29}\frac{13}{28}\frac{12}{27} = 0.950$

Exercice 45.
On dispose de données fournies par l'Office Fédéral (suisse) de la Statistique (Tableau 14.43) concernant le produit intérieur brut en francs suisses par habitant en 2011 pour chaque canton.

TABLE 14.43 – Produit intérieur brut par habitant

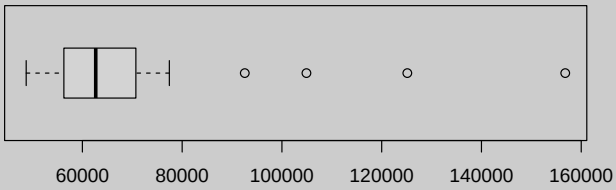
Pays	x_i	Pays	x_i
Uri	48733	Argovie	65174
Appenzell Rh.-Ext.	49329	Saint-Gall	65649
Fribourg	50235	Tessin	66611
Appenzell Rh.-Int.	50739	Vaud	67159
Valais	53867	Berne	67704
Schwiz	54317	Lucerne	69905
Thurgovie	56288	Bâle-Campagne	70721
Jura	56595	Neuchâtel	71126
Obwald	58083	Schaffhouse	77430
Glaris	58571	Zurich	92553
Nidwald	58968	Genève	104914
Grisons	59914	Zoug	125138
Soleure	60178	Bâle-Ville	156795

1. Calculez le premier quartile, la médiane et le troisième quartile.
2. Calculez la distance interquartile.
3. Calculez les bornes de la la boîte à moustaches. Indiquez les valeurs adjacentes et les valeurs extrêmes.
4. Dessinez la boîte à moustaches.

Solution

1. Premier quartile :
$$np = 26 \times \frac{1}{4} = 6.5$$
qui n'est pas un nombre entier. Donc,
$$x_{1/4} = x_{[6.5]} = x_{[7]} = 56\,288. \text{ (Thurgovie)}$$
Troisième quartile :
$$np = 26 \times \frac{3}{4} = 19.5$$
qui n'est pas un nombre entier. Donc,
$$x_{3/4} = x_{[19.5]} = x_{[20]} = 70\,721. \text{ (Bâle Campagne)}$$
La médiane n'étant autre que les deuxième quartile :
$$np = 26 \times \frac{1}{2} = 13$$
qui est un nombre entier. Donc,
$$x_{1/2} = \frac{1}{2}\left(x_{[13]} + x_{[14]}\right) = \frac{60\,178 + 65\,174}{2} = 62\,676.$$
2. $\text{IQ} = x_{3/4} - x_{1/4} = 70\,721 - 56\,288 = 14\,433.$
3. $\text{b}^- = x_{1/4} - 1.5\text{IQ} = 56\,288 - 14\,433 \times 1.5 = 34\,638.5.$
Valeur adjacente = 48 733. (Uri)
$\text{b}^+ = x_{3/4} + 1.5\text{IQ} = 70\,721 + 1.5 \times 14\,433 = 92\,370.3.$
Valeur adjacente = 77 430 (Schaffhouse)
Valeur(s) extrême(s) = (Zurich), (Genève), (Zoug) et (Bâle Ville).

4. Le boxplot :



En langage R

```
X=c(48733,49329,50235,50739,53867,54317,56288,56595,58083,58571,58968,59914,
60178,65174,65649,66611,67159,67704,69905,70721,71126,77430,92553,104914,
125138,156795)
boxplot(X,horizontal=TRUE)
```

Exercice 46.
On s'intéresse à l'impact que l'alcool pourrait potentiellement avoir sur l'espérance de vie. Pour cela, on dispose de données (Tableau 14.44) concernant l'espérance de vie de 14 pays et de leur consommation d'alcool respective (en litres par habitant). Ces observations ont été récoltées à partir de la base de données des statistiques de l'OCDE. On souhaite donc effectuer une régression linéaire en considérant l'espérance de vie comme la variable dépendante.

Table 14.44 – Variables

Espérance de vie(y_i)	Consommation d'alcool(x_i)
78.0	10.3
80.5	12.4
79.2	8.2
77.3	12.1
78.8	10.9
79.9	10.3
77.3	12.3
80.2	9.9
79.1	12.4
79.9	7.5
76.9	9.5
75.1	12.0
74.8	10.8
76.9	14.2

On connaît également les informations suivantes :

$$\sum_{i=1}^{14} y_i = 1093.9, \quad \sum_{i=1}^{14} x_i = 152.8, \quad \sum_{i=1}^{14} y_i^2 = 85516.05, \quad \sum_{i=1}^{14} x_i^2 = 1710.24, \quad r_{xy} = -0.336,$$

1. Calculez les moyennes et les variances marginales de X et Y.
2. Calculez la covariance entre les variables X et Y.
3. Donnez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez la qualité de l'ajustement à l'aide du coefficient de détermination. À votre avis, ça aurait du sens d'envisager une régression de X (la consommation en alcool) en fonction de Y, l'espérance de vie ? Justifiez votre réponse.
5. Donnez le résidu de l'observation $x = 12.1$.
6. Sur la base de ce modèle, établissez une prévision de l'espérance de vie pour une consommation d'alcool de 11 litres par habitant.

Solution

1. Moyennes et variances marginales :
   $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{152.8}{14} = 10.914.$$
   $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1093.9}{14} = 78.136.$$
   $$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{14} \times 1710.24 - (10.914)^2 = 3.038.$$
   $$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{14} \times 85516.05 - (78.136)^2 = 3.099.$$
2. Le covariance vaut :
   $s_{xy} = r_x y \times s_x \times s_y = -0.336 \times \sqrt{3.038} \times \sqrt{3.099} = -1.032.$

3. L'équation de la droite de régression de Y en fonction de X.

$$b = \frac{s_{xy}}{s_x^2} = \frac{-1.032}{3.038} = -0.340.$$

$$a = \bar{y} - b\ \bar{x} = 78.136 - (-0.340) \times 10.914 = 81.843.$$

$$y = a + b\ x = 81.843 - 0.340x.$$

4. Le coefficient de détermination vaut :

$$r_{xy}^2 = (r_{xy})^2 = (-0.336)^2 = 0.113.$$

Ce n'est pas un si bon modèle car que 11.3% de la variance de y est expliquée par x.
Ca n'aurait pas de sens d'effectuer la régression inverse car l'espérance de vie n'explique pas la consommation d'alcool.

5. On sait que : $e_{12.1} = y_{12.1} - y^*_{12.1} = y_{12.1} - (81.843 - 0.340 \times 12.1) = 77.3 - 77.733 = -0.433$.
6. $y^*_{11} = 81.843 - 0.340 \times 11 = 78.107$.

Exercice 47.
Considérons deux variables, le sexe du nouveau-né et la couleur de ses yeux. On a effectué 250 observations afin de déterminer si il y aurait une dépendance entre ces deux variables. De plus, on sait que parmi ces 250 bébés, il y en avait 150 qui étaient des filles et 80 qui avait les yeux noirs, 90 les yeux châtains et 50 les yeux bleus.

Table 14.45 – Tableau des profiles colonnes

Sexe	Châtain	Bleu	Vert	Noir	Total
Garçon	0.444	0.6	0.167	0.312	0.4
Fille	0.556	0.4	0.833	0.688	0.6
	1	1	1	1	1

1. Etablissez le tableau des effectifs.
2. Quel pourcentage de bébés ayant les yeux châtains sont des garçons ?
3. Etablissez le tableau des effectifs théoriques.
4. En sachant que $\chi^2_{obs} = 18.43$, calculez le V de Cramer et interprétez ce résultat.

Solution

1. Tableau des effectifs

Sexe	Châtain	Bleu	Vert	Noir	Total
Garçon	40	30	5	25	100
Fille	50	20	25	55	150
Total	90	50	30	80	250

2. Pourcentage d'enfants ayant des yeux châtains qui sont des garçons : $\frac{40}{90} = 44.4\%$.
3. Tableau des effectifs théoriques

Sexe	Châtain	Bleu	Vert	Noir	Total
Garçon	36	20	12	32	100
Fille	54	30	18	48	150
Total	90	50	30	80	250

4. Calcul du V de Cramer :

$$\phi^2 = \frac{\chi^2_{obs}}{n} = \frac{18.43}{250} = 0.074.$$

$$V = \sqrt{\frac{\phi^2}{\min(J-1.K-1)}} = \sqrt{\frac{0.074}{\min(1,3)}} = \sqrt{0.074} = 0.272.$$

Exercice 48.
Une banque voudrait revoir sa politique de carte de crédit. Il sait que par le passé environ 50% des détenteurs de cartes de crédit ont été insolvables et la banque a été incapable de recouvrer le solde impayé. Par conséquence la banque a estimé à 0.5 la probabilité qu'un détenteur d'une carte de crédit soit insolvable. La banque a également découvert que la probabilité d'oublier 1 ou plusieurs payements mensuels est de 0.3 pour les clients solvables. Bien entendu, la probabilité d'oublier 1 ou plusieurs paiements pour un client insolvable est égale à 1.

1. Calculez la probabilité qu'un nouveau client ait oublié un ou plusieurs paiements.
2. Etant donné qu'un client a oublié d'effectuer un paiement mensuel, calculez à posteriori la probabilité que le créditeur soit insolvable.
3. La banque voudrait reprendre sa carte de crédit si la probabilité qu'un client soit insolvable est supérieure à 0.3. La banque devrait-elle reprendre la carte de crédit si le client oublie un paiement mensuel ? Pourquoi ?
4. Trouver un système complet d'évènements. Justifiez votre réponse.

Solution

1. À partir de l'énoncé on peut définir les événements et les probabilités suivantes :
   I = "Le détenteur d'une carte de crédit est un client insolvable",
   S = "Le détenteur d'une carte de crédit est un client solvable",
   O = "Le client a oublié d'effectuer 1 ou plusieurs paiements".

   $$\Pr(I) = 0.5, \Pr(O|S) = 0.3, \Pr(S) = 0.5, \Pr(O|I) = 1.$$

   Par le théorème des probabilités totales, on a

   $$\Pr(O) = \Pr(O|S)\Pr(S) + \Pr(O|I)\Pr(I) = 0.3 \times 0.5 + 1 \times 0.5 = 0.65.$$

2. On cherche $\Pr(I|O)$. Par le théorème de Bayes, on a

   $$\Pr(I|O) = \frac{\Pr(I)\Pr(O|I)}{\Pr(O)} = \frac{1 \times 0.5}{0.65} = 0.769.$$

3. Suite à la réponse de la question 2 on sait que pour un détenteur d'une carte de crédit la probabilité d'être insolvable sachant qu'il a oublié un paiement mensuel est de 0.769, bien supérieur à 0.3, la banque devrait donc lui reprendre sa carte.
4. Un système complet d'évènements pourrait être :
   E_1="le détenteur d'une carte de crédit est un client solvable".
   E_2="le détenteur d'une carte de crédit est un client insolvable".

Exercice 49.
On s'intéresse à analyser le nombre de biens motorisés (ex. voitures, motos etc.) possédés par 50 ménages suisses.

Table 14.46 – Biens Motorisés

Nombre de biens motorisés x_j	Effectifs n_j	Effectifs cumulés N_j	Fréquences f_j	Fréquences cumulées F_j
0	3			
1	13			
2	16			
3	11			
4	5			
5	2			
Total	50	–	1.00	–

1. Complétez le tableau avec les effectifs cumulés, les fréquences et les fréquences cumulées.
2. Calculez le premier quartile, la médiane et le troisième quartile.
3. Calculez la moyenne et la mode.
4. Calculez la variance.
5. Calculez le coefficient d'asymétrie de Pearson et interprétez le résultat obtenu.
6. Représentez graphiquement la fonction de répartition.

Solution

1. Tableau complété :

Nombre de chemises vendues x_j	Effectifs n_j	Effectifs cumulés N_j	Fréquences f_j	Fréquences cumulées F_j
0	3	3	$3/50 = 0.06$	0.06
1	13	16	$13/50 = 0.26$	$0.06 + 0.26 = 0.32$
2	16	32	$16/50 = 0.32$	$0.32 + 0.32 = 0.64$
3	11	43	$11/50 = 0.22$	$0.64 + 0.22 = 0.86$
4	5	48	$5/50 = 0.10$	$0.86 + 0.10 = 0.96$
5	2	50	$2/50 = 0.04$	$0.96 + 0.04 = 1$
Total	31	–	1.00	–

2. Premier quartile :

$$np = 50 \times \frac{1}{4} = 12.5$$

qui n'est pas un nombre entier. Donc,

$$x_{1/4} = x_{[12.5]} = x_{[13]} = 1.$$

Troisième quartile :

$$np = 50 \times \frac{3}{4} = 37.5$$

qui n'est pas un nombre entier. Donc,

$$x_{3/4} = x_{[37.5]} = x_{[38]} = 3.$$

La médiane n'étant autre que le deuxième quartile :

$$np = 50 \times \frac{1}{2} = 25$$

qui est un nombre entier. Donc,

$$x_{1/2} = \frac{1}{2}\left(x_{[25]} + x_{[26]}\right) = \frac{2+2}{2} = 2.$$

3. Moyenne :

$$\bar{x} = \frac{1}{50}\sum_{j=1}^{6}\left(x_j n_j\right) = \frac{0\times 3 + 1\times 13 + 2\times 16 + 3\times 11 + 4\times 5 + 5\times 2}{50} = 2.16.$$

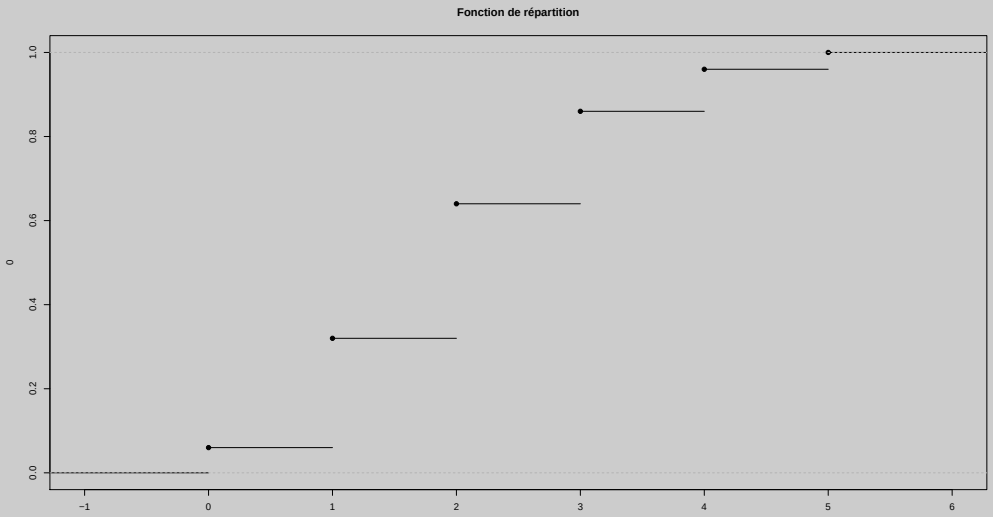
   et $x_M = 2$.

4. Variance :

$$s_x^2 = \frac{1}{50}\sum_{j=1}^{6} x_j^2 n_j - \bar{x}^2 = \frac{1}{50}(0^2\times 3 + 1^2\times 13 + 2^2\times 16 + 3^2\times 11 + 4^2\times 5 + 5^2\times 2) - 2.16^2 = 1.454.$$

5. Le coefficient de Pearson est égal à $\dfrac{\bar{x}-2}{\sqrt{1.454}} = 0.133$. Le coefficient est supérieur à 0. La distribution est allongée à droite.

6. Fonction de répartition :



Exercice 50.
Dans le tableau suivant, on dispose de données de 10 voitures en vente aux USA en 1993. On cherche à savoir si le prix de vente (y_i) dépend de la puissance du véhicule (x_i). Pour cela, on souhaite effectuer une régression linéaire pour expliquer le prix par la puissance du véhicule. On connaît également les informations suivantes :

Table 14.47 – Données de voitures en vente aux USA

y_i	x_i
15.9	140
33.9	200
29.1	172
37.7	172
30.0	208
15.7	110
20.8	170
23.7	180
26.3	170
34.7	200

$$\sum_{i=1}^{10}, y_i = 267.8, \quad \sum_{i=1}^{10} x_i = 1722, \quad \sum_{i=1}^{10} y_i^2 = 7706.72, \quad \sum_{i=1}^{10} x_i^2 = 304332, \quad \sum_{i=1}^{10} x_i y_i = 47675.6.$$

1. Calculez les variances marginales de X et Y.
2. Calculez le coefficient de corrélation.
3. Donnez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez la qualité de l'ajustement à l'aide du coefficient de détermination.
5. Donnez la variance résiduelle.
6. Sur la base de ce modèle, établissez une prévision du prix d'une voiture ayant une puissance de 190 chevaux.

Solution

1. Variance marginale de X : $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{10} \times 304332 - (172.2)^2 = 780.36.$

   Variance marginale de Y : $s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{10} \times 7706.72 - (26.78)^2 = 53.504.$
2. Le coefficient de corrélation vaut :
   $s_{xy} = 4767.56 - 26.78 \times 172.2 = 156.044,$
   $r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{156.044}{\sqrt{780.36}\sqrt{53.504}} = 0.764.$
3. L'équation de la droite de régression de Y en fonction de X.

$$b = \frac{s_{xy}}{s_x^2} = \frac{156.044}{780.36} = 0.2, \qquad a = \bar{y} - b\,\bar{x} = 26.78 - (0.2 \times 172.2) = -7.660,$$

$$y = a + b\,x = -7.660 + 0.2x.$$

4. Le coefficient de détermination vaut :

$$r_{xy}^2 = (r_{xy})^2 = 0.764^2 = 0.584.$$

   Ce n'est pas un si bon modèle car 58.4% de la variance de y est expliquée par x.
5. On sait que : $s_e^2 = s_y^2(1 - r_{xy}^2) = 53.504 \times (1 - 0.584) = 22.258,$
6. $y_{190}^* = -7.660 + 0.2 \times 190 = 30.340.$

Exercice 51.
Le Tableau 14.48 représente la répartition, par niveaux d'étude et par sexe, de 600 employés travaillant dans une entreprise horlogère.

TABLE 14.48 – Niveaux d'étude

Niveau d'étude	CFC	Bachelor	Master
Hommes	200	80	50
Femmes	150	100	20

1. Quel pourcentage d'employés a un niveau CFC ?
2. Parmi les femmes travaillant dans cette entreprise, quel pourcentage a un Bachelor ?
3. Etablir le tableau des fréquences.
4. Calculez le tableau des effectifs théoriques sous hypothèse d'indépendance entre le sexe et le niveau d'études.

Solution

1. Pourcentage d'employés ayant un CFC : $\dfrac{200+150}{600} = 58.33\%$.
2. Pourcentage de femmes ayant un Bachelor : $\dfrac{100}{150+100+20} = 37.04\%$.
3. Tableau des fréquences

Niveau d'étude	CFC	Bachelor	Master	Total
Hommes	0.34	0.13	0.08	0.55
Femmes	0.25	0.17	0.03	0.45
Total	0.59	0.3	0.11	1.00

4. Tableau des effectifs théoriques

Niveau d'étude	CFC	Bachelor	Master	Total
Hommes	192.5	99	38.5	330
Femmes	157.5	81	31.5	270
Total	350	180	70	600

Exercice 52.
Le Tableau 14.49 contient des données sur le nombre de voitures vendues par un concessionnaire neuchâtelois. Les données concernent 16 trimestres consécutifs de 2010 à 2013. On pense que la série est de la forme :

$$Y_t = T_t + S_t + e_t,$$

où T_t est la tendance, S_t est la composante saisonnière, qui ne dépend que du numéro de trimestre dans l'année et telle que la somme des S_t sur quatre trimestres consécutifs vaut 0, et les e_t sont des résidus.

Table 14.49 – Tableau des voitures vendues

Année	Trimestre	Y_t	T_t	$Y_t - T_t$
2010	1	48	–	–
2010	2	223	–	–
2010	3	105	103.125	1.875
2010	4	33	?	?
2011	1	55	116.625	−61.625
2011	2	270	118.375	151.625
2011	3	112	120.125	−8.125
2011	4	40	124.875	−84.185
2012	1	62	131.500	−69.500
2012	2	301	135.375	165.625
2012	3	134	138.500	−4.500
2012	4	49	144.000	−95.000
2013	1	78	150.750	−72.750
2013	2	329	154.875	174.125
2013	3	160	–	–
2013	4	56	–	–

1. Estimez la tendance T_4 (pour la 4ème observation) par une moyenne mobile $MM(4)$.
2. Isolez la composante saisonnière par la méthode additive (aidez-vous du Tableau).
3. Ajustez les composantes saisonnières obtenues au point 3 pour que la somme soit nulle.
4. Désaisonnalisez l'observation 1.

Solution

1. On applique la moyenne mobile $MM(4) = (L^2 + 2L + 2I + 2F + F^2)/8$.

$$MM(4)y_4 = \frac{1}{8}(223 + 2 \times 105 + 2 \times 33 + 2 \times 55 + 270) = \frac{879}{8} = 109.875.$$

2. On cherche la composante saisonnière. On commence par enlever la tendance en calculant $Y_t - T_t$, puis, pour chaque trimestre, on fait la moyenne des valeurs observées. Les observations numérotées, $1, 5, 9, 13$ correspondent à un trimestre, le premier de l'année, les observations numérotées $2, 6, 10, 14$ au deuxième trimestre, les observations numérotées $3, 7, 11, 15$ au troisième et celles numérotées $4, 8, 12, 16$ au quatrième. Pour chaque trimestre, on a trois observations disponibles de $Y_t - T_t$. On trouve :

$$S'_{1,5,9,13} = (-61.625 - 69.5 - 72.75)/3 = -67.958.$$

On trouve aussi :

$$S'_{2,6,10,14} = 163.792, \quad S'_{3,7,11,15} = -4.833 \text{ et } S'_{4,8,12,16} = -85.353.$$

3. On fait ensuite traditionnellement un ajustement de la série pour que la somme des composantes saisonnières soit nulle.

$$S_t = S'_t - \frac{1}{4}\sum_{i=1}^{4} S_i.$$

On a par exemple :

$$S_1 = S'_1 - \frac{1}{4}\sum_{i=1}^{4} S'_t = -67.958 - \frac{1}{4} \times (-67.958 + 163.792 + 4.833 - 85.353) = -69.370.$$

On trouve aussi :

$$S_2 = 162.380, \quad S_3 = -6.243 \text{ et } S_4 = -86.765.$$

4. La série désaisonnalisée est la série obtenue en ôtant la composante saisonnière : $\tilde{Y}_t = Y_t - S_t$.

$$\tilde{Y}_1 = Y_1 - S_1 = 48 - (-69.370) = 117.370.$$

Exercice 53.
On s'intéresse au nombre de personnes ayant descendu les pistes d'une montagne de ski. Le tableau suivant indique le nombre de passages pendant une journée sur les 20 pistes de la montagne :

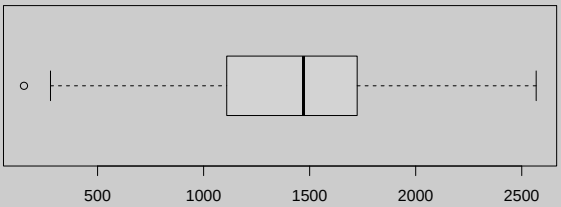
Table 14.50 – Nombre de passages sur les pistes d'une montagne de ski

Piste	Passages	Piste	Passages
Flamme	153	Dévaleuse	1500
La 68	278	Rouge-Gorge	1513
Geai-Bleu	904	La Plagne	1678
L'Expo	956	Soleil	1690
Tobogan	1022	Pancake	1705
Pingouin	1196	Edge	1743
Cascades	1203	Laurentienne	1788
Zig-Zag	1211	Belvédère	1814
Alpine	1337	Promenade	1899
Bouleaux	1442	Skibidou	2568

1. Calculez les premier, deuxième (la médiane) et troisième quartiles, ainsi que la distance interquartile.
2. Donnez le coefficient de Yule et commentez.
3. Construisez la boite à moustaches (box plot) de cette distribution.

Solution

1. Quartiles :
   — $np = 20\frac{1}{4} = 5$. Donc, $x_{\frac{1}{4}} = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(1022 + 1196) = 1109$.
   — $np = 20\frac{1}{2} = 10$. Donc, $x_{\frac{1}{2}} = \frac{1}{2}\left(x_{(10)} + x_{(11)}\right) = \frac{1}{2}(1442 + 1500) = 1471$.
   — $np = 20\frac{3}{4} = 15$. Donc, $x_{\frac{3}{4}} = \frac{1}{2}\left(x_{(15)} + x_{(16)}\right) = \frac{1}{2}(1705 + 1743) = 1724$.
   — $IQ = x_{\frac{3}{4}} - x_{\frac{1}{4}} = 1742 - 1109 = 615$.
2. $A_Y = \dfrac{x_{\frac{3}{4}} + x_{\frac{1}{4}} - 2x_{\frac{1}{2}}}{x_{\frac{3}{4}} - x_{\frac{1}{4}}} = \dfrac{1724 + 1109 - 2 \times 1471}{1724 - 1109} = -0.177$
   (faible allongement à gauche).
3. $b^- = x_{\frac{1}{4}} - 1.5IQ = 1109 - 1.5 \times 615 = 186.5$
   $b^+ = x_{\frac{3}{4}} + 1.5IQ = 1724 + 1.5 \times 615 = 2646.5$
   Valeurs adjacentes : 278 et 2568.
   Valeur extrême : 153.



En langage R

```
X=c(153,278,904,956,1022,1196,1203,1211,1337,1442,1500,1513,
1678,1690,1705,1743,1788,1814,1899,2568)
boxplot(X,horizontal=TRUE)
```

Exercice 54.
On s'intéresse à la grosseur des stations de ski d'une certaine chaîne de montagnes. On a une liste des 12 stations de ski de la chaîne de montagnes avec l'altitude en mètres du plus haut sommet (X) et le nombre de pistes de ski (Y) de chacune. La liste est donnée dans le Tableau 14.51 :

Table 14.51 – Données des stations de ski

i	x_i	y_i
1	1600	15
2	2120	17
3	2575	19
4	2864	24
5	2911	25
6	3005	20
7	3160	28
8	3200	33
9	3250	33
10	3415	39
11	3450	35
12	3500	41

On a les résultats suivants :

$$\sum_{i=1}^{12} x_i = 35050, \sum_{i=1}^{12} y_i = 329, \sum_{i=1}^{12} x_i^2 = 105994292, \sum_{i=1}^{12} y_i^2 = 9865, \sum_{i=1}^{12} x_i y_i = 1009341.$$

1. Calculez les variances marginales de X et Y.
2. Calculez la covariance entre X et Y.
3. Déterminez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez les coefficients de corrélation et de détermination entre les variables X et Y et ensuite déterminez la qualité de l'ajustement.
5. Donnez la valeur ajustée et le résidu pour la première observation du Tableau 14.51.

Solution

1. Variances marginales de X et Y.

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{11}105994292 - 2920.833^2 = 329007.607$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{11}9865 - 27.4167^2 = 76.811.$$

2. Covariance entre X et Y.

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{1}{12}1009341 - 2920.833 \times 27.4167 = 4032.148$$

3. Équation de la droite de régression de Y en fonction de X.

$$D_{y|x} \ : y = a + b\ x,$$

avec

$$b = \frac{s_{xy}}{s_x^2} = \frac{4032.148}{329007.6} = 0.012,$$

et

$$a = \bar{y} - \frac{s_{xy}}{s_x^2}x = 27.4167 - 0.012 \times 2920.833 = -7.633,$$

$$D_{y|x} \ : y = 0.012x - 7.633.$$

4. Coefficients de corrélation et de détermination entre les variables X et Y et qualité de l'ajustement.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{4032.148}{\sqrt{329007.6 \times 76.8106}} = 0.802.$$

$r^2 = 0.802^2 = 0.643 \Rightarrow$ bon ajustement.

5. Valeur ajustée et le résidu pour la première observation :

$$y_1^* = 0.012 \times 1600 - 7.633 = 11.567,$$

et

$$e_1 = 15 - 11.567 = 3.433.$$

Exercice 55.
Le propriétaire d'une station de ski aimerait savoir s'il existe une relation entre le sexe et le type de sport de glisse. Le tableau de contingence suivant présente la répartition, par sexe et par type de sport de glisse, de 1000 personnes ayant fréquenté cette station : Malheureusement, le fils du propriétaire s'est amusé à effacer

Table 14.52 – Tableau de contingence

	Ski alpin	Planche à neige	Télémark	Total
Femme	?	?	45	455
Homme	312	?	43	545
Total	612	300	?	?

quelques données du tableau.

1. Complétez le tableau de contingence.
2. Quel est le pourcentage de personnes pratiquant le télémark ?
3. Quel pourcentage de pratiquants de planche à neige sont des femmes ?
4. À l'aide de la table de contingence, déterminez, au moyen du V de Cramer, s'il existe une relation entre le sexe et le type de sport de glisse pratiqué.

Solution

1. Tableau de contingence :

Tableau de contingence

	Ski alpin	Planche à neige	Télémark	Total
Femme	300	110	45	455
Homme	312	190	43	545
Total	612	300	88	1000

2. T = "une personne pratique le télémark",
$\Pr(T) = \frac{88}{1000} = 0.088$ donc 8.8%.
3. P = "une personne pratique la planche à neige",
F = "une personne est une femme",
$\Pr(F|P) = \frac{110}{300} = 0.367$ donc 36.7%.
4. Tableau des effectifs théoriques :

Tableau des effectifs théoriques

	Ski alpin	Planche à neige	Télémark	Total
Femme	278.46	136.50	40.04	455
Homme	333.54	163.50	47.96	545
Total	612	300	88	1000

$$\chi^2_{obs} = \sum_{k=1}^{3}\sum_{j=1}^{2}\frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}} = 13.624$$

$$\phi^2 = \frac{\chi^2_{obs}}{n} = \frac{13.624}{1000} = 0.014$$

$$V = \sqrt{\frac{\phi^2}{\min(1,2)}} = \sqrt{\frac{0.014}{1}} = 0.117.$$

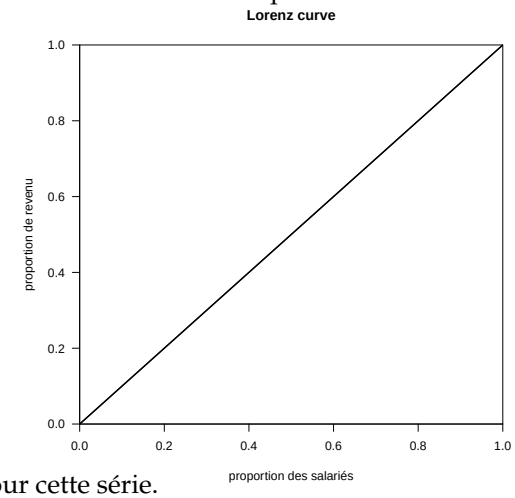
Comme $V = 0.117$ est proche de 0, la relation de dépendance est très faible.

Exercice 56.
L'équipe de comptabilité d'une station de ski est composée de 10 salariés. Le Tableau 14.53 contient leur revenu mensuel en milliers de francs (X) :

TABLE 14.53 – Revenu mensuel en milliers de francs

	i	revenu x_i	revenu dans la classe $x_{(i)}$	revenu cumulé $\sum_{j=1}^{i} x_{(j)}$
	1	2	2	2
	2	3	3	5
	3	4	4	9
	4	5	5	14
	5	10	10	24
Total	–	-	24	—

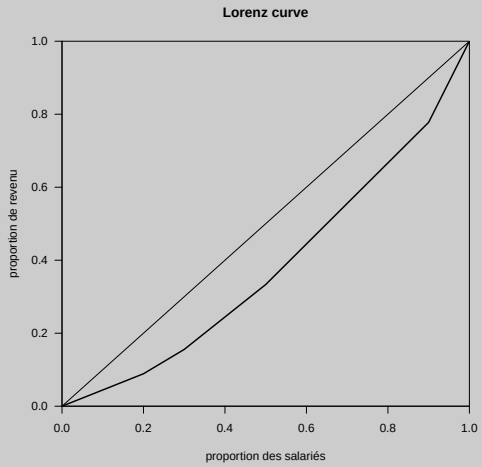
1. Tracez la courbe de Lorenz de cette série dans le repère ci-dessous.



2. Calculez l'indice de Gini pour cette série.

Solution

1. Courbe de Lorenz



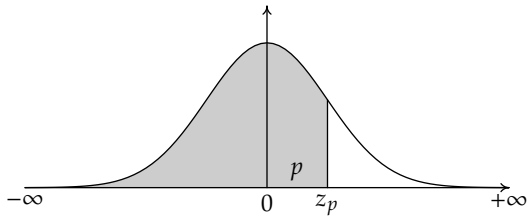
2. $G = \dfrac{1}{n-1}\left[\dfrac{2\sum_{i=1}^{n} i x_{(i)}}{n\bar{x}} - (n+1)\right] = \dfrac{1}{4}\left[\dfrac{2\times 90}{5\times 4.8} - 6\right] = 0.375.$

Quatrième partie

Annexes

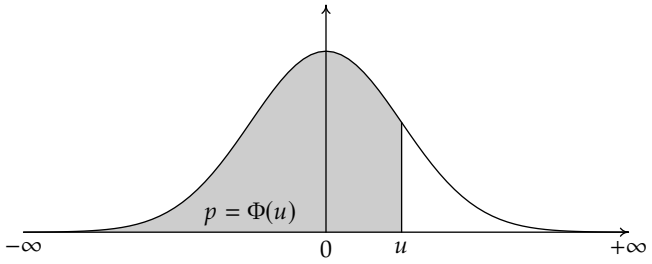
Tables statistiques

Table 14.54 – Table des quantiles $z_p = \Phi^{-1}(p)$ d'une variable normale centrée réduite



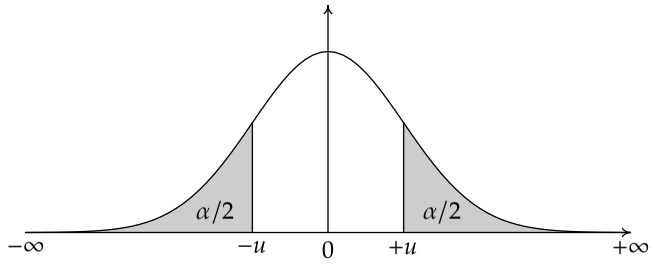
Ordre du quantile (p)	quantile (z_p)	Ordre du quantile (p)	Quantile (z_p)
0.500	0.0000	0.975	1.9600
0.550	0.1257	0.976	1.9774
0.600	0.2533	0.977	1.9954
0.650	0.3853	0.978	2.0141
0.700	0.5244	0.979	2.0335
0.750	0.6745	0.990	2.3263
0.800	0.8416	0.991	2.3656
0.850	1.0364	0.992	2.4089
0.900	1.2816	0.993	2.4573
0.950	1.6449	0.994	2.5121
0.970	1.8808	0.995	2.5758
0.971	1.8957	0.996	2.6521
0.972	1.9110	0.997	2.7478
0.973	1.9268	0.998	2.8782
0.974	1.9431	0.999	3.0902

Table 14.55 – Fonction de répartition de la loi normale centrée réduite
(*Probabilité de trouver une valeur inférieur à u*)



u	0.0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table 14.56 – Quantiles de la loi normale centrée réduite
(*u : valeur ayant la probabilité α d'être dépassé en valeur absolue*)



α	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	∞	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4958	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

Table 14.57 – Table des quantiles d'une variable χ^2 à n degrés de liberté

	ordre du quantile					
	0.01	0.025	0.05	0.95	0.975	0.99
n=1	0.000157	0.000982	0.003932	3.841	5.024	6.635
2	0.02010	0.05064	0.103	5.991	7.378	9.210
3	0.115	0.216	0.352	7.815	9.348	11.34
4	0.297	0.484	0.711	9.488	11.14	13.28
5	0.554	0.831	1.145	11.07	12.83	15.09
6	0.872	1.237	1.635	12.59	14.45	16.81
7	1.239	1.690	2.167	14.07	16.01	18.48
8	1.646	2.180	2.733	15.51	17.53	20.09
9	2.088	2.700	3.325	16.92	19.02	21.67
10	2.558	3.247	3.940	18.31	20.48	23.21
11	3.053	3.816	4.575	19.68	21.92	24.72
12	3.571	4.404	5.226	21.03	23.34	26.22
13	4.107	5.009	5.892	22.36	24.74	27.69
14	4.660	5.629	6.571	23.68	26.12	29.14
15	5.229	6.262	7.261	25.00	27.49	30.58
16	5.812	6.908	7.962	26.30	28.85	32.00
17	6.408	7.564	8.672	27.59	30.19	33.41
18	7.015	8.231	9.390	28.87	31.53	34.81
19	7.633	8.907	10.12	30.14	32.85	36.19
20	8.260	9.591	10.85	31.41	34.17	37.57
21	8.897	10.28	11.59	32.67	35.48	38.93
22	9.542	10.98	12.34	33.92	36.78	40.29
23	10.20	11.69	13.09	35.17	38.08	41.64
24	10.86	12.40	13.85	36.42	39.36	42.98
25	11.52	13.12	14.61	37.65	40.65	44.31
26	12.20	13.84	15.38	38.89	41.92	45.64
27	12.88	14.57	16.15	40.11	43.19	46.96
28	13.56	15.31	16.93	41.34	44.46	48.28
29	14.26	16.05	17.71	42.56	45.72	49.59
30	14.95	16.79	18.49	43.77	46.98	50.89
31	15.66	17.54	19.28	44.99	48.23	52.19
32	16.36	18.29	20.07	46.19	49.48	53.49
33	17.07	19.05	20.87	47.40	50.73	54.78
34	17.79	19.81	21.66	48.60	51.97	56.06
35	18.51	20.57	22.47	49.80	53.20	57.34
36	19.23	21.34	23.27	51.00	54.44	58.62
37	19.96	22.11	24.07	52.19	55.67	59.89
38	20.69	22.88	24.88	53.38	56.90	61.16
39	21.43	23.65	25.70	54.57	58.12	62.43
40	22.16	24.43	26.51	55.76	59.34	63.69
42	23.65	26.00	28.14	58.12	61.78	66.21
44	25.15	27.57	29.79	60.48	64.20	68.71
46	26.66	29.16	31.44	62.83	66.62	71.20
48	28.18	30.75	33.10	65.17	69.02	73.68
50	29.71	32.36	34.76	67.50	71.42	76.15
60	37.48	40.48	43.19	79.08	83.30	88.38
70	45.44	48.76	51.74	90.53	95.02	100.43
80	53.54	57.15	60.39	101.88	106.63	112.33
90	61.75	65.65	69.13	113.15	118.14	124.12
100	70.06	74.22	77.93	124.34	129.56	135.81
110	78.46	82.87	86.79	135.48	140.92	147.41
120	86.92	91.57	95.70	146.57	152.21	158.95

Table 14.58 – Table des quantiles d'une variable de Student à n degrés de liberté

	ordre du quantile			
	0.95	0.975	0.99	0.995
n=1	6.314	12.71	31.82	63.66
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
31	1.696	2.040	2.453	2.744
32	1.694	2.037	2.449	2.738
33	1.692	2.035	2.445	2.733
34	1.691	2.032	2.441	2.728
35	1.690	2.030	2.438	2.724
36	1.688	2.028	2.434	2.719
37	1.687	2.026	2.431	2.715
38	1.686	2.024	2.429	2.712
39	1.685	2.023	2.426	2.708
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
60	1.671	2.000	2.390	2.660
70	1.667	1.994	2.381	2.648
80	1.664	1.990	2.374	2.639
90	1.662	1.987	2.368	2.632
100	1.660	1.984	2.364	2.626
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.327	2.576

Table 14.59 – Table des quantiles d'ordre 0.95 d'une variable de Fisher à n_1 et n_2 degrés de liberté

	n_1=1	2	3	4	5	6	7	8	9	10	12	14	16	20	30	∞
n_2=1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.4	246.5	248.0	250.1	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.45	19.46	19.50
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.715	8.692	8.660	8.617	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.873	5.844	5.803	5.746	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.636	4.604	4.558	4.496	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.956	3.922	3.874	3.808	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.529	3.494	3.445	3.376	3.230
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.237	3.202	3.150	3.079	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.025	2.989	2.936	2.864	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.865	2.828	2.774	2.700	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.739	2.701	2.646	2.570	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.637	2.599	2.544	2.466	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.554	2.515	2.459	2.380	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.484	2.445	2.388	2.308	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.424	2.385	2.328	2.247	2.066
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.373	2.333	2.276	2.194	2.010
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.329	2.289	2.230	2.148	1.960
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.290	2.250	2.191	2.107	1.917
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.256	2.215	2.155	2.071	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.225	2.184	2.124	2.039	1.843
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.197	2.156	2.096	2.010	1.812
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.173	2.131	2.071	1.984	1.783
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.150	2.109	2.048	1.961	1.757
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	2.130	2.088	2.027	1.939	1.733
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.111	2.069	2.007	1.919	1.711
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	2.094	2.052	1.990	1.901	1.691
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	2.078	2.036	1.974	1.884	1.672
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	2.064	2.021	1.959	1.869	1.654
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	2.050	2.007	1.945	1.854	1.638
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.037	1.995	1.932	1.841	1.622
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142	2.070	2.015	1.972	1.908	1.817	1.594
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123	2.050	1.995	1.952	1.888	1.795	1.569
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106	2.033	1.977	1.934	1.870	1.776	1.547
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091	2.017	1.962	1.918	1.853	1.760	1.527
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.948	1.904	1.839	1.744	1.509
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.952	1.895	1.850	1.784	1.687	1.438
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.860	1.815	1.748	1.649	1.389
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.775	1.728	1.659	1.554	1.254
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831	1.752	1.692	1.644	1.571	1.459	1.000

Table 14.60 – Table des quantiles d'ordre 0.99 d'une variable de Fisher à n_1 et n_2 degrés de liberté

	n_1=1	2	3	4	5	6	7	8	9	10	12	14	16	20	30	∞
n_2=1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6143	6170	6209	6261	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.45	99.47	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.69	26.51	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.02	13.84	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.888	9.770	9.680	9.553	9.379	9.020
6	13.75	10.93	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.605	7.519	7.396	7.229	6.880
7	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.359	6.275	6.155	5.992	5.650
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.559	5.477	5.359	5.198	4.859
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	5.005	4.924	4.808	4.649	4.311
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.601	4.520	4.405	4.247	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.293	4.213	4.099	3.941	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.052	3.972	3.858	3.701	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.857	3.778	3.665	3.507	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.698	3.619	3.505	3.348	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.564	3.485	3.372	3.214	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.451	3.372	3.259	3.101	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.353	3.275	3.162	3.003	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.269	3.190	3.077	2.919	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.195	3.116	3.003	2.844	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.130	3.051	2.938	2.778	2.421
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.072	2.993	2.880	2.720	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	3.019	2.941	2.827	2.667	2.305
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.973	2.894	2.781	2.620	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.930	2.852	2.738	2.577	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.892	2.813	2.699	2.538	2.169
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.857	2.778	2.664	2.503	2.131
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.824	2.746	2.632	2.470	2.097
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.795	2.716	2.602	2.440	2.064
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.767	2.689	2.574	2.412	2.034
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.742	2.663	2.549	2.386	2.006
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934	2.798	2.696	2.618	2.503	2.340	1.956
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894	2.758	2.657	2.578	2.463	2.299	1.911
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946	2.859	2.723	2.622	2.543	2.428	2.263	1.872
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915	2.828	2.692	2.591	2.512	2.397	2.232	1.837
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.563	2.484	2.369	2.203	1.805
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.562	2.461	2.382	2.265	2.098	1.683
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.394	2.315	2.198	2.028	1.601
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.234	2.154	2.035	1.860	1.381
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.185	2.082	2.000	1.878	1.696	1.000

Liste des tableaux

1.1 Domaine de la variable . . . 14
1.2 Série statistique . . . 14
1.3 Tableau statistique . . . 14
1.4 Codification de la variable Y . . . 16
1.5 Série statistique de la variable Y . . . 16
1.6 Tableau statistique complet . . . 16
1.7 Nombre de personnes par ménage . . . 18
1.8 Tableau statistique du nombre de personnes par ménage . . . 18
1.9 Taille en centimètres de 50 élèves . . . 20
1.10 Bornes des classes . . . 20
1.11 Distribution groupée . . . 21

2.1 Tableau statistique . . . 25
2.2 Tableau des valeurs distinctes . . . 26
2.3 Notes d'un étudiant . . . 29

3.1 Poids Y et taille X de 20 individus . . . 41
3.2 Tableau de contingence . . . 49
3.3 Tableau des effectifs n_{jk} . . . 49
3.4 Tableau de fréquences . . . 49
3.5 Tableau des fréquences . . . 49
3.6 Tableau des profils lignes . . . 50
3.7 Tableau des profils colonnes . . . 50
3.8 Tableau des effectifs théoriques n^*_{jk} . . . 51
3.9 Tableau des écarts à l'indépendance e_{jk} . . . 51
3.10 Tableau des e^2_{jk}/n^*_{jk} . . . 51
3.11 Tableau de contingence : effectifs n_{jk} . . . 52
3.12 Tableau des fréquences f_{jk} . . . 52
3.13 Tableau des profils lignes . . . 52
3.14 Tableau des profils colonnes . . . 52
3.15 Tableau des effectifs théoriques n^*_{jk} . . . 52
3.16 Tableau des écarts à l'indépendance e_{jk} . . . 53
3.17 Tableau des e^2_{jk}/n^*_{jk} . . . 53

4.1 Tableau du prix d'un bien de consommation de 2000 à 2006 . . . 55
4.2 Tableau de l'indice simple du prix du Tableau 4.1 . . . 55
4.3 Exemple : prix et quantités de trois bien pendant 3 ans . . . 56
4.4 Mesures de l'inégalité par pays : source Eurostat . . . 61

5.1 Biens manufacturés aux USA . . . 64
5.2 Indice des prix à la consommation, France (Source : Gouriéroux and Monfort, 1983) . . . 65
5.3 Trafic du nombre de voyageurs SNCF . . . 67
5.4 Hauteur du lac de Neuchâtel : moyennes mensuelles (Source : Office fédéral de l'environnement) 68
5.5 Décomposition de la variable FRIG, méthode additive . . . 78
5.6 Moyenne des composantes saisonnières . . . 78
5.7 Décomposition de la variable FRIG, méthode multiplicative . . . 79
5.8 Moyenne des composantes saisonnières . . . 79
5.9 Prix moyen du Mazout pour 100 ℓ (achat entre 800 et 1500 ℓ) . . . 82
5.10 Lissage exponentiel simple et double de la série temporelle Prix moyen du Mazout pour 100 litres (achat entre 800 et 1500 litres) en CHF . . . 83

6.1 Illustration du théorème des probabilités totales . . . 89
6.2 Factorielle des nombres de 1 à 10 . . . 90

7.1 Nombre de noisettes mangées par 27 écureuils . . . 114
7.2 Branche choisie par 24 étudiants (B=Biologie, C=chimie, M=mathématique, F=français) . . . 114
7.3 Note moyenne de 22 étudiants . . . 114
7.4 Poids des élèves . . . 118
7.5 Taux d'occupation professionnelle de 20 individus . . . 120
7.6 Nombre d'enfants par couple (40 couples) . . . 120
7.7 Domicile des 20 élèves d'une classe . . . 120
7.8 Loyers mensuels en francs suisses de 25 appartements . . . 120
7.9 Série continue . . . 125
7.10 Âge des 50 employés d'une entreprise . . . 127
7.11 Nombre de jours de chômage pour 40 personnes . . . 129
7.12 Répartition de la population d'un pays par groupe d'âge . . . 129
7.13 Qualité de production de 30 produits (D = défectueux, Q = de bonne qualité) . . . 129

8.1 Classe A . . . 138
8.2 Classe B . . . 138
8.3 Nombre de véhicules par habitant selon les communes . . . 143
8.4 Salaire mensuels de 5 hommes et 5 femmes . . . 145
8.5 Nombre d'enfants par couple . . . 146
8.6 Taux d'intérêt d'un investissement . . . 147
8.7 Primes d'assurance accident payées par quatre femme et quatre hommes . . . 149
8.8 Distribution des salaires . . . 151
8.9 Série d'une variable discrète . . . 153
8.10 Principaux stades anglais . . . 156
8.11 Valeurs prises par une variable quantitative discrète . . . 158
8.12 Quotient intellectuel de 100 enfants . . . 160
8.13 Nombre de jours de chômage pour 40 personnes . . . 160
8.14 Tableaux statistiques des variables x et y . . . 160

9.1 Revenu mensuel et nombre d'années . . . 167
9.2 Ancienneté et absence . . . 169
9.3 Consommation de crèmes glacées . . . 171
9.4 Poids et tailles . . . 172
9.5 Cours du dollar en Suisse (en CHF) de 1982 à 2007. . . . 174
9.6 Avis et notes des étudiants . . . 177
9.7 Consommation de médicaments, source : Murray et al. (1981, pp. 551–560). . . . 179
9.8 Tableau des effectifs observés n_{jk} . . . 181
9.9 Tableau des contingences . . . 182
9.10 Tableau des effectifs observés n_{jk} . . . 183
9.11 Tableau des effectifs . . . 184
9.12 Tableau des profils lignes . . . 186
9.13 Sensation de manque de sommeil selon le sexe . . . 188
9.14 Profils colonnes . . . 190

10.1 Indice simple d'un bien . . . 191
10.2 Tableau des prix et quantités . . . 193
10.3 Dépenses de Monsieur Durand . . . 194
10.4 Tableau des prix et quantités . . . 196
10.5 Prix (en milliers de francs suisses) et quantité de montres vendues . . . 198
10.6 Évolution des prix du lait . . . 200
10.7 Revenus annuels en milliers de francs de 200 personnes . . . 203
10.8 Tableau année 2002 . . . 205
10.9 Revenu annuel en milliers de francs suisses de 100 personnes . . . 207
10.10 Argent à disposition par mois pour 100 étudiants . . . 209
10.11 Tableau des revenus 2004 . . . 211

11.1 Nombre de téléphones portables . . . 213
11.2 Nombre d'élèves dans le degré secondaire 1 . . . 214

11.3 Tableau des ventes . . . 215
11.4 Nombre d'entrées de travailleurs saisonniers dans un pays . . . 219

12.1 Distribution de probabilité de X . . . 233
12.2 Distribution de probabilité de X . . . 255
12.3 Distribution de probabilité de X . . . 255
12.4 Distribution de probabilité de X . . . 258
12.5 Distribution de probabilité de X . . . 260

13.1 Ancienneté de 20 professeurs . . . 297

14.1 Distribution des salaires . . . 323
14.2 Données de la livraison d'ampoules . . . 325
14.3 Données trimestrielles . . . 327
14.4 Distribution des loyers . . . 328
14.5 Bruit et productivité . . . 330
14.6 Tableau des prix et quantités . . . 331
14.7 Données trimestrielles . . . 332
14.8 Structure de la population . . . 334
14.9 Âge et coût de la santé . . . 336
14.10 Tableau de contingence . . . 337
14.11 Salaires des employés de deux entreprises . . . 338
14.12 Données trimestrielles . . . 340
14.13 Population des pays de l'UE . . . 342
14.14 Données de la livraison de fruits . . . 343
14.15 Précipitations (en mm/an) relevées pendant 10 ans . . . 344
14.16 Nombre de tickets gratuits reçu par un individu . . . 345
14.17 Nombre de voitures par ménage . . . 346
14.18 Tableau de contingence . . . 349
14.19 Distribution en classes d'âge . . . 351
14.20 Données par Pays . . . 352
14.21 Tableau de contingence . . . 353
14.22 Distribution de probabilité de X . . . 354
14.23 Population (en milliers d'individus) des 20 régions de France . . . 355
14.24 Prix et poids des 50 modèles de voitures . . . 356
14.25 Revenu mensuel (X) de chacun des 5 salariés . . . 358
14.26 Distribution de probabilité de la récompense . . . 360
14.27 Loyer moyen selon le canton par m^2 pour un studio, en francs suisses, en 2000 . . . 362
14.28 Prix trimestriels d'un titre pendant les années 2006 à 2008 . . . 363
14.29 Estimation de la tendance par une moyenne mobile . . . 364
14.30 Longueur du lancer au concours de lancer de javelot . . . 367
14.31 Longueur du lancer au concours de lancer de javelot . . . 367
14.32 Taux de chômage (X, en %) de huit pays d'Europe . . . 369
14.33 Répartition, par degré et par sexe, de 800 enseignants . . . 371
14.34 Fonction de répartition . . . 372
14.35 Distribution de probabilité $p_X(x)$ . . . 372
14.36 Pour ou contre la construction d'une nouvelle salle de spectacle . . . 373
14.37 Superficie totale des cantons suisses, en km^2 . . . 375
14.38 Nombre de but marqués . . . 377
14.39 Gain à ce jeu de loterie : répartition . . . 381
14.40 Gain à ce jeu de loterie : distribution . . . 381
14.41 Données par Pays . . . 382
14.42 Préférence des femmes pour produits A, B, C . . . 383
14.43 Produit intérieur brut par habitant . . . 385
14.44 Variables . . . 387
14.45 Tableau des profiles colonnes . . . 389
14.46 Biens Motorisés . . . 391
14.47 Données de voitures en vente aux USA . . . 393
14.48 Niveaux d'étude . . . 394
14.49 Tableau des voitures vendues . . . 395
14.50 Nombre de passages sur les pistes d'une montagne de ski . . . 397

14.51 Données des stations de ski 398
14.52 Tableau de contingence 400
14.53 Revenu mensuel en milliers de francs 401
14.54 Table des quantiles $z_p = \Phi^{-1}(p)$ d'une variable normale centrée réduite 405
14.55 Fonction de répartition de la loi normale centrée réduite 406
14.56 Quantiles de la loi normale centrée réduite 407
14.57 Table des quantiles d'une variable χ^2 à n degrés de liberté 408
14.58 Table des quantiles d'une variable de Student à n degrés de liberté 409
14.59 Table des quantiles d'ordre 0.95 d'une variable de Fisher à n_1 et n_2 degrés de liberté 410
14.60 Table des quantiles d'ordre 0.99 d'une variable de Fisher à n_1 et n_2 degrés de liberté 411

Table des figures

1.1 Diagramme en secteurs des fréquences . . . 15
1.2 Diagramme en barres des effectifs . . . 15
1.3 Diagramme en secteurs des fréquences . . . 17
1.4 Diagrammes en barres des effectifs et des effectifs cumulés . . . 18
1.5 Diagramme en bâtonnets des effectifs pour une variable quantitative discrète . . . 19
1.6 Fonction de répartition d'une variable quantitative discrète . . . 19
1.7 Histogramme des fréquences . . . 21
1.8 Histogramme des fréquences avec les deux dernières classes agrégées . . . 22
1.9 Fonction de répartition d'une distribution groupée . . . 23

2.1 Médiane quand n est impair . . . 30
2.2 Médiane quand n est pair . . . 31
2.3 Asymétrie d'une distribution . . . 35
2.4 Distributions mésokurtique et leptokurtique . . . 36
2.5 Boîtes à moustaches pour la variable superficie en hectares (HApoly) des communes du canton de Neuchâtel . . . 39
2.6 Boîtes à moustaches du "revenu moyen des habitants" des communes selon les provinces belges 40

3.1 Le nuage de points . . . 42
3.2 Exemples de nuages de points et coefficients de corrélation . . . 43
3.3 Le nuage de points, le résidu . . . 44
3.4 La droite de régression . . . 46

4.1 Courbe de Lorenz . . . 59

5.1 Dépenses en biens durables USA (milliards de dollars de 1982) . . . 64
5.2 Nombre de réfrigérateurs vendus de 1978 à 1985 . . . 65
5.3 Indice des prix à la consommation p_t . . . 66
5.4 Rapport mensuel des indices de prix p_t/p_{t-1} . . . 66
5.5 Rapport en glissement annuel des indices de prix p_t/p_{t-12} . . . 66
5.6 Trafic du nombre de voyageurs SNCF . . . 67
5.7 Hauteur du lac de Neuchâtel . . . 68
5.8 Exemple de fonction logistique avec $c = 0.5$ . . . 69
5.9 Série avec une tendance linéaire dépendant du temps . . . 71
5.10 Différence d'ordre un de la série avec une tendance linéaire . . . 71
5.11 Différence d'ordre 4 de la variable vente de 'réfrigérateurs' . . . 72
5.12 Trafic du nombre de voyageurs SNCF . . . 72
5.13 Différence d'ordre 12 sur la série trafic du nombre de voyageurs SNCF . . . 73
5.14 Logarithme du rapport d'ordre 12 sur la série trafic du nombre de voyageurs SNCF . . . 73
5.15 Nombre de réfrigérateurs et moyenne mobile d'ordre 4 . . . 75
5.16 Décomposition de la série de ventes de réfrigérateurs 5.1 . . . 77
5.17 Evolution du prix du mazout en CHF (achat entre 800 et1500 ℓ), lissage exponentiel double et lissage exponentiel simple . . . 84

6.1 Système complet d'événements . . . 86
6.2 Distribution de "faces" obtenus. . . . 91
6.3 Distribution d'une variable aléatoire binomiale avec $n = 5$ et $p = 0.6$. . . . 94
6.4 Distribution d'une variable de Poisson avec $\lambda = 1$. . . . 95
6.5 Probabilité que la variable aléatoire soit inférieure à a . . . 96
6.6 Fonction de densité d'une variable uniforme . . . 97
6.7 Fonction de répartition d'une variable uniforme . . . 97

6.8 Fonction de densité d'une variable normale 98
6.9 Fonction de répartition d'une variable normale 98
6.10 Densité d'une normale centrée réduite, symétrie 99
6.11 Fonction de densité d'une variable exponentielle avec $\lambda = 1$. 100
6.12 Densité d'une variable de chi-carré avec $p = 1, 2, \ldots, 10$ 105
6.13 Densités de variables de Student avec $p = 1, 2$ et 3 et d'une variable normale 105
6.14 Densité d'une variable de Fisher 106
6.15 Densité d'une normale bivariée 106
6.16 Nuage de points de réalisations d'une normale bivariée 107

9.1 Nuage de points 175

10.1 Courbe de Lorenz 207
10.2 Courbes de Lorenz 208

13.1 Courbes de Lorenz des pays A et B 296
13.2 Courbe de Lorenz des revenus 309
13.3 Courbes de Lorenz des pays A et B 316

14.1 Histogramme 334

Bibliographie

Boudon, R. (1979). *La logique du social : introduction à l'analyse sociologique*, volume 325. Hachette Paris.

Gouriéroux, C. and Monfort, A. (1983). *Cours de séries temporelles*. Collection Économie et Statistiques Avancées. Série Ecole nationale de la statistique et de l'administration et du Centre d'études des programmes économiques. Economica.

Murray, J., Dunn, G., Williams, P., and Tarnopolsky, A. (1981). Factors affecting the consumption of psychotropic drugs. *Psychological Medicine*, 11(3) :551–560.

Paradis, E. (2002). R pour les débutants. Montpellier (F) : University of Montpellier https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf.

Index

analyse combinatoire, 90
arrangement, 90
axiomatique, 86

Bernoulli, 92
bernoullienne, 92
binôme de Newton, 92
Boudon, 52
boxplot, 39
boîte à moustaches, 39

changement d'origine et d'unité, 36
circularité, 55
coefficient
    d'asymétrie de Fisher, 35
    d'asymétrie de Pearson, 35
    d'asymétrie de Yule, 35
    de corrélation, 43
    de détermination, 43
combinaison, 91
complémentaire, 85
composante saisonnière, 74
corrélation, 43
courbe
    de Lorenz, 58
    leptokurtique, 36
    mésokurtique, 36
    platykurtique, 36
covariance, 42, 109

densité, 21, 96
    conditionnelle, 101
    marginale, 109
diagramme
    en barres, 15
        des effectifs, 17
    en bâtonnets des effectifs, 19
    en boite, 39
    en feuilles, 38
    en secteurs, 15, 17
    en tiges, 38
différence, 70, 85
    saisonnière, 71
distance interquartile, 32
distribution
    binomiale, 92, 94
    bivariée, 100
    bivarée, 107
    conditionnelle, 101, 102
    de probabilité, 92
    exponentielle, 99
    groupée, 20
    leptokurtique, 36
    marginale, 101, 107
    mésokurtique, 36
    normale bivariée, 105, 107, 109
domaine, 13
données observées, 48
droite de régression, 43
décile, 31
    share ratio, 60
dérivées partielles, 44
désaisonnalisation, 76

écart
    moyen absolu, 34
    médian absolu, 34
    à l'indépendance, 50
écart-type, 33
    marginal, 42
effectif, 14
    d'une modalité, 14
    d'une valeur disctincte, 14
    marginal, 48
    théorique, 50
ensemble
    parties d'un ensemble, 86
    système complet, 86
espérance, 92, 102
    conditionnelle, 108
    d'une variable
        binomiale, 93
        indicatrice, 92
    propriétés, 102
étendue, 32
événements, 85
    indépendants, 89
    mutuellement exclusifs, 86
exercice
    Activités après le bachelor, 230
    Anagrammes, 247
    Ancienneté et absence, 169
    Années d'études et revenus, 167
    Assurance et prime, 259
    Avion, 249
    Avis pédagogiques, 177
    Bières, 183
    Boules de Noël, 233
    Boxplot, 153
    Boîte de conserve et loi normale, 276
    Boîtes de chocolats, 231
    Cadenas, 223
    Calcul d'indices d'inégalité, 209
    Calcul de paramètres, 144

Cartes et mains, 254
Cartes et évènements, 228
Cartes, 224
Changement d'origine, 142
Chaussures et probabilités, 236
Choix de films, 253
Classe et échantillon, 261
Classes d'élèves, 138
Cluedo, 223
Comparaison de courbes de Lorenz, 208
Consommation d'eau et loi normale, 275
Consommation de médicaments, 179
Courbe de Lorenz et inégalités, 203
Cours du dollar, 174
Course de chevaux, 248
Crèmes glacées, 171
Cylindres et loi normale, 277
Daltonie, 181
Distribution de salaires, 151
Décomposition et désaisonnalisation, 219
Dépendance entre variables dichotomiques, 186
Désaisonnalisation, 218
Employés absents pour cause de maladie et loi de Poisson, 266
Espérance et variance, 255
Espérance et écart-type, 255
Euro Millions, 251
Excès de vitesse et loi de Poisson, 265
File d'attente et loi de Poisson, 267
Germination, 262
Habitudes alimenaires selon le sexe, 190
Histogramme et classes, 127
Histogrammes, 125
Indice chaîne, 202
Indice de Gini, 205
Indice de Laspeyres et de Paasche, 194
Indice de Laspeyres, 193
Indice et lait, 200
Indices composites, 196
Indices d'inégalité et courbe de Lorenz, 207
Indices de prix et montres, 198
Indices simples, 191
Jeu avec des cartes, 232
Jeu de dés, 257
Lancer de dés, 235
Lecture des tables statistiques, 279
Lecture inverse de la table de la loi normale, 270
Loterie, 251
Loyers et cantons, 182
Manque de sommeil, 188
Monstres, 241
Moyennes arithmétique, géometrique et harmonique, 141
Moyennes géométrique, harmonique ou arithmétique, 147
Nombre d'enfants, 146
Notation binaire, 250
Opinion, 225
Opérateur de sommation 1, 133
Opérateur de sommation 2, 136
Opérateurs de décalage 1, 213
Opérateurs de décalage 2, 214
Paramètres dans une distribution, 158
Pièce de monnaie, 263
Places à table, 249
Poids d'élèves, 118
Poids et tailles, 172
Poker, 252
Pratiques culturelles, 237
Primes d'assurance, 149
Probabiliés d'accident, 240
Qualité d'ampoules, 238
Quelle moyenne ?, 143
Revenus dans les États, 211
Répartition et espérance, 258
Résultat et loi normale, 271
Salaires hommes et femmes, 145
Spectateurs dans les stades, 156
Séquence d'enfants, 229
Séries et calcul de paramètres, 160
Séries statistiques et graphiques, 114
Tabagisme et cancer, 243
Table de la loi normale 1, 268
Table de la loi normale 3, 269
Table de la loi normale 4, 270
Table de la loi normale 2, 269
Tailles et loi normale, 273
Taux de réussite, 264
Temps de travail, 242
Temps et loi normale, 273
Test de grossesse, 245
Test pharmaceutique, 246
Théorème de Bayes, 279
Théorème des probabilités totales et de Bayes, 244
Théorème des probabilités totales, 241, 278
Tirage de jetons, 227
Types de variables, 113
Urne et boules, 239
Urne et jetons, 260
Variable normale standardisation, 272
Variables et graphiques, 120
Variables, types et graphiques, 129
Variance avec une double somme, 137
Variances, 154
Villas vendues, 215
Vitesse et loi normale, 274
Vol dans les magasins, 256
Âges dans les familles, 155
État civil et nationalité, 184
expérience aléatoire, 85

filtre linéraire, 74
fonction
de densité, 96, 99
conditionnelle, 101, 102
d'une variable exponentielle, 100
d'une variable uniforme, 97
marginale, 101
de répartition, 19, 22, 29
discontinue, 31
jointe, 101
par palier, 30

forward operator, 70
fréquence, 14

groupe, 37

histogramme, 21
histogramme des fréquence, 21
homoscédastique, 108

identité, 55
indice, 55
  chaine, 58
  d'équirépartition, 60
  de Fisher, 57
  de Gini, 59
  de Hoover, 60
  de Laspeyres, 56
  de Paasche, 57
  de pauvreté, 60
  de Sidgwick, 58
  propriétés, 55
  selon les pays, 60
  simple, 56
  synthétique, 56
indépendance, 102
intersection, 85

khi-carré, 50

lag operator, 70
lissage exponentiel, 78
  double, 80
  simple, 78

médiane, 29
mesures d'inégalité, 55
mise en évidence, 27
modalités, 13
mode, 25
modèle linéaire, 70
moindres carrés, 44, 80
moment, 34
  centré, 35
  d'ordres supérieurs, 35
  à l'origine, 34
moyenne, 25, 26, 28, 30, 36, 37, 42
  arithmétique, 25
  conditionnelle, 101, 102
  géométrique, 28, 57
  harmonique, 28, 57
  marginale, 42, 101, 106, 107
  mobile, 74
    Henderson, 75, 76
    non-pondérée, 74
    Spencer, 75
    symétrique, 74
    Van Hann, 75
  pondérée, 29, 37
médiane, 31
  mobile, 76
méthode
  additive, 76
  multiplicative, 77

normale bivariée, 106

opérateur
  avance, 70
  de différence, 70
  de décalage, 70
  forward, 70
  identité, 70
  lag, 70
  retard, 70

paramètres
  d'aplatissement, 36
  de dispersion, 32
  de forme, 35
  de position, 25
  marginaux, 42
percentile, 31
permutation
  avec répétition, 90
  sans répétition, 90
piechart, 15
probabilité, 85, 86
  conditionnelle et indépendance, 88
  théorème des probabilités totales, 89
profils
  colonnes, 50
  lignes, 50
propriétés, 104
propriétés des espérances et des variances, 102

quantile, 31, 158, 405, 407–409
quartile, 31
quintile, 31
  share ratio, 60

résidus, 45, 46
réversibilité, 55

SC somme de carrés, 46, 47
série
  chronologique, 69
  statistique, 13
    bivariée, 41
  temporelle, 63
signe de sommation, 26
skewness, 35
somme
  d'une constante, 27
  des carrés, 27
    de la régression, 47
    des résidus, 44, 47
    totale, 46
statistique, 13
  descriptive
    bivariée, 41
    univariée, 25
système complet d'événements, 86

tableau

de contingence, 48
de fréquences, 49
des profils colonnes, 50
des profils lignes, 50
statistique, 15–17, 20
tendance, 68
linéaire, 69, 70
logistique, 69
parabolique, 69
polynomiale, 69
quadratile, 69
quadratique, 70
théorème
de Bayes, 89
de la variance totale, 38
transitivité, 55

union, 85
unités
d'observation, 13
statistiques, 13

valeurs
adjacentes, 39
ajustées, 45
possibles, 13
variable, 13
aléatoire, 91
continue, 95
discrète, 91
indépendante, 102
binomiale, 92
de Fisher, 105
de Poisson, 94
de Student, 104
espérance, 91
indicatrice, 92
khi-carrée, 104
normale, 98
centrée réduite, 99
ordinale, 16
qualitative, 13
nominale, 13, 14
ordinale, 13, 16
quantitative, 13, 41
continue, 13, 20
discrète, 13, 17
uniforme, 96
variance, 32–34, 36–38, 42, 46, 47, 91–93, 95, 96, 98, 99, 101, 102, 104
conditionnelle, 101, 102, 108
d'une variable
binomiale, 93
indicatrice, 92
de régression, 47
marginale, 42, 101, 106, 107
propriétés, 102
résiduelle, 47, 48