UNIVERSITÉ
DE GENÈVE
FACULTÉ DES SCIENCES
DE LA SOCIÉTÉ

Statistique pour sciences sociales

Cahier d'exercices
dont une sélection avec solutions

Cours de Gilbert Ritschard

février 2015

Table des matières

N.B. : Les exercices avec un astérisque sont tirés d'anciens examens.

1 Les méthodes quantitatives en sciences sociales

Exercice 1.1

Citer, en les justifiant brièvement, trois raisons pour lesquelles le recours aux méthodes quantitatives s'avère important en sciences sociales.

Exercice 1.2

Préciser la nature (nominale, ordinale, mesurable discrète ou continue, intervalle ou ratio) des caractères suivants :

1. La région d'origine
2. La langue maternelle
3. Le numéro de code dans le cas d'une nomenclature hiérarchisée
4. La commune de domicile
5. Le nombre de passagers d'un train
6. Le vote d'un électeur
7. La durée d'un voyage Genève-Londres
8. Les recettes journalières d'un supermarché
9. La grandeur d'un pied
10. La pointure d'une chaussure
11. La durée de vie d'une ampoule électrique
12. L'heure de fermeture d'un café-restaurant

Exercice 1.3

Soient les données suivantes sur le nombre de chômeurs en 1990 dans 3 cantons (moyenne annuelle) :

i	Cantons	Hommes (x_i)	Femmes (y_i)
1	ZH	1186	814
2	TI	1045	1096
3	GE	1472	1035

(Source : La Vie Économique 7/91)

1. Déterminer le nombre total z_i de chômeurs pour chaque canton et vérifier numériquement que :
$$\sum_{i=1}^{3} z_i = \sum_{i=1}^{3}(x_i + y_i) = \sum_{i=1}^{3} x_i + \sum_{i=1}^{3} y_i$$

2. Supposons maintenant que l'on ne dispose que du nombre total z_i de chômeurs par canton. En admettant que les chômeurs femmes représentent dans chaque canton 50 % du nombre z_i, calculer la prédiction $w_i = 0,5 \cdot z_i$ de leur nombre. Vérifier numériquement ensuite que :
$$\sum_{i=1}^{3} w_i = \sum_{i=1}^{3} a z_i = a \sum_{i=1}^{3} z_i$$
avec a = 50% = 0,5.

3. Chaque canton reçoit une subvention de la confédération pour créer $b = 300$ emplois. Déterminer le nombre v_i de chômeurs par canton restant après la création de ces emplois. Vérifier que :
$$\sum_{i=1}^{3}(z_i - b) = \sum_{i=1}^{3} z_i - 3b = \sum_{i=1}^{3} v_i$$

Exercice 1.4

1. Expliciter l'inconnue x des équations suivantes
   (a) $a + bx = c$
   (b) $(x - m)/s = z$
   (c) $d/x = g$
   (d) $\alpha^x = \gamma$

2. Déterminer la valeur de la solution de x des équations précédentes pour
   (a) $a = 10$, $b = -0.5$ et $c = 2$
   (b) $m = 10$, $s = 4$, $z = 1.28$
   (c) $d = 0.5$, $g = -2$
   (d) $\alpha = 10$, $\gamma = 1000$

Exercice 1.5

Représenter graphiquement dans le plan (x, y) les courbes

1. $y = 2 + 0.5\,x$, pour $x \in [-6; 4]$
2. $y = 1/x$, pour $x \in]0; 10]$

3. $y = \ln x$ pour $x \in]0; 10]$
4. $y = e^x$ pour $x \in [-5; 2]$

Première partie

Statistique descriptive

2 Données portant sur un caractère unique

Exercice 2.1

Un journaliste a relevé le nombre d'interviews qu'il a réalisé la première semaine d'un festival de cinéma. Les résultats obtenus sont les suivants :

jour	nombre d'interviews
L	20
M_a	5
M_e	10
J	15
V	10
S	20
D	20

On considère l'interview comme unité statistique et le jour comme caractère observé.

1. Déterminer les fréquences relatives des interviews par jour.
2. Est-il possible de représenter les résultats par un histogramme ? Pourquoi ?
3. Faire une représentation graphique appropriée.

On considère à présent le jour comme unité statistique et le nombre d'interviews comme variable observée.

1. Donner la répartition des jours selon le nombre d'interviews.
2. Faire l'histogramme de la distribution en considérant les classes [0 - 12,5[, [12,5 - 17,5[, [17,5 - 22,5].

Exercice 2.2

On dispose d'une série de données sur les demandes d'autorisations de construire dans 72 communes du Valais :

32	43	21	36	21	36	26	34
48	55	60	75	23	62	55	68
55	63	100	66	72	62	70	71
22	36	29	18	23	29	23	18
5	10	23	11	4	6	17	14
34	35	43	30	27	37	33	28
37	43	34	65	37	46	29	48
30	26	25	26	27	31	19	24
24	23	20	26	23	24	23	17

Établir une présentation tige-feuilles (steam and leaf) de ces données.

Exercice 2.3

Le tableau suivant présente les données groupées de 44 observations.

classe	n_k
$[0-2[$	15
$[2-10[$	17
$[10-60[$	12

Dessiner l'histogramme de cette distribution.

Exercice 2.4

Les valeurs suivantes correspondent au nombre de personnes à charge relevées dans un échantillon de dossiers d'employés

4 2 4 1 3 0 3 3 1 3 2

Déterminer la valeur des mesures statistiques suivantes de l'échantillon :

1. le mode
2. la moyenne
3. la médiane
4. l'étendue
5. les quartiles et l'écart interquartile
6. l'écart moyen
7. la variance
8. l'écart type

Exercice 2.5

On dispose d'une série de données x_i représentant les dépenses totales en santé, en % du PIB, dans 22 pays, pour l'année 1986 :

Europe	Allemagne	8.1	Finlande	7.5	Luxembourg	6.9
	Autriche	8.3	France	8.5	Norvège	6.8
	Belgique	7.1	Irlande	8.0	Pays-Bas	8.3
	Danemark	6.0	Islande	7.5	Suisse	7.7
	Espagne	6.0	Italie	6.7	Royaume-Uni	6.1
	Suède	9.1	Portugal	5.6		
Autres	Australie	6.9	Etats-Unis	10.9	Nouvelle-Zélande	8.3
	Canada	8.7	Japon	6.7		

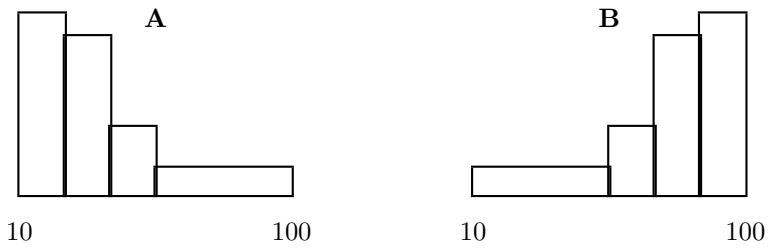
Source : OCDE en chiffres 1989

1. Que valent la médiane et les quartiles de cette distribution ? Tracer le box-plot correspondant.
2. Sachant que la somme des données des pays se trouvant en Europe est de 124.2, donner la moyenne $\bar{x}_E$ des pays européens.
   Calculer ensuite la moyenne $\overline{x}$ pour l'ensemble des données.
3. Sachant que la somme des carrés des écarts $(x_i - \overline{x})^2$ des pays se trouvant en Europe vaut 17.42, calculer la variance et l'écart type de la distribution totale.
4. Rassembler les données selon les classes suivantes : [0 - 3[; [3 - 6[; [6 - 9[; [9 - 12]. Calculer l'approximation de la moyenne et de la variance que l'on obtient à partir des données groupées. Comparer ces résultats avec la moyenne $\overline{x}$ et la variance calculées précédemment.

   Que pensez vous du choix des classes ?

Exercice 2.6

Considérez les deux distributions suivantes :



1. Quelle distribution a la plus forte dispersion ?
   a) A
   b) A et B ont la même dispersion
   c) B
   d) On n'a pas suffisamment d'information
2. Pour quelle distribution la moyenne est-elle supérieure à la médiane ?
   a) A
   b) B
   c) Moyenne et médiane sont égales dans les 2 cas
   d) On n'a pas suffisamment d'information

Exercice 2.7 Interpolation linéaire

Le tableau suivant donne les taux de participation à une votation fédérale pour cinq catégories (années) d'âges. Déterminez les valeurs manquantes par interpolation linéaire.

Age	Taux de participation
23	28%
30	?
32	40%
?	45%
44	55%

Exercice 2.8 Approximations sur données groupées

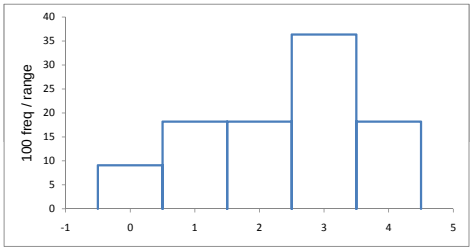
On dispose d'une série de données concernant le salaire mensuel en francs de 100 personnes. On organise les données en classes pour obtenir le tableau suivant :

classe k	effectif (n_k)
$[0 - 1'000[$	15
$[1'000 - 3'000[$	20
$[3'000 - 4'000[$	30
$[4'000 - 6'000[$	20
$[6'000 - 7'000]$	15

1. Approximer la médiane par interpolation linéaire.
2. Calculer l'approximation du salaire moyen.

Exercice 2.9 Asymétrie et aplatissement

Voici l'histogramme des données de l'exercice 2.4 sur le nombre de personnes à charge par employé.



Répondre aux questions suivantes sans faire de calculs.

1. Quel est le signe de l'indice d'asymétrie pour cette distribution ? Que devient cet indice si l'on remplace deux des valeurs '3' par des '1' et l'un des '4' par '0' ?
2. L'indice d'aplatissement (kurtosis) de la distribution vaut -0.6. Que devient-il suite au changement envisagé au point précédent ? Et si l'on avait simplement remplacé un '3' par '0' ?

Exercice 2.10 *

Au cours d'une enquête menée dans le canton de Berne, un échantillon de 10 exploitations agricoles a été prélevé. Nous donnons ci-dessous leur taille en hectares :

20 4 15 21 8 3 5 14 3 10

1. Calculer la moyenne, la médiane et les quartiles.
2. Construire le box-plot de ces données.
3. Commenter la dispersion en vous basant sur les informations calculées dans les deux premiers points.

Exercice 2.11 *

Nous donnons ci-dessous les tirages en semaine, pour février 1994, des 13 plus importants quotidiens romands :

56′494	31′222	60′066	31′339	91′137	22′160	44′070
58′642	32′544	29′476	15′312	35′418	25′000	

1. Construire le boxplot de ces données.
2. Regrouper les données ci-dessus dans les 3 classes suivantes : [0 - 30'000[, [30'000 - 60'000[, [60'000 - 100'000] et construire l'histogramme correspondant.
   Indication : Les hauteurs de l'histogramme seront construites à partir des fréquences.
3. Calculer l'approximation de la moyenne que l'on obtient à partir des données groupées.

Exercice 2.12 *

Nous donnons ci-dessous la plus haute température relevée chaque mois de 1993 à Genève :

Janvier	13, 8	Avril	23, 0	Juillet	32, 1	Octobre	22, 2
Février	7, 6	Mai	29, 0	Août	31, 8	Novembre	15, 5
Mars	20, 1	Juin	29, 1	Septembre	23, 5	Décembre	14, 4

1. Construire et commenter le boxplot de ces données.
2. Regrouper les données selon les classes [6-20[, [20-26[, [26-34], et construire l'histogramme.
3. Calculer l'approximation de la moyenne et de la variance que l'on obtient en considérant les données groupées.

Exercice 2.13 *

Le tableau suivant donne la répartition des personnes vivant seules selon le groupe d'âge. Les données sont en milliers et concernent la Suisse en 1960 et 1980.

âge	1960	1980
$[18-25[$	11	93
$[25-45[$	45	203
$[45-65[$	80	157
$[65-80[$	89	257

1. Préciser la nature de ces données : s'agit-il de données nominales (qualitatives) ou mesurables (quantitatives) ? Comment ont-elles été recueillies (sondage, données administratives, etc...) ?
2. Dessiner sur un même graphique l'histogramme de chacune des distributions 1960 et 1980 (travailler avec les fréquences relatives pour faciliter la comparaison).
3. Calculer les âges moyens et les écarts types de 1960 et 1980. Comparer et commenter les distributions.

Exercice 2.14 *

La présentation tige-feuilles ci-dessous, générée par SPSS, concerne la population des 37 communes genevoises ayant moins de 10'000 habitants .

```
POPULATI Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    11.00        0 .  33555677889
    10.00        1 .  0122677789
     7.00        2 .  0011669
     1.00        3 .  0
     1.00        4 .  6
      .00        5 .
     1.00        6 .  0
     6.00 Extremes    (>=6114)

 Stem width:  1000.0
 Each leaf:       1 case(s))
```

Comme indiqué, les tiges (stem) ont une largeur de 1000, ce qui signifie par exemple que la plus petite commune a environ $0.3 \cdot 1000 = 300$hab.

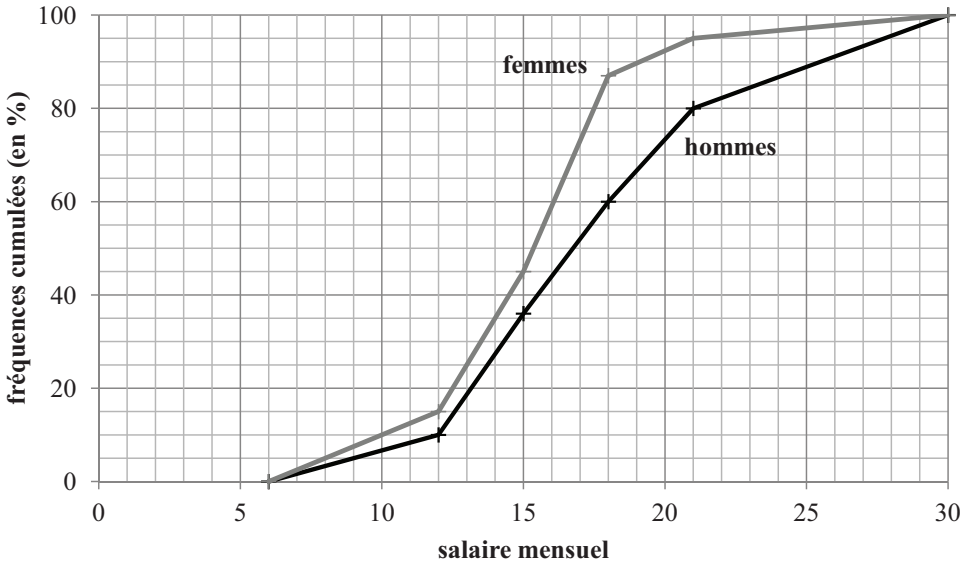
1. Déterminer la médiane et l'écart interquartile du nombre d'habitants.
2. Compléter le tableau suivant

k	population	effectif n_k	fréquence relative f_k	fréquence relative cumulée F_k
1	$[0, 1'000[$	.	.	.
2	$[1'000, 3'000[$	.	.	.
3	$[3'000, 10'000]$	.	.	.
	total	.	.	—

3. Déterminer la moyenne et l'écart type approximés qui résultent du tableau précédent.

Exercice 2.15 Fréquences cumulées

Vous êtes chargé de rédiger un rapport sur les différences salariales entre hommes et femmes parmi les cadres supérieurs d'une entreprise. Vous disposez des courbes de fréquences relatives cumulées des salaires masculins et féminins suivantes, où les salaires sont indiqués en milliers de francs.



1. Déterminer graphiquement la médiane des salaires masculins et féminins.
2. Quel est le pourcentage de femmes cadres supérieures qui gagnent moins de 15'000.- ? Et pour les hommes ?
3. Que pouvez-vous en conclure sur les différences salariales des hommes et des femmes cadres supérieures dans cette entreprise ?

3 Données portant sur deux ou plusieurs caractères

Exercice 3.1

Le tableau suivant donne la répartition de 100 personnes entre 15 et 30 ans selon leur âge et leurs goûts musicaux. On a donc deux caractères : "âge" et "goût musical". Chaque caractère a deux modalités.

goût musical / âge	Rock	Techno	Total
15 à 20 ans	18	35	53
21 à 30 ans	25	22	47
Total	43	57	100

1. Calculer les fréquences relatives qui caractérisent la distribution conjointe des individus selon les deux caractères.
2. Calculer la distribution (marginale) des goûts musicaux.
   Donner une présentation graphique de cette distribution (en utilisant le carré unitaire).
3. Calculer la distribution (marginale) de l'âge.
   Donner une présentation graphique de cette distribution.
4. Déterminer les pourcentages lignes (distributions conditionnelles du goût musical selon l'âge). Représentez ces distributions sur le graphique de la question 3. Les deux caractères sont ils indépendants ?
5. Déterminer les pourcentages colonnes (distributions conditionnelles de l'âge selon le goût musical).
6. Calculer le khi-deux de Pearson et comparer avec le seuil critique de 3.84. Calculer le v de Cramer. Commenter.

Exercice 3.2

A la suite des élections au Grand Conseil genevois du 14 novembre 1993, on s'intéresse aux résultats de quatre candidats représentatifs respectivement des partis Libéral, Démocrate chrétien, Socialiste et Alliance de Gauche. Le tableau suivant donne, pour trois quartiers de la ville de Genève, le nombre d'électeurs ayant voté pour ces quatre candidats ainsi que le pourcentage des voix obtenu par chacun.

		Candidats				
quartiers	nombre de votants	Brunschwig-Graf %	Maître %	Ziegler %	Spielmann %	total %
Champel	5230	37	35	18	10	100
Pâquis	2417	24	25	28	23	100
Mail-Jonction	3382	19	21	31	29	100
Total	11029	29	29	24	18	100

1. Quels sont les deux variables étudiées et quelle est leur nature (nominale, ordinale, discrète, continue) ?
2. Présenter graphiquement les données de ce tableau dans un carré unitaire. Commenter.

Exercice 3.3

Voici trois tableaux à double entrée. Dans quel cas les caractères A et B sont-ils indépendants ? Que vaut le khi-deux de Pearson pour ce cas ?

a)

		B	
A		2	8
		4	16

b)

		B	
A		2	8
		16	4

c)

		B	
A		2	8
		2	18

d) aucun

Exercice 3.4

Le tableau suivant présente les données relatives aux votations populaires de 1988 (pourcentage de “oui").

Canton	x pourcentage de oui à l'initiative sur la réduction de la durée de travail	y pourcentage de oui à l'initiative sur la limitation de l'immigration
Zurich	36,8	36,2
Berne	31,4	34,8
Genève	45,7	33,2
Jura	56,1	28,9
Tessin	59,4	37,8
Fribourg	38,3	33,5

1. Faire le diagramme de dispersion de ces données. Représenter le point moyen sur ce diagramme.
2. Quel est, selon ce graphique, le sens de l'association entre les deux séries x et y ?
3. Calculer la covariance entre x et y.
4. Centrer et réduire les données x et y, puis calculer les covariances des variables standardisées ainsi obtenues. Que représente ce résultat ?
5. Vérifier que la covariance calculée ci-dessus est égale à $cov(x, y)/(s_x\, s_y)$ où s_x et s_y sont les écarts types de x et y.
6. Au vu du résultat trouvé, vous paraît-il raisonnable de postuler une relation linéaire entre les deux séries ? Une telle hypothèse permettrait de prédire le pourcentage de oui à la seconde initiative pour un canton dont on ne connaîtrait que le résultat pour la première initiative.

Exercice 3.5 Lien entre longueur et altitude de cols

Le tableau suivant donne la répartition des 26 cols routiers des Alpes suisses selon la longueur du col et selon son altitude. On a donc deux caractères, à savoir “longueur" et “altitude".

		longueur du col			
		court (15-25 km)	moyen (25-35 km)	long (35-50 km)	TOTAL
altitude en mètres	1400-2000	3	2	5	10
	2000-2300	0	4	5	9
	2300-2600	1	4	2	7
TOTAL		4	10	12	26

1. Calculer la longueur moyenne approximée des cols que l'on obtient à partir de ces données groupées. Calculer également l'altitude moyenne approximée.

2. Calculer les fréquences relatives qui caractérisent la distribution conjointe des individus selon les deux caractères “longueur" et “altitude".

3. Calculer la distribution marginale selon la longueur du col. Donner une présentation graphique de cette distribution.

4. Déterminer la distribution de l'altitude (distribution conditionnelle) au sein de chaque catégorie de longueur (court, moyen, long). Compléter le graphique de la question 3 avec ces distributions conditionnelles.

5. Commenter ces distributions du point de vue de l'indépendance des caractères “altitude" et “longueur".

6. Calculer la covariance approximée entre l'altitude et la longueur des cols.

Exercice 3.6 Variances, covariances, et tableau de corrélations

Les caractéristiques de 30 enfants choisis au hasard parmi les élèves inscrits en 6ème primaire dans une école privée genevoise sont données par trois variables :

x_1 : âge en années
x_2 : moyenne annuelle en 5ème primaire
x_3 : nombre de frères et sœurs

On a calculé les quantités suivantes :

$$\sum_{i=1}^{30}(x_{1i}-\bar{x}_1)^2 = 11.18 \qquad \sum_{i=1}^{30}(x_{2i}-\bar{x}_2)^2 = 145.87$$
$$\sum_{i=1}^{30}(x_{3i}-\bar{x}_3)^2 = 48.8 \qquad \sum_{i=1}^{30}(x_{1i}-\bar{x}_1)(x_{3i}-\bar{x}_3) = 3.24$$

1. Déterminer les variances et covariances pour l'âge et le nombre de frères et sœurs.

2. Compléter le tableau (matrice) des corrélations ci-dessous et interpréter la valeur des coefficients.

	x_1	x_2	x_3
x_1	.	-0.11	.
x_2	.	.	.
x_3	.	-0.82	.

Exercice 3.7 Données centrées et normées

Soit deux séries (vecteurs colonnes) de données centrées et normées.

$$x = \begin{pmatrix} -0.22 \\ -0.67 \\ 0.22 \\ 0.67 \end{pmatrix} \qquad y = \begin{pmatrix} 0.11 \\ 0.57 \\ 0.11 \\ -0.80 \end{pmatrix}$$

1. Calculer la moyenne et la variance de chacune des séries. Que constatez-vous ?

2. On appelle produit scalaire et l'on note $x'y$ la somme des produits terme à terme $\sum_i x_i y_i$. Calculer ce produit et vérifier qu'il donne la corrélation entre x et y. Interpéter la corrélation trouvée.

Exercice 3.8 *

Le tableau suivant est tiré des résultats d'un sondage sur "la présence de femmes au Conseil d'Etat" publiés par la *Tribune de Genève* le 25 mai 1993.

	Opinion (en pourcents)				
	pas nécessaire	utile sans plus	indispensable	ne sait pas	Total
Femmes	2.8	28.3	63.8	5.1	100
Hommes	4.9	44.3	45.9	4.9	100

On admet que l'échantillon sondé contient 60% de femmes et 40% d'hommes.

1. Déterminer la distribution marginale des réponses, puis donner le tableau des distributions conditionnelles "lignes" que l'on aurait en cas d'indépendance entre le sexe et l'opinion.

	pas nécessaire	utile sans plus	indispensable	ne sait pas	Total
Femmes	.	.	.	.	100
Hommes	.	.	.	.	100
Marginale	.	.	.	.	100

2. Donner une présentation graphique des données. Commenter, notamment du point de vue de l'indépendance des deux caractères considérés.

Exercice 3.9 *

Le tableau ci-dessous donne le nombre de crises cardiaques frappant les hommes et les femmes de Cardioville selon leur classe d'âge.

	<30 ans	30-60 ans	>60ans
hommes	10	50	30
femmes	10	30	20

Donner une présentation graphique des données faisant apparaître la distribution (marginale) selon le sexe, et pour chaque sexe la répartition par classes d'âge.

Deuxième partie

Échantillonnage et probabilités

4 Inférence, incertitude et probabilités

Exercice 4.1

Le tableau suivant donne la répartition des participants à une enquête selon la catégorie socio professionnelle et le revenu familial :

		Revenu familial			
		Faible	Moyen	Élevé	Total
Occupation	Au Foyer	8	26	6	40
	Ouvrier	16	40	14	70
	Cadre	6	62	12	80
	Professionnel	0	2	8	10
	Total	30	130	40	200

On choisit une personne au hasard :

1. Trouver la probabilité que cette personne se classe dans les catégories suivantes :
   a) au foyer b) ouvrier c) cadre d) professionnel
2. Trouver la probabilité que le revenu familial soit :
   a) faible b) moyen c) élevé
3. Trouver la probabilité que cette personne soit :
   a) un cadre ayant un revenu élevé
   b) une personne au foyer ayant un revenu faible
   c) un professionnel ayant un revenu moyen
4. Que valent les probabilités conditionnelles suivantes :
   a) P(élevé | au foyer) b) P(cadre | élevé) c) P(ouvrier | faible)
5. Soient les événements A et B où A est l'événement "la personne choisie est un ouvrier" et B est l'événement "la personne choisie a un revenu faible".
   Que peut-on dire sur l'indépendance des événements A et B ?

Exercice 4.2

Soient les probabilités :

$p(A) = 1/2$	$p(B) = 1/3$	$p(C) = 1/4$
$p(A \text{ et } B) = 1/8$	$p(B \mid C) = 1/2$	$p(C \mid A) = 1/5$
$p(A \mid B \text{ et } C) = 2/3$	$p(A \text{ ou } B \text{ ou } C) = 49/60$	

1. En déduire les probabilités suivantes :

a)	$p(A \text{ et } C)$	*d)*	$p(A \mid B)$	*g)*	$p(A \text{ ou } B)$
b)	$p(B \text{ et } C)$	*e)*	$p(C \mid B)$	*h)*	$p(A \text{ ou } C)$
c)	$p(A \text{ et } B \text{ et } C)$	*f)*	$p((B \text{ et } C) \mid A)$		

2. a) Y a-t-il deux événements mutuellement exclusifs parmi les événements A, B et C ?
   b) Quelles paires d'événements ne sont pas indépendants ?

Exercice 4.3

Une crèche fréquentée par 40 garçons et 60 filles est frappée par une épidémie. Un jour où 15% des filles et 40% des garçons sont absents, on se propose d'osculter un enfant présent choisi au hasard. On considère les événements F « être une fille »et A « être absent », ainsi que leur négation $\bar{F}$ et $\bar{A}$.

1. Présenter les informations disponibles sous forme d'arborescence.
2. Déterminer et interpréter les probabilités $p(F, \bar{A})$, $p(\bar{F}, \bar{A})$ et $p(\bar{A})$.
3. Quelle est la probabilité que l'enfant choisi parmi les présents soit une fille ?

Exercice 4.4

Parmi les personnes ayant répondu à un questionnaire, 40% déclarent avoir visité Berne, 35% avoir visité Lucerne, et 15% affirment avoir visité les deux villes. Un voyage à Berne est offert par tirage au sort à l'un des répondants. Quelle est la probabilité que le gagnant ait déjà visité Lucerne si l'on sait qu'il n'a pas visité
Berne ?

a)	$0,500$	b)	$0,333$	c)	$0,250$
d)	0,375	e)	aucune de ces valeurs		

Exercice 4.5

Le tableau suivant donne le nombre de postes TV par ménage pour un immeuble de 10 appartements.

i	1	2	3	4	5	6	7	8	9	10
x_i	0	3	2	0	2	1	2	1	1	1

où i : désigne le ménage et x_i : le nombre de postes TV du ménage i.

1. Soit X la variable aléatoire représentant le nombre de postes TV d'un ménage choisi au hasard dans cet immeuble. Compléter le tableau de la distribution de probabilités de X.

X	0	1	2	3
$P(X = x)$	·	·	·	·

2. Si l'on trouve 3 appartements dans chacun des trois premiers étages de l'immeuble et un seul appartement au dernier étage, quelle est la probabilité de sélectionner un des ménages du 2ème étage ?
3. Sachant que le ménage choisi a au moins un poste TV. Quelle est la probabilité qu'il en ait trois ?
4. Calculer l'espérance mathématique et la variance de X.

Exercice 4.6

Voici trois distributions conjointes entre deux variables aléatoires X, Y.

a)

X \ Y	-1	1	
-1	$1/8$	$1/8$	$1/4$
1	$3/8$	$3/8$	$3/4$
	$1/2$	$1/2$	1

b)

X \ Y	-1	1	
-1	$1/4$	$1/4$	$1/2$
1	0	$1/2$	$1/2$
	$1/4$	$3/4$	1

c)

X \ Y	-1	0	1	
0	0	$1/2$	0	$1/2$
1	$1/4$	0	$1/4$	$1/2$
	$1/4$	$1/2$	$1/4$	1

1. Calculer dans chaque cas la covariance entre X et Y.
2. Les variables sont-elles indépendantes ? Justifier.

Exercice 4.7 *

A 10h15, 80% des fonctionnaires de l'université Karl Marx de Budapest, se trouvent à la cafétéria. La population de l'université est composée de 30% de fonctionnaires. Or, parmi toutes les personnes qui fréquentent ou travaillent à l'université, 28% sont à la cafétéria à 10h15. Quelle est la probabilité qu'une personne choisie au hasard dans la cafétéria à 10 h15 soit un fonctionnaire ?

Exercice 4.8 *

Nous avons recueilli les renseignements suivants concernant les étrangers résidant à Genève :

- Parmi les étrangers originaires d'Europe, 55,2 % proviennent de pays limitrophes de la Suisse.
- Les étrangers en provenance de pays limitrophes représentent 44,4 % des étrangers.
- 3,3 % des étrangers sont originaires d'Afrique.

1. Calculer la probabilité qu'un étranger résidant à Genève soit originaire d'un pays africain ou d'un pays limitrophe de la Suisse.
2. Calculer la probabilité qu'un étranger résidant à Genève soit originaire d'Europe.

Exercice 4.9 *

En science politique, 75% des garçons et 83% des filles suivent les Compléments de statistique. Les filles constituent par ailleurs 68% de l'effectif total des étudiants en science politique. On choisit un étudiant au hasard.

1. Quelle est la probabilité qu'il s'agisse d'une fille et qu'il suive les Compléments ?
2. Quelle est la probabilité qu'il s'agisse d'un garçon et qu'il suive les Compléments ?
3. Quelle est la probabilité qu'il suive les Compléments ?

Exercice 4.10 *

Soit X la fréquentation d'un restaurant un jour choisi au hasard. Sa distribution de probabilité est :

x	100	110	120
$P(X = x)$	0.2	.	0.5

Compléter le tableau, puis calculer l'espérance mathématique $\mathrm{E}(X)$ et l'écart type σ de la variable X.

Exercice 4.11 *

Mon ami Podz, qui n'est pas un habile cycliste, a tendance à être ivre quelques soirs par semaine. On sait que cet ami, lorsqu'il est en soirée, chute un soir sur trois et qu'il est ivre trois soirs sur sept. On sait en outre qu'il est ivre et qu'il tombe la même soirée deux soirs sur sept.

1. Préciser les deux événements simples considérés et leurs contraires. Donner le tableau des probabilités simples et conjointes.
2. J'ai rendez-vous ce soir avec cet ami. Or, celui-ci ayant terminé ses examens, je sais qu'il va être ivre. Quelle est la probabilité qu'il chute à vélo ?
3. La consommation d'alcool de mon ami influence-t-elle son habileté à vélo ?

Exercice 4.12 *

D'après une étude réalisée sur une fameuse plage de France, on a pu constater que 70% des femmes blondes ont eu un coup de soleil, alors que seulement 40% de brunes en ont eu un.
Parmi ces femmes, 2/5 étaient blondes ; les autres brunes.

1. Présenter les informations données sous forme d'un arbre de probabilités.
2. Peut-on affirmer que le fait d'avoir eu un coup de soleil est indépendant de la couleur des cheveux ?
3. En choisissant une femme au hasard, quelle est la probabilité qu'elle soit brune et qu'elle n'ait pas de coup de soleil ?
4. Quelle est la probabilité qu'une femme ait eu un coup de soleil ?
5. Quelle est la probabilité qu'une femme ayant eu un coup de soleil soit blonde ?

5 Échantillonnage

Exercice 5.1

Quatre individus d'un groupe sont interrogés sur leur âge. Le résultat obtenu est le suivant :

Individu i	1	2	3	4
Age i	35	20	15	20

Soit X l'âge d'un individu choisi au hasard.

1. Donner la distribution des probabilités de la variable X et calculer l'espérance mathématique et la variance de cette variable.
2. On désire se faire une idée du groupe en n'examinant que deux individus. On tire un échantillon avec remises de taille 2 parmi les quatre individus.
   (a) Donner tous les échantillons de taille 2 possibles ainsi que leur probabilité.
   (b) Calculer pour chaque échantillon la moyenne $\bar{x}_r$, la variance s_r^2 et le minimum $x_{\min}$.
   (c) Donner le tableau des probabilités de la moyenne $\bar{X}_r$, de la variance S_r^2 et du minimum $X_{\min}$ d'un échantillon, choisi au hasard.
   (d) Calculer $\mathrm{E}(\bar{X}_r); \mathrm{Var}(\bar{X}_r)\,; \mathrm{E}(S_r^2)$ et $\mathrm{E}(X_{\min})$.
3. A présent on choisit l'échantillon de taille 2 sans remises.
   (a) Donner tous les échantillons de taille 2 possibles ainsi que leur moyenne $\bar{x}_s$.
   (b) Donner le tableau des probabilités de la moyenne $\bar{X}_s$ d'un échantillon choisi au hasard.
   (c) Calculer $\mathrm{E}(\bar{X}_s)$ et $\mathrm{Var}(\bar{X}_s)$. Comparer ces résultats avec $\mathrm{E}(\bar{X}_r)$ et $\mathrm{Var}(\bar{X}_r)$ trouvés en 2.(d). Commenter.

Exercice 5.2

On s'intéresse à une grandeur X qui est distribuée comme suit dans une population de taille 10 :

x	-4	2	4
$p(X=x)$	1/10	5/10	.

On se propose de tirer un échantillon de taille 7 avec remises dans cette population.

1. Calculer la variance $\mathrm{Var}\,\bar{X}$ de la moyenne d'échantillon $\bar{X}$.
2. Déterminer l'espérance mathématique $\mathrm{E}(S^2)$ de la variance d'échantillon $S^2 = 1/n \sum (X_i - \bar{X})^2$.

On se propose, à présent, de tirer un échantillon de taille 7 sans remises dans cette population.

3. Calculer l'espérance mathématique $\mathrm{E}(\bar{X}_s)$
4. Calculer la variance $\mathrm{Var}\,\bar{X}_s$ de la moyenne d'échantillon $\bar{X}_s$.

Exercice 5.3 *

Le comité d'organisation d'une manifestation sportive a reçu des propositions de partenariat d'un groupe de 4 sponsors. Chacun est disposé à verser une certain montant (en milliers de francs) :

Sponsor i	1	2	3	4
Montant i	12	8	4	8

Comme le comité est surchargé, il veut se faire une idée du groupe en ne s'intéressant qu'à 3 différents sponsors.

1. Donner tous les sous-groupes possibles ainsi que pour chacun la somme totale y des montants offerts.
2. Donner le tableau des probabilités de la somme Y d'un sous-groupe choisi au hasard.
3. Calculer l'espérance et la variance de la somme Y d'un sous-groupe tiré au hasard.

Exercice 5.4 *

Durant les soldes, un magasin de disques propose à la clientèle un grand panier contenant en proportions égales des disques à 10, 15, 20 et 30 francs. Nous savons de plus que le panier contient un total de 200 disques.

1. Calculer l'espérance et la variance du prix des disques.
2. On prélève un échantillon de 20 disques par tirages sans remises. En supposant qu'au moment du tirage le prix moyen des disques restant dans le panier est $\mu = 21$ francs avec une variance de $\sigma^2 = 60$, calculer l'espérance $\mathrm{E}(\bar{X})$ et la variance $\mathrm{Var}(\bar{X})$ du prix moyen $\bar{X}$ de l'échantillon.

Exercice 5.5 *

Un porte-monnaie contient 2 pièces de 5 francs, 1 pièce de 2 francs et 1 pièce de 1 franc. On se propose de tirer un échantillon de deux pièces par tirages sans remises.

1. Enumérer les différents échantillons possibles.
2. Soit X le nombre de pièces de 5 francs dans l'échantillon. Donner la distribution de probabilité de X.
3. Sachant qu'on a obtenu une pièce de 5 francs au premier tirage ($X_1 = 5$), quelle est la probabilité d'obtenir une pièce de 1 franc au deuxième ($X_2 = 1$) ? Les deux événements "$X_1 = 5$" et "$X_2 = 1$" sont-ils indépendants ?

Exercice 5.6 *

Parmi trois sujets soumis à votation, le nombre de sujets acceptés par une commune choisie au hasard est distribué comme suit

x	0	1	2	3
$P(X = x)$	.3	.4	.1	.2

Soit $\bar{X}$ et S^2 la moyenne et la variance d'un échantillon de 8 communes choisies au hasard par tirages avec remises. Calculer $\mathrm{Var}(\bar{X})$ et $\mathrm{E}(S^2)$.

Exercice 5.7 *

Soit une population de *trois* mères de famille ayant respectivement 1, 5 et 3 enfants. Soit X le nombre d'enfants d'une mère choisie au hasard dans cette population. On désire prélever un échantillon de taille 2 par tirages au hasard *avec* remises. Soit $\bar{X}$ la moyenne de l'échantillon. Déterminer.

a) la probabilité $P(\bar{X} > 3)$ b) la variance de $\bar{X}$.

6 Distributions continues

Exercice 6.1

Soit Z une variable aléatoire qui suit une loi normale, centrée et réduite $N(0,1)$.

1. Déterminer la probabilité des événements :
   a) $Z < 2.14$ b) $Z < -2.14$
   c) $Z \in [-2.14 \quad ; \quad 2.14]$
2. Pour quelle valeur de a a-t-on $P(Z < a) = p$ lorsque p prend les valeurs suivantes :
   a) $p = 0,8686$ b) $p = 0,9719$ c) $p = 0,2912$ d) $p = 0,7$

Soit X une variable aléatoire qui suit une loi normale $N(2,1)$.

3. Déterminer les probabilités suivantes :
   a) $P(X < 3,98)$ b) $P(X < 1,5)$ c) $P(X \in [2;3])$
4. Pour quelle valeur de a a-t-on :
   a) $P(X < a) = 0,7454$ b) $P(2 - a < X < 2 + a) = 0,7458$

Soit Y une variable aléatoire qui suit une loi normale $N(2,9)$.

5. Déterminer les probabilités suivantes :
   a) $P(Y < 4,25)$ b) $P(Y < 1,5)$
6. Pour quelle valeur de a a-t-on :
   a) $P(Y < a) = 0,64$ b) $P(a < Y < 2,5) = 0,3788$
   c) $P(2 - a < Y < 2 + a) = 0,9$

Exercice 6.2

Le nombre de spectateurs aux matchs d'un club de football est supposé suivre une loi normale $N(\mu = 24'000, \sigma^2 = 16'000'000)$. L'assistance maximale a été l'an passé de 31'000 spectateurs. Quelle est la probabilité d'atteindre au moins le 90% de ce maximum lors d'un match choisi au hasard ?

Exercice 6.3

Les poids de 600 bonbons suivent une distribution normale dont la moyenne est de 68 grammes et l'ecart-type de 3 grammes. Combien de bonbons ont un poids :

1. supérieur à 72 grammes,
2. inférieur à 64 grammes,
3. compris entre 65 et 71 grammes,
4. égal à 68 grammes.

Exercice 6.4

Considérons la variable aléatoire T_{15} qui suit une loi de Student à 15 degrés de liberté. Trouvez les différents seuils a tels que :

1. $P(T_{15} < a) = 0,9$
2. $P(T_{15} < -a) = 0,9$
3. $P(T_{15} < a) = 0,05$
4. $P(-a < T_{15} < a) = 0,99$

Refaites les calculs mais avec une loi de Student à 10 degrés de liberté.

Exercice 6.5

Déterminer les seuils a et b tels qu'on ait pour une variable Q_{10} distribuée selon une loi du khi-2 à 10 degrés de liberté :

1. $p(Q_{10} < a) = 0.95$
2. $p(Q_{10} < b) = 0.1$
3. Que vaut la probabilité $p(b < Q_{10} < a)$?

Exercice 6.6 *

100 skieurs professionnels ont descendu une célèbre piste de ski des alpes suisses. Le temps X en secondes que chacun a réalisé suit une distribution normale $N(140; 100)$.

1. Déterminer le nombre de skieurs réalisant un temps compris entre 130 et 145 secondes.
2. Quel temps t faut-il réaliser pour être parmi les 20 premiers ?

Exercice 6.7

1. Le service statistique du Ramtchaka dispose d'un recensement complet des salaires. Selon celui-ci, la moyenne des salaires de l'ensemble des employés de la branche de la vente s'élève à $\mu = 175$ pesos par semaine et la variance des salaires est de $\sigma^2 = 160$. Toujours selon le service des statistiques, on peut raisonnablement faire l'hypothèse que les salaires sont normalement distribués. Qu'elle est la probabilité d'interroger quelqu'un ayant un salaire en dessous du seuil de pauvreté fixé à 170 pesos ?
2. Selon leur expert en statistique, la moyenne $\bar{X}$ d'un échantillon de 8 personnes (i.i.d.) suit de ce fait une loi normale $N(\mu = 175, \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = 20)$. Cette affirmation vous semble-t-elle correcte ?
3. En supposant que cette affirmation est correcte, qu'elle est la probabilité d'obtenir un échantillon avec une moyenne inférieure au seuil de pauvreté (170) ?
4. Quelle est la probabilité que l'écart entre la moyenne d'un échantillon et μ soit supérieur à 10 (en valeur absolue) ?

Troisième partie

Statistique inférentielle

7 Estimation

Exercice 7.1

Soit Y le nombre de personnes qui assistent à la représentation d'une pièce de théâtre.

1. Pour 5 jours de représentation au hasard on a observé :

i	1	2	3	4	5
y_i	300	280	290	310	295

   Estimer l'espérance μ de Y par la médiane Y_{med} de l'échantillon, puis par la moyenne $\bar{Y}$.

2. Pour 4 nouveaux jours choisis au hasard on a observé :

i	6	7	8	9
y_i	305	318	290	280

   Recalculer ces deux estimations de μ en considérant l'échantillon formé par l'ensemble des 9 observations. Commenter.

3. Pour les cinq premières observations on trouve :

$$\sum_{i=1}^{5} y_i^2 = 435'625$$

   et pour l'ensemble des 9 observations :

$$\sum_{i=1}^{9} y_i^2 = 792'274.$$

   Donner l'estimation non biaisée de la variance de Y_i, ainsi que l'estimation non biaisée de l'écart type de la moyenne $\bar{Y}$ dans chacun des deux cas considérés.
   Commenter.

4. Supposons maintenant qu'à la suite d'une erreur le nombre d'entrées du 9ème jour a été mal relevé. De ce fait nous avons :

i	1	2	3	4	5	6	7	8	9
y_i	300	280	290	310	295	305	318	290	28

   Recalculer les deux estimations de μ. Lequel des deux estimateurs utilisés est le plus pertinent ?

Exercice 7.2

Un sondage auprès de personnes choisies au hasard a permis d'établir l'estimation $\hat{p} = 65\%$ de la proportion p d'individus favorables à la décision du conseil de sécurité de l'ONU concernant le retrait des casques bleus de Somalie.

Dans le cadre de ce sondage une variable aléatoire X_i a été associée à chaque individu i. Cette variable prend les valeurs suivantes :

$$X_i = \begin{cases} 1 & \text{si l'individu est favorable} \\ 0 & \text{sinon} \end{cases}$$

1. Sachant qu'un total de 700 personnes ont répondu favorablement, sur combien de personnes portait le sondage ?
2. Donner le tableau de la distribution de la variable X_i.
   Calculer $\mathrm{E}(X_i)$ et $\mathrm{Var}(X_i)$ en fonction de p.
3. Donner une estimation non biaisée de p.
4. Proposer une estimation de $\sigma^2 = \mathrm{Var}(X_i)$. En déduire celle de l'écart type $\sigma_{\hat{P}}$ de l'estimateur utilisé en 3.

Exercice 7.3

On dispose de deux estimateurs pour estimer le nombre moyen de personnes qui regardent une émission de télévision. Le premier $\hat{\mu}_1$ est non biaisé, le deuxième $\hat{\mu}_2$ a un biais de 2 personnes ($\mathrm{Biais}(\hat{\mu}_2) = 2$).

Sachant que la variance de $\hat{\mu}_1$ est égale à 20 et que celle de $\hat{\mu}_2$ vaut 5, quel est l'estimateur le plus efficace ?

Exercice 7.4

Le temps X mensuel en minutes consacré à la lecture de magazines est supposé suivre une loi normale $N(\mu, \sigma^2)$. Voici les valeurs observées pour huit individus choisis au hasard :

330 547 461 380 412 356 502 484

1. Calculer pour cet échantillon, le temps moyen $\bar{x}$ de lecture. S'agit-il d'une estimation non biaisée ? Proposer d'autres estimations de μ.
2. Donner un intervalle de confiance à 90% pour μ en supposant que σ^2 vaut 5'850.
3. Calculer la variance s^2 du temps de lecture dans l'échantillon. En déduire une estimation non biaisée de σ^2.
4. Comparer cette estimation avec la valeur supposée de σ^2 utilisée en 2. Déterminer un intervalle de confiance à 95% pour σ^2. Commenter.
5. Déterminer l'intervalle de confiance à 90% pour μ sans tenir compte de l'hypothèse sur σ^2. Comparer avec l'intervalle trouver en 2.

Exercice 7.5

Une compagnie d'aviation s'intéresse à la distance moyenne en kilomètres μ parcourue annuellement par les voyageurs qui se déplacent en première classe. Pour un échantillon de taille 100, on a déterminé que la moyenne $\bar{x}$ des distances parcourues vaut 76'400 et l'écart-type s des distances 5'250.

La taille de l'échantillon étant grande, utilisez le Théorème de la limite centrale pour trouver un intervalle de confiance pour μ à 95% (on admettra que la variance de la population σ^2 est égale à la variance de l'échantillon).

Exercice 7.6

Le diamètre en centimètres X d'une pizza suit une loi $N(\mu, \sigma^2)$. On admet que σ^2 vaut 9. Pour un échantillon de 15 pizzas on a calculé la moyenne $\bar{x} = 30$.

1. Calculer l'intervalle de confiance à 95% pour μ.
2. On observe 11 pizzas supplémentaires. La moyenne du nouvel échantillon de taille 26 reste égale à 30.
   (a) Comment évolue la taille de l'intervalle de confiance à 95% pour μ ?
   (b) Que devient l'intervalle si l'on retient un degré de confiance de 90% ? Commenter.

Exercice 7.7

Afin d'inférer sur une proportion p d'individus favorables à l'ouverture des commerces jusqu'à 21 heures, on désire procéder à un sondage. Pour cela on s'adresse à n personnes choisies au hasard dans la population genevoise. On associe à chaque individu i une variable aléatoire X_i qui prend la valeur 1 si l'individu est favorable et 0 sinon.

1. Discuter l'indépendance des X_i dans le cas où l'on procède à des tirages sans remises. Donner le tableau des probabilités de X_i. Exprimer $\mathrm{E}(X_i)$ et $\mathrm{Var}(X_i)$ en fonction du paramètre p.

Pour un échantillon de taille $n = 500$, on obtient 300 avis favorables.

2. Proposer un estimateur $\hat{P}$ absolument correct de p (utiliser le fait que $\mu = p$), puis donner l'estimation résultant de l'échantillon.
3. On admet que $\hat{P}$ est distribué selon une loi normale. Cette hypothèse est-elle raisonnable ? Justifier.
4. On peut montrer qu'ici la variance de l'échantillon est $s^2 = \hat{p}(1 - \hat{p})$. La vraie proportion p n'est pas connue, et par conséquent on ne connaît pas non plus la vraie variance σ^2. Pour simplifier, on retiendra ici l'hypothèse que σ^2 est égal à la variance d'échantillon s^2.
   Déterminer l'écart type de $\hat{P}$, puis donner un intervalle de confiance à 90% pour p. Commenter.

Exercice 7.8 *

Un mélomane vient d'acheter 7 disques. Leurs prix sont donnés ci-dessous :

89 32 48 30 34 28 69

1. On fait l'hypothèse que les prix sont distribués selon une loi normale $N(\mu, \sigma^2)$. Cette hypothèse est-elle pertinente ?
2. En supposant $\sigma^2 = 400$, calculer un intervalle de confiance à 80 % pour μ.

Un second mélomane a aussi acheté 7 disques et en a noté les prix :

12 19 10 30 10 30 25

3. En supposant $\sigma^2 = 400$, calculer un intervalle de confiance à 80 % pour μ.
4. Comparer et commenter les intervalles de confiance obtenus aux points 2 et 3.

Exercice 7.9 *

Soit une population de variance $\sigma^2 = 400$ dont on se propose d'estimer la moyenne μ inconnue. On considère deux estimateurs : $\bar{X}_{20}$ la moyenne d'un échantillon de taille 20 obtenu par tirages avec remises, et la moyenne $\bar{X}_Q$ de 2'000 réponses à un questionnaire. Le biais de $\bar{X}_Q$ est évalué à 5. Calculer l'erreur quadratique moyenne des deux estimateurs. Commenter.

Exercice 7.10 *

Pour estimer un paramètre θ, on dispose de deux estimateurs $\hat{\Theta}_1$ et $\hat{\Theta}_2$. On a

$$\begin{array}{lcllcl} \mathrm{E}(\hat{\Theta}_1) & = & \theta & \mathrm{Var}\,\hat{\Theta}_1 & = & 100 \\ \mathrm{E}(\hat{\Theta}_2) & = & \theta + 9 & \mathrm{Var}\,\hat{\Theta}_2 & = & 10 \end{array}$$

Lequel est le plus efficace et pourquoi ?

Exercice 7.11 *

Soit un échantillon de 4 licenciés en sciences sociales, pour lesquels on dispose des données suivantes :

âge	22	30	21	26
années d'études en licence	3	4	3	5

5. Calculer la corrélation entre l'âge et le nombre d'années d'études des quatre licenciés choisis.
6. Sachant que la variance de la durée d'études est $\sigma^2 = 0.96$, construire un intervalle de confiance à 90% pour le nombre moyen d'années nécessaires pour obtenir la licence. Quelle hypothèse doit-on faire ? Est-elle pertinente ?

Exercice 7.12 *

On s'intéresse à la proportion p de suisses favorables à l'adhésion à la CE. Sur un échantillon de 730 personnes, 422 se déclarent favorables. A chaque réponse i on associe une variable X_i, qui prend la valeur 1 si la réponse est positive et 0 sinon.

Déterminer un intervalle de confiance pour p à 95 % en admettant que la variance des X_i vaut 0,25.

8 Test d'hypothèses

Exercice 8.1

Les données suivantes concernent le taux d'acceptation X d'une initiative d'un échantillon de communes en Suisse.

$$\begin{array}{cccccc} 62,3 & 44,4 & 49,2 & 63,3 & 47,6 & 60,1 \\ 37,4 & 55,8 & 57,5 & 58,3 & 56,2 & 54,3 \end{array}$$

On suppose que le taux X est distribué normalement et que l'écart type des taux pour l'ensemble des communes suisses vaut 5.
Remarque : L'unité statistique est ici le canton pour lequel on a un taux, et non l'électeur pour qui on aurait une variable binaire accepte/refuse. On n'est donc pas dans le contexte particulier du test d'une proportion, mais dans le cas général du test d'une moyenne.

1. Au seuil de signification de $\alpha = 0,05$ la moyenne μ des taux est-elle significativement supérieure à 50% ? Commenter.
   Formellement, on testera :

$$\begin{array}{ll} H_0 : \mu = \mu_0 = 50 & \text{contre} \\ H_1 : \mu = \mu_1 > 50 & \end{array}$$

2. Soit un échantillon de taille 100 pour lequel on a obtenu la même moyenne que pour l'échantillon précédent de taille 12. Déterminer la région critique dans ce cas et comparer avec celle obtenue en 1. Commenter.

Exercice 8.2

Le temps consacré au repas de midi par les employés d'une entreprise qui prennent leurs repas au restaurant est supposé suivre une loi normale $N(\mu, \sigma^2)$ avec $\sigma^2 = 144$.

1. On veut tester l'hypothèse $H_0 : \mu = \mu_0 = 45$ contre l'hypothèse $H_1 : \mu = \mu_1 = 50$ avec un échantillon de taille sept. Pour un risque de première espèce $\alpha = 10\%$, la région d'acceptation est :

   (a) $\{\bar{x} \mid \bar{x} < 50,8\}$
   (b) $\{\bar{x} \mid 36,1 < \bar{x} < 53,9\}$
   (c) $\{\bar{x} \mid \bar{x} < 52,1\}$
   (d) $\{\bar{x} \mid 35,5 < \bar{x} < 54,9\}$
   (e) Aucun de ces ensembles.

2. Pour un échantillon observé donné, le test précédent nous a conduit à accepter H_0. Quelle serait la conclusion pour un risque $\alpha = 0,05$?

   (a) Acceptation de H_0.
   (b) Rejet de H_0.
   (c) On ne peut pas se prononcer sans refaire le test.

3. Toujours pour le même test, on considère une seconde hypothèse alternative $H_2 : \mu = \mu_2 = 47$. La région d'acceptation est plus grande lorsque l'hypothèse alternative est :

   (a) H_1
   (b) H_2
   (c) La région d'acceptation est la même.
   (d) On n'a pas suffisamment d'information.

Exercice 8.3

La durée de vie moyenne d'un échantillon de 10 tubes néons est de 157 heures avec un écart type de 12 heures. On admet que la durée de vie X d'un tube suit une loi normale $N(\mu, \sigma^2)$ et l'on suppose que l'écart type de la population est égal à celui de l'échantillon.

1. tester l'hypothèse $\mu = 160$ relativement à l'hypothèse $\mu \neq 160$ heures, avec un niveau de signification
   a) $\alpha = 5\%$ b) $\alpha = 1\%$
2. On se propose de calculer la puissance du test avec un niveau de signification $\alpha = 5\%$ pour
   a) $\mu_1 = 150$ b) $\mu_1 = 155$ c) $\mu_1 = 160$
   d) Par symétrie ($\mu_0 = 160$), donner les puissances pour $\mu_1 = 165$ et $\mu_1 = 170$.

Exercice 8.4

Le 7 mars 1993, les Suisses ont accepté par 73,73% de oui la réouverture des casinos. On aimerait déterminer si les cantons qui n'ont pas de frontière avec l'étranger ont été plus favorables à cette décision que la moyenne suisse. A cet effet, on dispose du pourcentage x de oui des 10 cantons et demi-cantons sans frontière avec l'étranger :

Lucerne	76	Glaris	74
Uri	76	Zoug	74
Schwytz	75	Fribourg	75
Obwald	75	Appenzell R.-Ext.	77
Nidwald	78	Appenzell R.-Int.	77

$\sum_{i=1}^{10} x_i^2 = 57'321$

On admet que le pourcentage X de oui par canton est distribué selon une loi $N(\mu, \sigma^2)$.

1. En supposant σ^2 inconnu, tester, au seuil de signification $\alpha = 0,05$, si le pourcentage de oui de ces 10 cantons est supérieur à celui de l'ensemble de la Suisse. Formellement, on testera :

$$H_0 : \mu = \mu_0 = 73,73 \text{ contre}$$
$$H_1 : \mu = \mu_1 > \mu_0$$

2. En supposant $\sigma^2 = 1,8$, refaites le test précédent.
3. Tester au seuil $\alpha = 5\%$ si l'hypothèse faite en 2 sur σ^2 est raisonnable.

Exercice 8.5

Sur l'emballage des biscuits 'Madame K' le poids du contenu indiqué est de 180 grammes. Afin de contrôler cette affirmation, on a pesé le contenu de 400 paquets choisis au hasard dans un très grand lot de cette production. Le poids total des 400 paquets est égal à 71'440 grammes et la somme des carrés des écarts à la moyenne vaut 123'594.
A l'aide d'un test d'hypothèses vérifier avec un risque $\alpha = 5\%$) si, en moyenne, le poids d'un paquet de biscuit correspond bien au poids indiqué sur l'emballage.

Exercice 8.6

Le propriétaire d'une vidéothèque affirme que le nombre de vidéocassettes louées quotidiennement est de 1500. Un employé du magasin veut vérifier l'exactitude de l'affirmation de son patron. La moyenne d'un échantillon de 36 jours choisis au hasard est de 1450 vidéocassettes louées par jour. On se propose de tester l'affirmation au seuil $\alpha = 0,05$ en supposant que les ventes suivent une loi normale et que σ vaut 120.

1. Quelle statistique proposez-vous d'utiliser ?
2. L'hypothèse que l'on désire vérifier revient à effectuer le test suivant :

$$\begin{array}{ll} H_0 : \mu = \mu_0 = 1500 & \text{contre} \\ H_1 : \mu = \mu_1 < \mu_0 & \end{array}$$

   Déterminer numériquement la région critique R pour la statistique considérée.
3. Quelle devrait être la conclusion de l'employé ?
4. Calculer la puissance du test pour les valeurs suivantes de μ_1 : $\mu_1 = 1440$; $\mu_1 = 1460$; $\mu_1 = 1480$.
5. Tracer la courbe d'efficacité du test.

Exercice 8.7

Sur la base d'un sondage réalisé en 2000 au lendemain de la votation sur les accords bilatéraux auprès de 752 personnes, l'Hebdo a publié des estimations de la proportion de suisses favorables à l'adhésion de la Suisse à l'UE ainsi que de la proportion de favorables à l'adhésion à l'ONU.

1. En admettant que l'échantillon a été obtenu par un tirage purement aléatoire, déterminer la demi-longueur (marge d'erreur) d'un intervalle de confiance à 90% conservateur (utilisant la variance maximale) pour les proportions publiées.
2. Déterminer cette même marge d'erreur en utilisant l'estimation de la variance pour la proportion p des personnes favorables à l'UE, sachant que 385 des personnes interrogées se sont déclarées favorables.
3. Tester l'hypothèse $H_0 : p = 50\%$ contre $H_1 : p > 50\%$ en considérant successivement les deux marges d'erreur calculées précédemment. A quel risque de première espèce correspondent ces marges d'erreur ? La majorité favorable des personnes interrogées est-elle statistiquement significative ?

Exercice 8.8 *

Une association de consommateurs s'intéresse à la quantité de lipides (graisses) dans une boisson light (le terme "light" suppose que la boisson contient peu de graisses). Sur l'étiquette d'une bouteille de 1 litre, on peut lire que la quantité de lipides est égale à 60 décigrammes par litre. Après avoir analysé 100 bouteilles d'un litre de cette boisson, on a mesuré un total de 6100 décigrammes de lipides. De

plus, on sait que l'écart type de la quantité de graisses par bouteille à la production est de $\sigma = 7,5$ décigrammes.
L'association de consommateurs veut tester si les bouteilles ne contiennent pas trop de graisses. Formellement, on effectue un test au seuil d'efficacité $\alpha = 10\%$, dont l'hypothèse nulle est :

$$H_0 : \mu = \mu_0 = 60$$

Les calculs se feront en décigrammes.

1. Proposez l'hypothèse alternative adéquate pour le but proposé.
2. Déterminez la région d'acceptation.
3. Donnez la conclusion du test et commentez.
4. Calculez la puissance du test pour $\mu_1 = 60$, $\mu_1 = 61$.

Exercice 8.9 *

Un disquaire affirme que le prix de ses disques suit une loi normale $N(\mu, \sigma^2)$ avec $\mu = 26$ francs et $\sigma^2 = 245$. Pour vérifier ses dires, on tire l'échantillon de prix suivant :

19 22 30 52 48 32 36 33

1. Effectuer le test suivant avec un risque de première espèce $\alpha = 5\%$:

$$\begin{array}{ll} H_0 : \mu = \mu_0 = 26 & \text{contre} \\ H_1 : \mu = \mu_1 > 26 & \end{array}$$

2. Refaire le même test au seuil $\alpha = 10\%$.
3. Comparer et commenter les résultats obtenus aux points 1 et 2.

Exercice 8.10 *

On dispose des moyennes générales d'un échantillon de dix étudiants choisis au hasard parmi ceux de première année de sciences sociales.

3.9 2.7 3 5 4 5.5 5.8 2.9 3.5 3.5

Selon certaines rumeurs, la moyenne générale se situerait aux alentours de 3,5. On désire tester cette hypothèse contre l'hypothèse alternative d'une moyenne μ supérieure à 3,5 avec un risque α de 10%.

1. Faire le test en supposant la variance σ^2 connue et égale à 1.1. On précisera notamment formellement les hypothèses nulle H_0 et alternative H_1, la statistique utilisée et sa distribution, la forme de la région critique et le (ou les) seuils critique(s).
2. Calculer la puissance du test précédent pour $\mu_1 = 4$. Commenter.
3. Refaire le test, sans faire d'hypothèse sur la valeur de σ^2. Préciser la statistique utilisée et sa distribution. Commenter.

Exercice 8.11 *

Dans un parti donné, la proportion Y de députés qui ne respecte pas les consignes du parti lors de votes au parlement est supposée suivre une loi normale $N(p = 0,2; \sigma^2 = 0,16)$. Pour 10 votes choisis au hasard, la moyenne des proportions observées est $\bar{y} = 0,18$. Tester, au seuil de 5%($\alpha = 0,05$), si la proportion moyenne p est significativement inférieure à 0,2.

Exercice 8.12 *

Selon les données passées d'un centre de recrutement militaire le poids des recrues suit une loi normale $N(\mu = 80 \text{ kg}, \sigma^2 = 100)$. Le poids moyen d'un échantillon de 25 recrues de l'année courante atteint 85 kg.

1. Tester, au seuil de signification de 5 % ($\alpha = 0,05$) si le poids moyen de l'année courante dépasse 80 kg.
2. Calculer la puissance du test précédent pour $\mu_1 = 85$.

Exercice 8.13

Le propriétaire d'une vidéothèque affirme que le nombre de vidéocassettes louées quotidiennement est de 1500. Un employé du magasin veut vérifier l'exactitude de l'affirmation de son patron. Pour ce faire, il retient un échantillon de 36 jours choisis au hasard dont la moyenne est de 1450 vidéocassettes louées par jour et l'écart type s au sein de l'échantillon de $s = 120$. Il veut tester l'affirmation au seuil $\alpha = 0.05$ en supposant que les ventes suivent une loi normale.

1. Quelle statistique proposez-vous d'utiliser ?
2. L'hypothèse que l'employé désire vérifier revient à effectuer le test suivant :
   $H_0 : \mu = \mu_0 = 1500$, contre
   $H_1 : \mu = \mu_1 < \mu_0$
   Déterminer numériquement la région critique R pour la statistique considérée.
3. Quelle devrait être la conclusion de l'employé ?

Exercice 8.14

En se basant sur un sondage réalisé auprès de 752 personnes, on s'intéresse à la proportion de personnes favorables à l'introduction d'un congé paternité dans un pays donné.

1. Sachant que 385 personnes ont répondu favorablement, peut-on affirmer que les personnes de ce pays sont favorables l'introduction d'un congé paternité ? Tester cette affirmation à l'aide d'un test d'hypothèse conservateur (en utilisant la variance maximale).
2. Effectué le même test en utilisant l'estimation de l'erreur standard.

Quatrième partie

Exercices récapitulatifs

A Statistique descriptive

Exercice 1.1 *

Le tableau suivant donne l'âge en années et le revenu mensuel en milliers de francs de 27 individus.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
âge	26	31	55	32	31	25	28	32	35	34	24	22	24	57
revenu	3.1	4.2	5.2	6.2	5.7	7.5	3.8	2.8	6.1	4.5	2.6	3.3	3.2	4.2

i	15	16	17	18	19	20	21	22	23	24	25	26	27
âge	28	39	27	23	42	41	38	25	22	21	24	44	41
revenu	4.9	6.3	5.2	3.8	4.8	7.1	5.1	2.9	3.3	2.2	3.5	3.9	5.2

1. Donner une présentation tige-feuilles de l'âge des individus. Utiliser les six tiges correspondants aux classes $20-24, 25-29, 30-34, 35-39, 40-49, 50-59$. Commenter.
2. Représenter le boxplot de la distribution des âges :
3. Compléter la table de contingence suivante (il suffit de trouver l'effectif manquant à l'intersection de la 1ère ligne et de la 2ème colonne, les autres s'en déduisent par différence), puis déterminer les pourcentages ligne. Commenter du point de vue de l'indépendance des variables.

	âge			
revenu	[20 - 30[	[30 - 40[	[40 - 60]	total
[2'000 - 4'000[	10	.	.	12
[4'000 - 8'000]	.	.	.	.
total	13	8	6	27

4. En utilisant les données du tableau de la question 3, construire l'histogramme de la distribution des âges. Commenter.
5. Sur la base du tableau de la question 3 (c'est-à-dire sans tenir compte du détail des données initiales), calculer l'âge moyen $\bar{x}_{age}$ (écart type $s_{age} = 9.8$), l'écart type de la distribution des revenus s_{rev} (revenu moyen $\bar{x}_{rev} = 4'666.67$), et finalement la corrélation $r_{age,rev}$ entre l'âge et le revenu.

6. Toujours à partir du tableau de la question 3, dessiner la courbe de concentration des revenus et calculer l'indice de Gini correspondant.

Exercice 1.2 *

Le 20 février 1994, le peuple suisse a accepté l'initiative dite "des Alpes", demandant le transfert du trafic marchandise de transit de la route au rail. Le tableau ci-dessous répartit les 26 cantons et demi-cantons selon leur pourcentage de "oui" et leur appartenance linguistique :

Appartenance linguistique	Pourcentages de "oui" [0-40[	[40-60]	]60-100]	Total
Suisse romande	4	2	0	6
Suisse alémanique	0	15	3	18
Suisse italienne et romanche	0	1	1	2
Total	4	18	4	26

1. Représenter les données ci-dessus sur un carré unitaire, en faisant apparaître la distribution marginale des appartenances linguistiques et les distributions conditionnelles des pourcentages de "oui".
2. Calculer la variance des pourcentages de "oui" par canton (à partir des données groupées du tableau ci-dessus).
3. Nous désirons avoir une représentation graphique plus précise des 15 cantons alémaniques ayant voté "oui" entre 40 et 60 %. Les pourcentages de "oui" de ces 15 cantons sont donnés ci-dessous :

   56 51 57 52 60 60 59 55
   57 52 55 60 56 48 53

   Construisez le boxplot de ces données.
4. Au vu des chiffres et du boxplot du point 3, le groupement en classes des pourcentages de "oui" donné dans l'énoncé vous semble-t-il judicieux ?

Exercice 1.3 *

Le tableau suivant donne la surface en m^2 nets par étudiant pour chaque faculté et école de l'Université de Genève en 1994-95.

Faculté/école	m^2 nets/étud	étudiants
Sciences	21.5	2'036
Médecine	23.5	1'435
Lettres	3.0	2'408
SES	2.3	2'705
Droit	4.8	1'170
Théologie	5.8	77
FPSE	2.6	2'034
IA	18.3	223
ETI	5.3	365
ELCF	1.1	170

1. Calculer la surface moyenne par étudiant pour l'ensemble de l'Université.

2. Déterminer la médiane, les premier et troisième quartiles et les valeurs extrêmes des surfaces par étudiant. Représenter ces valeurs sous forme de boxplot. Commenter la forme du boxplot ainsi que la différence entre la médiane et la surface moyenne trouvée précédemment.

On considère à présent le groupe "Sciences et Médecine" formé des facultés de Sciences et de Médecine, le groupe "Écoles" formé de l'IA, l'ETI et l'ELCF et le groupe "Autres facultés". Le tableau suivant donne la surface disponible et le nombre d'étudiants par 100 m^2.

	Surface en milliers de m^2	Étudiants par 100 m^2
Sciences et Médecine	78	4.5
Autres facultés	25	33
Écoles	6	12

7. Calculer l'écart type du nombre d'étudiants par 100 m^2.

8. Dessiner la courbe de concentration des étudiants par 100 m^2 par rapport aux trois groupes et calculer l'indice de Gini correspondant. Commenter.

Exercice 1.4 *

Le tableau suivant donne le taux d'acceptation par canton de chacune des deux initiatives fédérales soumises à votation le 6 juin 1993.

		% oui à l'initiative				% oui à l'initiative	
canton	votants	contre F/A-18	contre places d'armes	canton	votants	contre F/A-18	contre places d'armes
ZH	446'400	43.9	46.2	SH	36'444	42.9	42.1
BE	395'040	39.4	41.7	AR	22'106	39.8	37.3
LU	129'450	32.0	34.1	AI	5'258	30.2	29.0
UR	13'877	25.1	28.2	SG	160'979	41.0	39.8
SZ	38'842	33.7	34.3	GR	54'905	39.9	41.0
OW	11'211	25.2	26.4	AG	175'317	33.0	34.3
NW	15'667	22.4	25.3	TH	75'458	36.7	34.3
GL	12'773	33.7	34.8	TI	101'282	55.1	56.5
ZG	35'714	36.2	39.5	VD	172'793	46.9	50.3
FR	75'885	47.7	50.2	VS	80'069	40.4	41.5
SO	99'366	40.9	44.0	NE	50'087	47.5	52.1
BS	75'160	58.4	60.2	GE	108'363	58.3	63.5
BL	92'133	51.4	52.8	JU	25'442	69.4	73.3

1. Donner une présentation tige-feuilles (steam and leaf) des pourcentages de oui pour l'initiative contre les F/A-18. Commenter.

2. Déterminer la médiane ainsi que les premier et troisième quartiles des taux d'acceptation de l'initiative contre le F/A-18. Dessiner ensuite le boxplot de la distribution de ces taux (2cm = 10%). Commenter.

3. Calculer, pour les F/A-18, le taux moyen d'acceptation des votants des trois cantons suivants : Vaud (VD), Valais (VS), Genève (GE).

4. Sans faire de calcul, donner une évaluation à cinq centièmes près (1, 0.95, 0.9, ..., -0.95, -1) de la corrélation entre les taux d'acceptation des deux initiatives. Si vous deviez calculer cette corrélation, tiendriez-vous compte, et pourquoi, du nombre de votants par canton ?

5. Faire le diagramme de dispersion des cantons ZH, BE, UR, VS, GE, JU, selon les taux d'acceptation des deux initiatives. Prendre les F/A-18 en abscisse et les places d'armes en ordonnées.

6. En considérant les cinq cantons ZH, BE, LU, UR, SZ, calculer l'indice de Gini de la concentration des votants par canton.

7. La participation au scrutin du 6 juin 1993 a été de 58% à Genève et de 54.9% pour l'ensemble de la Suisse. Compléter le tableau suivant d'indices de taux de participation.

	Par rapport à		
Indice	Genève	Vaud	Suisse
Genève (GE)	.	.	.
Vaud (VD)	83.6	.	.
Suisse	.	.	.

Exercice 1.5 *

Le tableau suivant donne l'âge en années, le sexe (H = homme, F = femme) et la religion (C = catholique, P = protestant, A = autres) de 31 individus.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
âge	26	20	21	30	31	20	28	32	35	34	24	22	24	27	32
sexe	H	H	F	F	H	H	H	F	H	F	F	F	H	F	F
religion	A	P	C	C	C	C	P	A	P	C	P	C	P	P	A

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
âge	18	19	27	23	42	41	38	25	22	21	19	18	40	38	29	27
sexe	H	F	F	H	F	H	H	H	F	F	F	H	H	F	H	H
religion	P	C	P	P	C	C	P	P	P	C	C	P	C	C	P	A

1. Donner une présentation tige-feuilles de l'âge des individus. Utiliser les six tiges correspondants aux classes $10-19, 20-24, 25-29, 30-34, 35-39, 40-49$.

2. Représenter le boxplot de la distribution des âges :

3. Compléter le tableau suivant, puis donner l'approximation de l'âge moyen et de l'écart type de la distribution qui découle de ces données groupées.

âge	$[15-25[$	$[25-35[$	$[35-45]$
nombre d'individus	.	.	.

4. Compléter le tableau de contingence suivant

	catholique	protestant	autre	total
homme	.	.	.	.
femme	.	.	.	.
total	.	.	.	31

puis donner une présentation graphique permettant de comparer la distribution selon la religion des hommes à celle des femmes. Commenter.

Exercice 1.6 *

Le tableau suivant donne pour 21 cantons le pourcentage de oui recueillis par l'initiative "pour une politique raisonnable en matière de drogue" (*droleg*).

Canton	% oui	canton	% oui	canton	% oui
ZH	32	BS	34	TG	25
BE	27	BL	30	TI	20
LU	24	SH	33	VD	17
SZ	25	AR	25	VS	17
ZG	28	SG	26	NE	15
FR	18	GR	29	GE	25
SO	28	AG	26	JU	17

1. Donner une présentation tige-feuilles du pourcentage de oui.

2. Déterminer, selon la méthode (Tukey) vue au cours, la médiane et les quartiles du pourcentage de oui. Dessiner le boxplot de la distribution. (prendre 2 carrés = 10%)

3. Au vu du boxplot, préciser, en le justifiant mais sans calculs, si la moyenne des pourcentages observés est plus petite ou plus grande que la médiane.

Exercice 1.7 *

Le 22 mai 1992, le Département militaire fédéral (DMF) annonçait la suppression d'ici 1995 d'environ 1700 emplois. Le tableau suivant donne la répartition par région de 1696 postes menacés.

Emplois	Région		
	Suisse romande et Tessin (FR,VS,VD,TI)	Suisse orientale (GR,ZH,BE, AG,SG,TH)	Suisse centrale (LU,UR,NW, DW,SZ,GL)
menacés	112	1067	517
non menacés	2'653	11'159	3'847
Total	2'765	12'226	4'364

1. (a) Calculer pour chaque région le pourcentage d'emplois menacés ainsi que le pourcentage moyen suisse.

   (b) Déterminer l'écart type des pourcentages régionaux d'emplois menacés.

2. Donner une présentation graphique qui fait ressortir simultanément i) la répartition des emplois actuels par région et ii) la répartition entre menacés et non menacés pour chaque région. Commenter du point de vue de l'indépendance des deux variables.

3. Par rapport au pourcentage de la région "Suisse centrale", l'indice du pourcentage d'emplois menacés à Lucerne vaut 0,52.

   (a) Calculer ce même indice pour Uri, où le pourcentage d'emplois menacés est 26,1 % puis

   (b) déterminer l'indice d'Uri par rapport à Lucerne.

4. Calculer l'indice de Gini mesurant la concentration des emplois du DMF par région.

Exercice 1.8 *

Le tableau suivant donne la répartition en 1985 de la population américaine de plus de 25 ans selon l'âge et le niveau d'éducation. Les effectifs sont donnés en milliers. Pour les 65 ans et plus, on retiendra l'intervalle 65-80 ans.

Niveau d'éducation	Classe d'âges [25-35[	[35-45[	[45-55[	[55-65[	≥ 65	Total
Ecole secondaire non achevée	5416	5030	5777	7606	13746	37575
Ecole secondaire achevée	16431	1855	9435	8795	7558	44074
Université,1-3 ans	8555	5576	3124	2524	2503	22282
Université,4 ans ou plus	9771	7596	3904	3109	2483	26863
Total	40173	20057	22240	22034	26290	130794

1. Déterminer l'âge moyen (approximé) des personnes qui ont fait 4 ans ou plus d'études universitaires.
2. Tracer l'histogramme de la distribution selon l'âge des personnes qui n'ont pas achevé l'école secondaire. Multiplier vos hauteurs par une constante de sorte que la plus grande hauteur soit comprise entre 0.5 et 1.
3. Le tableau suivant donne, pour chaque classe d'âges, la distribution selon le niveau d'éducation (distribution colonnes). Compléter le tableau, comparer les distributions et commenter.

Niveau d'éducation	Classe d'âges [25-35[	[35-45[	[45-55[	[55-65[	≥ 65
Ecole secondaire non achevée	0.13	0.25	0.26	0.35	...
Ecole secondaire achevée	0.41	0.09	0.42	0.40	...
Université,1-3 ans	0.21	0.28	0.14	0.11	...
Université,4 ans ou plus	0.24	0.38	0.18	0.14	...

B Échantillonnage et probabilités

Exercice 2.1 *

Le tableau ci-dessous donne, pour 1993, la répartition par département du budget de la Confédération (en milliards de francs) :

Prévoyance sociale	11.295
Trafic	6.239
Défense nationale	5.753
Agriculture	3.416
Formation/Recherche	2.971
Étranger	2.070
Divers	8.856
	40.600

1. On désire tirer un échantillon de 4 budgets par tirages avec remises. Sachant que $\sigma^2 = 9.618$, calculer :

   (a) La variance de la moyenne d'échantillon : $\mathrm{Var}(\bar{X})$.

   (b) L'espérance de la variance d'échantillon : $\mathrm{E}(S^2)$.

2. On désire tirer un échantillon de 4 budgets par tirages sans remises. Calculer :

   (a) L'espérance de la moyenne d'échantillon : $\mathrm{E}(\bar{X})$.

   (b) La variance de la moyenne d'échantillon : $\mathrm{Var}(\bar{X})$.

3. Soit Y le budget et soit la variable X dont les trois modalités sont :

   x_1 : budgets compris entre 0 et 4 milliards de francs

   x_2 : budgets compris entre 4 et 8 milliards de francs

   x_3 : budgets compris entre 8 et 12 milliards de francs

   (a) Donner la distribution de probabilité de X.

   (b) Sachant que $X = x_1$, quelle est la probabilité que le budget Y soit supérieur à 2.5 milliards de francs ?

Exercice 2.2 *

En 1993, la télévision suisse a réparti ses ressources entre les 4 groupes linguistiques de la façon suivante :

Groupes linguistiques	Pourcentages de ressources
Allemand	$42,1$
Italien	$23,7$
Français	$33,0$
Divers	$1,2$

On désire prélever un échantillon de deux groupes linguistiques parmi les quatre groupes considérés (allemand, italien, français, divers), par tirages avec remises.

1. Énumérer les différents échantillons possibles.
2. Soit Z, le nombre de groupes linguistiques de l'échantillon qui reçoivent plus de 30% des ressources totales. Donner la distribution de probabilité de Z.

C Statistique inférentielle

Exercice 3.1 *

On désire estimer la moyenne du contrôle continu d'un cours de statistique et tester diverses hypothèses sans consulter l'ensemble des données. Pour cela, on tire aléatoirement 6 échantillons de 30 personnes ayant participé au contrôle. Les moyennes des notes des échantillons sont :

$$4.18 \quad 4.41 \quad 4.32 \quad 4.42 \quad 4.3 \quad 4.58$$

1. Bien que l'on ne connaisse pas la distribution des notes, pensez-vous que l'on puisse tout de même utiliser la loi normale pour effectuer de l'inférence sur la moyenne des notes ?

Pour les questions 2 à 6 on vous demande d'effectuer vos calculs avec trois chiffres après la virgule. Les questions peuvent se résoudre indépendamment. De plus on suppose que toutes les hypothèses classiques d'indépendance et de normalité sont satisfaites.

2. Donnez un intervalle de confiance à 95% pour la moyenne.
3. On veut savoir si, avec un risque $\alpha = 5\%$, les étudiants ont en moyenne obtenu une note inférieure à 4. L'hypothèse nulle est :

$$H_0 : \mu = \mu_0 = 4$$

   (a) Proposez l'hypothèse alternative adéquate pour le but proposé.
   (b) Donnez la région critique du test.
   (c) Donnez la conclusion du test et expliquez pourquoi celle-ci est ici triviale au vu des données.

4. L'écart type des notes de tous les élèves est 1.589.

   (a) Calculez l'écart type de la moyenne d'un échantillon de taille 30.
   (b) Calculez l'écart type de la moyenne des moyennes des 6 échantillons. C'est cet écart type, noté $\sigma_{\bar{x}}$, qu'il faut utiliser pour la suite de l'exercice.

5. Pour des raisons d'arrondis, nous posons $\sigma_{\bar{x}} = 0.1$. Nous voulons tester maintenant :

$$\begin{array}{rcl} H_0 & : & \mu = \mu_0 = 4.4 \\ H_a & : & \mu = \mu_1 > \mu_0 \end{array}$$

   (a) Déterminez la région d'acceptation de H_0 et donnez la conclusion du test.
   (b) Calculez la puissance du test pour
      - $\mu_1 = 4.45$
      - $\mu_1 = 4.5$
      - $\mu_1 = 4.6$

6. La moyenne des notes de tous les élèves est en réalité $\mu = 4.305$.
   (a) Calculez la puissance des deux tests (question 3 et 4).
   (b) Commentez l'intervalle de confiance trouvé en 2, ainsi que les autres résultats à la lumière de cette information supplémentaire sur μ.

Exercice 3.2 *

Dans un but d'inférence statistique, on admet que les revenus d'ouvriers qualifiés sont distribués selon une loi normale $N(\mu, \sigma^2)$, avec $\mu = 60$ et $\sigma^2 = 160$. On dispose d'un échantillon de 5 salaires (donnés en milliers de francs) :

$$56 \quad 72 \quad 48 \quad 64 \quad 80$$

1. Construire un intervalle de confiance à 80 % pour μ. Commenter.
2. On aimerait vérifier la pertinence de l'hypothèse $\mu = 60$. Pour ce faire, effectuer le test suivant, avec $\alpha = 10\%$:
$$H_0 : \mu = \mu_0 = 60 \quad \text{contre}$$
$$H_1 : \mu = \mu_1 > 60$$
Commenter.

Exercice 3.3 *

Le pourcentage X de voix obtenu par un candidat aux élections présidentielles dans un district choisi au hasard est supposé suivre une loi $N(\mu; \sigma^2)$.
Les valeurs suivantes de X ont été observées dans 9 districts choisis au hasard

$$40 \quad 50 \quad 55 \quad 95 \quad 60 \quad 45 \quad 48 \quad 38 \quad 54$$

1. Construire le boxplot de ces données.
2. Donner les estimations de μ que l'on obtient en prenant successivement comme estimateur la moyenne $\bar{X}$, et la médiane $\text{med}(X_1, X_2, \ldots, X_n)$ de l'échantillon. Discuter les qualités et inconvénients de chacun de ces estimateurs.
3. Donner une estimation non biaisée de σ^2. En déduire une estimation de l'écart type de $\bar{X}$.
4. Construire un intervalle de confiance à 90 % pour μ en supposant $\sigma^2 = 300$. Commenter.
5. Tester l'hypothèse $H_0 : \mu = 49\%$ contre $H_1 : \mu > 49\%$ avec un risque de première espèce de 5%. Commenter.

Exercice 3.4 *

Afin d'étudier le montant mensuel X dépensé pour l'alimentation par les rentiers AI, on dispose de l'échantillon suivant

$$900 \quad 1'100 \quad 1'250 \quad 800 \quad 950 \quad 750 \quad 1'000$$

On admet que X est distribué selon une loi $N(\mu, \sigma^2)$.

1. Donner deux estimations non biaisées de μ fondées respectivement sur la moyenne et la médiane de l'échantillon. Des deux estimateurs utilisés, lequel est le plus efficace sous l'hypothèse de normalité ?
2. Donner une estimation non biaisée de la variance σ^2 de X. En déduire une estimation de l'écart type σ.
3. Construire un intervalle de confiance à 90% pour μ.
4. Construire un intervalle de confiance à 90% pour σ^2. Donner l'intervalle correspondant pour σ.

Exercice 3.5 *

L'élection de deux conseillers fédéraux en 1999 a soulever le problème de la représentation du Tessin. Dans cette perspective on a relevé, pour le Tessin et trois régions, les écarts entre le pourcentage de "oui" (ou de "non" en cas d'écart négatif au Tessin) par rapport à l'ensemble de la Suisse. Les régions sont la Suisse romande « romand » (FR, VD, VS, NE, GE, JU), la Suisse alémanique alpines « alpes » (UR, SZ, OW, NW, GL, AI, AR, GR) et la Suisse alémanique industrielle « indust » (ZH, BE, LU, ZG, SH, SO, BS, BL, SG, AG, TG).

	TI	romand	alpes	indust
redPL	10.0	-3.8	-7.1	1.5
avs	19.0	10.7	-8.3	-3.4
finTP	10.1	3.4	-6.3	-0.9
droleg	6.3	7.3	-0.5	-2.6
LFtrav	3.1	11.6	-0.8	-3.7
moyenne	9.7	5.9	-4.6	-1.8
variance	28.191	31.488	10.912	3.727

1. En admettant que l'écart moyen pour la région « alpes » est $\mu_0 = -3$ et que la variance de ces écarts est $\sigma^2 = 130$, tester avec $\alpha = 5$ % si la moyenne observée pour le Tessin ($\bar{x} = 9.7$) est significativement supérieure à cette valeur. Préciser les hypothèses, la statistique utilisée et sa réalisation pour l'échantillon des cinq écarts observés, la forme de la région critique et la conclusion du test.
2. Construire un intervalle de confiance à 80 % pour la variance σ^2 de la distribution des écarts du Tessin.
3. En admettant que l'écart moyen pour la Suisse romande est $\mu_0 = 5$, tester avec $\alpha = 5$ % si la moyenne observée pour le Tessin ($\bar{x} = 9.7$) est significativement supérieure à cette valeur. Préciser les hypothèses, la statistique utilisée et sa réalisation pour l'échantillon des cinq écarts observés, la forme de la région critique et la conclusion du test.

D Ensemble du cours

Exercice 4.1 *

Du 1er mars au 30 juin 1993, 518 titres ont été décernés par l'Université de Genève. Le tableau suivant donne la répartition par faculté, école ou institut.

Nombre de titres décernés par facultés, écoles et instituts

Sciences	102	EA (Architecture)	4
Médecine	99	ETI (Traduction)	5
Lettres	57	ELCF (Langue française)	14
SES	123	IUHEI	40
Droit	27	Etudes européennes	6
PSE	37	IUED (Développement)	4

1. Regrouper les données ci-dessus en quatre classes, les trois premières d'amplitude 25 et la dernière d'amplitude 50, en choisissant 0 comme seuil inférieur de la première classe. Calculer les fréquences relatives f_k de chaque classe k, puis présenter les données sous forme d'histogramme (on déterminera les hauteurs à partir des fréquences n_k).
2. Calculer l'approximation de l'écart type de la distribution que l'on obtient à partir des données groupées.
3. On se propose de tirer un échantillon de taille 2 parmi les 12 facultés, écoles et instituts ci-dessus.
   (a) Quelle est la probabilité que les deux valeurs tirées soient strictement inférieures à 6 si l'on procède i) avec remises et ii) sans remises ?
   (b) Quelle est la probabilité que la première valeur soit inférieure à 30 et la seconde supérieure à 41 si l'on procède i) avec remises et ii) sans remises ?

On considère à présent la répartition des titres décernés entre mars et juin 1993 au sein de la Faculté des SES.

Licences en sciences sociales	27
Licences en sciences économiques	10
Licences en gestion	61
Diplômes et doctorats SES	25

4. Calculer l'indice de Gini pour les données ci-dessus. Commenter.
5. Les doctorats en SES représentent 20% de la catégorie "Diplômes et doctorats" des SES, et 7,14% de l'ensemble des doctorats décernés par l'Université de Genève. Calculer
   (a) la probabilité qu'un titre choisi au hasard parmi ceux décernés en SES soit un doctorat ;

(b) le nombre de doctorats décernés par l'Université de Genève.

Soit X le nombre d'étudiants obtenant un type de licence choisi au hasard parmi ceux offerts par l'Université de Genève. Afin d'étudier la distribution de X, notamment son espérance mathématique μ et sa variance σ^2, on choisit au hasard un échantillon de 7 types de licence. Voici 6 des 7 valeurs observées

$$5 \qquad 12 \qquad 20 \qquad 2 \qquad 12 \qquad 17 \qquad .$$

6. La moyenne $\bar{x}$ de l'échantillon vaut 11. En déduire la valeur manquante x_7, puis déterminer la médiane et le 1er quartile de l'échantillon.

Pour les questions suivantes, on admet que la valeur manquante est $x_7 = 14$.

7. Donner une estimation absolument correcte de μ et de σ^2.
8. On pense que le nombre moyen de licences décernées par type est 15. Tester cette hypothèse, avec un risque de première espèce de 5%, contre l'hypothèse que cette moyenne est inférieure à 15.

Exercice 4.2 *

Le tableau suivant donne le taux de participation et le taux d'acceptation par canton de l'article agricole (AGRI) soumis à votation le 9 juin 1996.

Cantons	% de oui	Participation	Cantons	% de oui	Participation
ZH	83.5	33.7	SH	79.5	54.7
BE	78.9	28.2	AR	75.7	42.0
LU	74.4	38.0	AI	69.2	27.2
UR	70.8	24.6	SG	76.6	32.9
SZ	65.1	26.7	GR	81.3	21.3
OW	67.2	32.4	AG	74.1	26.4
NW	72.1	33.2	TH	69.7	28.6
GL	78.8	25.1	TI	81.5	15.8
ZG	77.9	33.2	VD	67.0	30.8
FR	71.1	33.4	VS	66.2	13.8
SO	73.7	36.7	NE	76.5	23.7
BS	87.5	48.1	GE	85.5	60.1
BL	81.5	29.2	JU	70.3	39.4

1. Déterminer la médiane ainsi que les premier et troisième quartiles du taux d'acceptation de l'article agricole. Dessiner ensuite le boxplot de la distribution de ce taux. Commenter.
2. Compléter la table de contingence suivante qui résume les données précédentes. (*Il suffit de trouver les valeurs de deux cases, les autres s'en déduisent par calcul !*)

	Acceptation			
Participation	[60 - 70[	[70 - 80[	[80 - 90]	total
[0 - 33.3[	.	.	.	17
[33.3 - 66.6]	.	.	.	9
total	6	14	6	26

3. Calculer les fréquences relatives des taux d'acceptation pour chacune des deux catégories de participation (distributions conditionnelles par lignes). Représenter ces deux distributions ainsi que la distribution du taux de participation sur un même carré unitaire. Commenter du point de vue de l'indépendance.
4. Sur la base du tableau de la question 2 (c'est-à-dire sans tenir compte du détail des données initiales), calculer la covariance entre la participation et l'acceptation.

Pour la votation sur la réforme du gouvernement (autre sujet soumis à votation le 9 juin), on dispose uniquement des taux d'acceptation pour 5 cantons (tirés au hasard avec remise).

33.4 56.3 27.8 28.2 40.3

5. Donner un estimateur non biaisé de l'écart-type pour tous les cantons suisses et calculer l'estimation obtenue avec l'échantillon ci-dessus.

Pour les deux questions suivantes, on admettra que la variance du taux d'acceptation par canton pour la réforme du gouvernement vaut, pour l'ensemble de la Suisse, $\sigma^2 = 110.25$.

6. En utilisant l'échantillon précédent, estimer le taux moyen d'acceptation des cantons suisses par un intervalle de confiance à 90%.
7. Calculer la variance de la moyenne $\bar{X}$ d'un échantillon de taille 5.

   On dispose par ailleurs d'un autre estimateur, noté $\tilde{X}$, de la moyenne des taux de tous les cantons, avec $\text{Biais}(\tilde{X}) = 5$ et $\text{Var}(\tilde{X}) = 1$. Quel estimateur choisissez-vous ? Justifier.
8. On sait que la moitié des 6 cantons romands ont accepté la réforme du gouvernement, alors que seuls 11.54% de l'ensemble des 26 cantons suisses l'ont acceptée. Quel est le pourcentage de cantons romands parmi ceux qui ont refusé la réforme ?

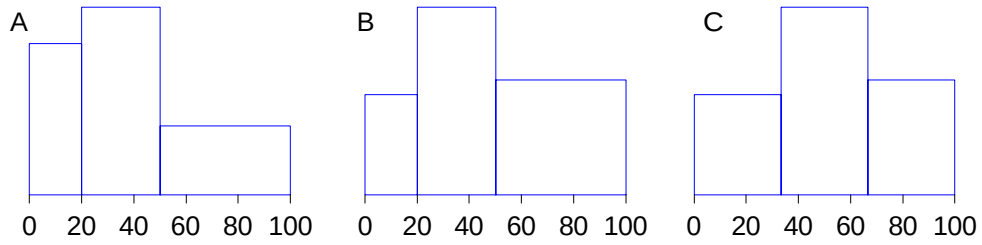
E Exercices Q.C.M.

Exercice 5.1 *

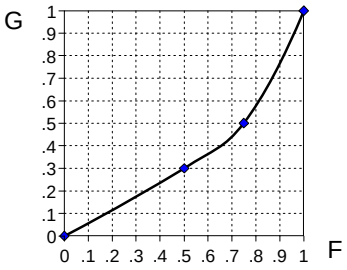
Le tableau ci-dessous donne la distribution de 28 entreprises selon leur nombre d'employés

nombre d'employés	$0-20$	$20-50$	$50-100$
nombre d'entreprises	7	13	8

et voici trois histogrammes :



1. Quel est l'histogramme des données considérées ?
   a) A b) B c) C d) aucun
2. Pour quel histogramme a-t-on la plus forte dispersion ?
   a) A b) B c) C d) on ne peut pas se prononcer
3. De A et C, quel histogramme représente une distribution où la médiane est inférieure à la moyenne ?
   a) A b) C c) les deux d) aucun
4. La figure ci-contre illustre la courbe de concentration des enfants par classes (F fréquences cumulées des classes, G fréquences cumulées des élèves).
   On y lit que :



   a) La moitié des élèves est dans le 70% des classes les moins nombreuses.

b) 70% des élèves sont dans le 50% des classes les moins nombreuses.

c) La moitié des classes les plus nombreuses comprend 75% des élèves.

d) 70% des élèves sont dans le 50% des classes les plus nombreuses

5. Soit x la population en milliers d'habitants et y le taux de divorce en pourmilles. Pour quatre villes on a observé :

Ville	A	B	C	D	moyenne
population x	50	800	300	100	312.5
divorce y	5	9	7	8	7.25

Calculer la covariance entre x et y

a) 2587,5 b) 0,734 c) 438,67 d) 321,875

6. la corrélation entre deux variables x et y est égale à -1. Vous en concluez que

a) y est proportionnel à x

b) y est inversement proportionnel à x.

c) y croit de façon inversement proportionnelle à x

d) y diminue proportionnellement à x

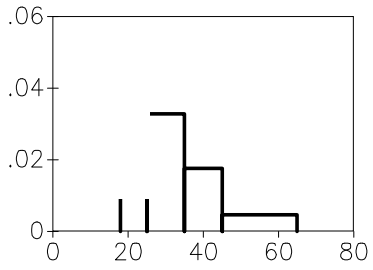
7. Selon un récent sondage, 60% des romands et tessinois sont favorables au traité sur l'EEE, contre 39% des alémaniques. En admettant que 2/3 des personnes interrogées sont alémaniques, quelle est la probabilité qu'un individu interrogé choisi au hasard soit alémanique et non favorable (individu ou opposé) à l'EEE ?

a) 61% b) 40,7% c) 91,5% d) 26%

Exercice 5.2 *

1. L'histogramme partiellement effacé ci-dessous représente la distribution suivante des accidentés hommes de la circulation.

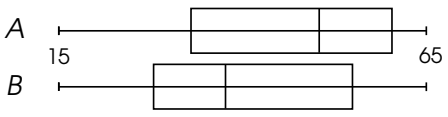
âge	$[18-25[$	$[25-35[$	$[35-45[$	$[45-65]$
hommes	2255	1830	980	515



Quelle est la hauteur du premier rectangle de l'histogramme ?

a) 322,143 b) 0,404 c) 0,040 d) 0,058

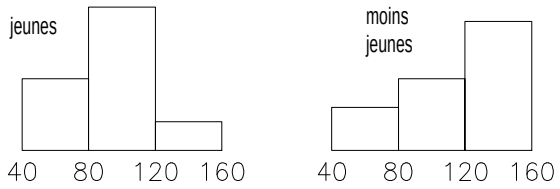
2. Voici deux boxplots. Dans quel cas la proportion de données inférieures au 1er quartile est-elle la plus importante ?



a) A b) B

c) Elle est identique dans les 2 cas d) on n'a pas suffisamment d'information

Les histogrammes ci-dessous schématisent la distribution des jeunes (< 40 ans) et moins jeunes (> 40 ans) d'une entreprise selon leur revenu.



3. Dans quel cas a-t-on la plus grande dispersion ?
   a) jeunes b) moins jeunes
   c) c'est la même dans les 2 cas d) on n'a pas suffisamment d'information
4. Dans quel cas a-t-on la plus forte concentration du revenu ?
   a) jeunes b) moins jeunes
   c) c'est la même dans les 2 cas d) on n'a pas suffisamment d'information
5. Le tableau suivant donne les quantités et les valeurs de deux biens en 1986 et 1991

	quantités		valeurs	
biens	1986	1991	1986	1991
A	80	150	800	3′000
B	220	100	660	1′000

   L'indice Laspeyres des prix de 1991 par rapport à 1986 vaut :
   a) 219,89 b) 123,29 c) 222,22 d) 260,27
6. Dans une ville donnée, 40% de la population a les cheveux bruns, 25% a les yeux marron et 15% à la fois les cheveux bruns et les yeux marron. On choisit au hasard un résident de cette ville. Quelle est la probabilité qu'il ait les yeux marron si l'on sait qu'il n'a pas les cheveux bruns ?
   a) 0,167 b) 0,333 c) 0,250 d) 0,375

Exercice 5.3 *

1. Voici une présentation tige-feuilles des notes obtenues aux exercices imposés par les 23 concurrents d'un concours de patinage artistique :

tige	feuilles (dixièmes)
3	.8 .9
4	.5 .7 .3 .2 .2 .5 .9 .9 .8
5^-	.4 .2 .3 .3 .0 .1 .2
5^+	.7 .8 .5 .9 .9

   Quel est le premier quartile de cette distribution ?
   a) 4.2 b) 5 c) 4.5 d) 6
2. Le tableau ci-dessous donne pour trois votations fédérales, le nombre de questions soumises et le taux d'abstention.

nombre de questions :	3	2	4
% abstention :	60	51	66

   La corrélation (arrondie au millième) entre les deux séries est :
   a) 0,197 b) 5,000 c) 0,986 d) 0,993

On dispose des données suivantes pour deux biens :

	prix		quantités	
	1985	1991	1985	1991
bien 1	3	4	10	8
bien 2	5	6	7	8

3. L'indice (simple) des quantités du bien 1 de 1991 par rapport à 1980 $I^1_{91/80}(q)$ vaut 3. Que vaut $I^1_{85/80}(q)$?

   a) 2,4 b) 3,75 c) 2,25 d) 3,05

4. L'indice Laspeyres des prix de 1990 par rapport à 1985 vaut 1,14. Trouver le taux d'inflation entre 1991 et 1990. (Utiliser 4 décimales pour les calculs intermédiaires.)

   a) 26,2% b) 10,7% c) 1,3% d) 12,2%

5. Voici la répartition des députés des quatre principales formations politique d'un parlement :

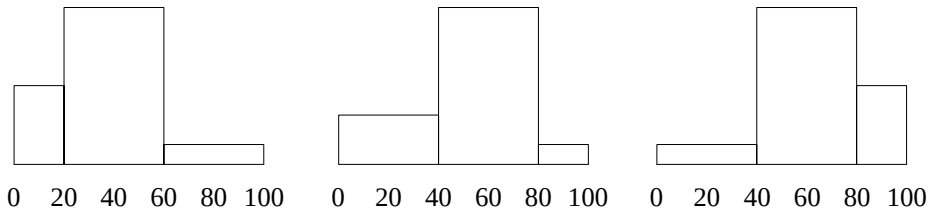
Parti :	nationaliste	républicain	écologiste	socialiste
Nombre de députés :	40	60	10	90

   Calculer l'indice de concentration de Gini de cette distribution.

   a) 0,325 b) -0,325 c) -0,125 d) 0,125

6. Parmi les distributions suivantes de revenu, laquelle correspond à la plus forte concentration du revenu ?

   a) b) c)



7. On admet que la probabilité qu'il pleuve un jour donné est égale à 0,1 et que la probabilité que Monsieur Dupont ait un accident un jour quelconque vaut 0,005 et 0,012 un jour de pluie. Déterminer la probabilité qu'un jour choisi au hasard il pleuve et que M. Dupont ait un accident.

   a) 0,24 b) 0,12 c) 0,012 d) 0,0012

8. Soit une population de trois étudiants ayant obtenu respectivement les moyennes de maturité 3,5 , 5 et 5. On envisage de prélever un échantillon de taille 2 par tirages sans remises. Quelle est la variance de la médiane $X_{\text{med}} = \text{med}(X_1, X_2)$ de l'échantillon ?

   a) 0,25 b) 0,5 c) 0,125 d) 0,35

Exercice 5.4 *

La température X à 14 heures en août à Genève est supposée suivre une loi $N(\mu = 28,\ \sigma^2 = 10)$

1. Quelle est, selon cette hypothèse, le nombre moyen de jours où la température dépasse 31 degrés en août ?

   a) 0,95 b) 0,83 c) 25,7 d) 5,3

   e) 1,6

2. Voici 6 relevés effectués au hasard en août,

$$22 \quad 28 \quad 33 \quad 27 \quad 32 \quad 25$$

Quel est, en maintenant l'hypothèse $\sigma^2 = 10$, l'intervalle de confiance à 90% pour μ ?

a) [26,18 ; 29,49] b) [25,30 ; 30,36] c) [25,71 ; 29,96] d) [22,63 ; 33,04]

3. Un échantillon donné vous amène à accepter une hypothèse H_o au dépend d'une hypothèse H_1 au seuil de signification de 10%. La puissance du test est de 30%. Quelle serait si H_1 était vraie, la probabilité que votre échantillon vous ait amené à rejeter à tort H_1 ?

a) 10% b) 30% c) 70%

d) on ne dispose pas de suffisamment d'information

Exercice 5.5 *

La proportion de non-fumeurs Y (en pourcents) parmi les passagers d'une compagnie aérienne est supposée suivre la loi normale $N(\mu, \sigma^2)$.

1. Pour $\mu = 54$ et $\sigma^2 = 144$, quelle est la probabilité que sur un vol choisi au hasard, il y ait au moins trois fois plus de non-fumeurs que de fumeurs ?

a) 0,96 b) 0,25 c) 0,75 d) 0,04

2. Pour 30 vols choisis au hasard on a obtenu

$$\bar{y} = 54 \qquad \text{et} \qquad \sum (y_i - \bar{y})^2 = 4'950$$

Donner l'estimation de l'écart-type $\sigma_{\bar{y}}$ de $\bar{Y}$ fondée sur l'estimation non biaisée de σ^2.

a) 2,385 b) 2,345 c) 5,690 d) 13,065

3. En considérant le même échantillon ($n = 30, \bar{y} = 54$), et en admettant que σ^2 vaut 144, déterminez l'intervalle de confiance à 90 % pour μ.

a) $[49,71; 58,29]$ b) $[50,40; 57,60]$

c) $[50,33; 57,67]$ d) $[51,20; 56,80]$

4. On se propose de prélever un échantillon de taille $n = 6$, par tirage avec remises, dans une population distribuée comme suit :

x	0	1	2	4
$P(X = x)$	0,3	0,1	0,2	0,4

Donner l'espérance mathématique $\mathrm{E}(S^2)$ de la variance de l'échantillon S^2.

a) 0,481 b) 2,89 c) 2,408 d) 3,468

5. Un échantillon donné vous amène à rejeter une hypothèse H_0 au profit d'une hypothèse H_1 au seuil de signification $\alpha = 10\%$. Quelle sera la conclusion si vous diminuez le risque α ?

a) rejet de H_0 b) acceptation de H_0

c) on ne peut pas se prononcer

Réponses aux questions QCM

Exercice E.1
1 a), 2 b), 3 a), 4 d), 5 d), 6 d), 7 b).

Exercice E.2
1 d), 2 c), 3 b), 4 a), 5 d), 6 a).

Exercice E.3
1 c), 2 d), 3 b), 4 b), 5 a), 6 a), 7 d), 8 c).

Exercice E.4
1 d), 2 c), 3 c).

Exercice E.5
1 d), 2 a), 3 b), 4 c), 5 c).

F Exercices EXCEL

Exercice 6.1

Le tableau suivant représente une feuille Excel.

	A	B	C
1	taux de change	1.7	
2		francs	dollars
3	moyenne	100	
4	écart type	8	
5	variance		

3. Quelle formule Excel faut-il mettre en C3 pour obtenir la moyenne en dollars de façon à pouvoir ensuite la copier coller en C4 pour obtenir l'écart type en dollars ?

4. Quelle formule Excel permet en B5 d'obtenir la variance en francs ?

5. Laquelle des deux formules précédentes faut-il copier en C5 pour obtenir la variance en dollars ?

Exercice 6.2

Le tableau suivant représente une feuille Excel.

	A	B	C
1		3	=(B1–B$4)^2
2		5	
3		10	
4		=MOYENNE(B1:B3)	

1. Vous copiez le contenu de la cellule C1 en C2 et C3 et celui de la cellule B4 en C4. Indiquer dans le tableau ci-dessus les formules Excel qui en résultent en C2, C3 et C4.

2. Quelle valeur obtient-on en C3 ? Que représente-t-elle ?

3. Donner et interpréter la valeur qui en résulte en C4.

Exercice 6.3

	A	B	C
1	*dép. Mén*	*chômage*	*véhicules*
2	18	0.6	339
3	23	0.5	323
4	22	1	310

1. Calculer la valeur que donne la formule Excel « =VAR(A2:A4) »placée en B7.

Exercice 6.4

Refaire l'exercice 1.5 avec Excel.

Cinquième partie

Solutions de quelques exercices

Solution de l'exercice 1.4

1. Expliciter l'inconnue x des équations suivantes

   (a) $a + bx = c$

   $$\begin{aligned} a + bx &= c \\ bx &= c - a \\ x &= \frac{c-a}{b} \end{aligned}$$

   (b) $(x - m)/s = z$

   $$\begin{aligned} (x-m)/s &= z \\ (x-m) &= sz \\ x &= sz + m \end{aligned}$$

   (c) $d/x = g$

   $$\begin{aligned} d/x &= g \Leftrightarrow d = gx \\ x &= d/g \end{aligned}$$

   (d) $\alpha^x = \gamma$

   $$\begin{aligned} \alpha^x &= \gamma \\ x\log(\alpha) &= \log(\gamma) \\ x &= \log(\gamma)/\log(\alpha) \end{aligned}$$

2. Déterminer la valeur de la solution de x des équations précédentes pour

   (a) $a = 10$, $b = -0.5$ et $c = 2$

   $$x = \frac{c-a}{b} = \frac{2-10}{-0.5} = 16$$

   (b) $m = 10$, $s = 4$, $z = 1.28$

   $$x = sz + m = 4 \cdot 1.28 + 10 = 15.12$$

   (c) $d = 0.5$, $g = -2$

   $$x = d/g = 0.5/(-2) = -0.25$$

   (d) $\alpha = 10$, $\gamma = 1000$

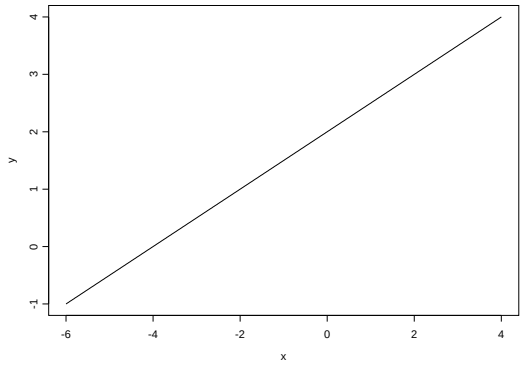
   $$x = \log(\gamma)/\log(\alpha) = \log(1000)/\log(10) = 3$$

Solution de l'exercice 1.5

Représenter graphiquement dans le plan (x, y) les courbes

1. $y = 2 + 0.5\,x$, pour $x \in [-6; 4]$

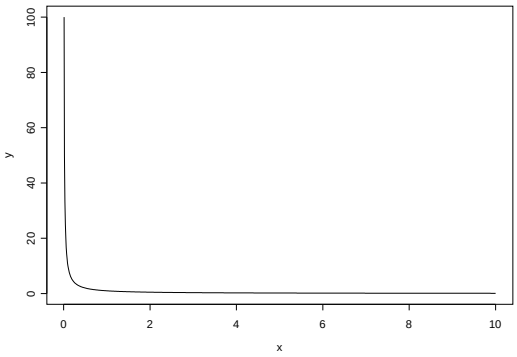
   Il suffit de calculer deux points, par exemple,
   $x = -6 \Rightarrow y = -1$ et
   $x = 4 \Rightarrow y = 4$.
   La droite passe par ces deux points.



2. $y = 1/x$, pour $x \in]0; 10]$

   On fait passer la courbe par quelques points :

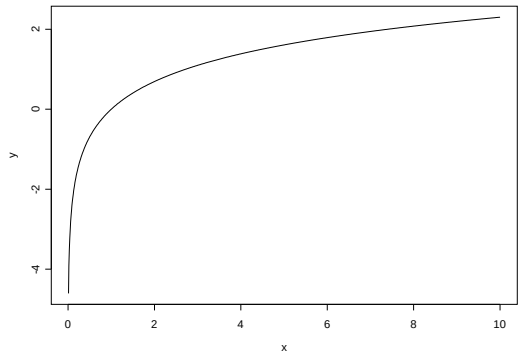
x	$y = 1/x$
.01	100
.1	10
1	1
2	.5
10	.1



3. $y = \ln x$ pour $x \in]0; 10]$

   On fait passer la courbe par quelques points :

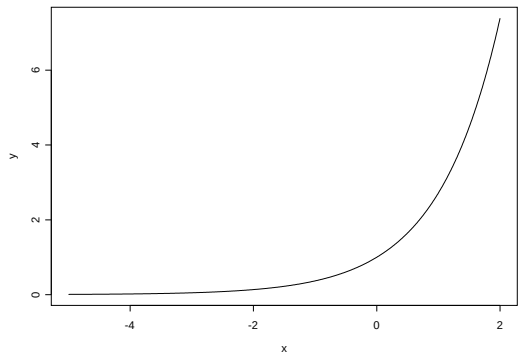
x	$y = \ln x$
.01	-4.6
.3	-1.2
1	0
3	1.1
10	2.3



4. $y = e^x$ pour $x \in [-5; 2]$

   On fait passer la courbe par quelques points :

x	$y = e^x$
-5	.007
-1	.37
0	1
1.5	4.5
2	7.4



Solution de l'exercice 2.2

Établir une présentation tige-feuilles (steam and leaf) du nombre d'autorisation de construire de 72 communes valaisannes. On peut, sans que cela ne soit requis cependant, d'abord ordonner les données en ordre croissant :

4	5	6	10	11	14	17	17	18	18	19	20	21	21	22	23	23	23	23	23
23	23	24	24	24	25	26	26	26	26	27	27	28	29	29	29	30	30	31	32
33	34	34	34	35	36	36	36	37	37	37	43	43	43	46	48	48	55	55	55
60	62	62	63	65	66	68	70	71	72	75	100								

Il faut choisir des "tiges" *raisonnables*, de sorte que la présentation ne soit ni trop étendue, ni trop condensée. Par exemple, on peut choisir comme tige les dizaines :

Tige	Feuilles
0	4 5 6
1	0 1 4 7 7 8 8 9
2	0 1 1 2 3 3 3 3 3 3 3 4 4 4 5 6 6 6 6 7 7 8 9 9 9
3	0 0 1 2 3 4 4 4 5 6 6 6 7 7 7
4	3 3 3 6 8 8
5	5 5 5
6	0 2 2 3 5 6 8
7	0 1 2 5
8	
9	
10	0

Un autre choix des tiges peut être les demi-dizaines :

Tige	Feuilles
0-	4
0+	5 6
1-	0 1 4
1+	7 7 8 8 9
2-	0 1 1 2 3 3 3 3 3 3 3 4 4 4
2+	5 6 6 6 6 7 7 8 9 9 9
3-	0 0 1 2 3 4 4 4
3+	5 6 6 6 7 7 7
4-	3 3 3
4+	6 8 8
5-	
5+	5 5 5
6-	0 2 2 3
6+	5 6 8
7-	0 1 2
7+	5
8-	
8+	
9-	
9+	
10-	0

Cette deuxième présentation plus fine est évidemment plus étalée. Elle donne une vision plus détaillée de la distribution des données. Il faut remarquer que les deux présentations sont cependant acceptables.

Solution de l'exercice 2.5

Les données x_i concernent les dépenses pour la santé en % du PIB de 22 pays en 1986 :

Europe	Allemagne	8.1	Finlande	7.5	Luxembourg	6.9
	Autriche	8.3	France	8.5	Norvège	6.8
	Belgique	7.1	Irlande	8.0	Pays-Bas	8.3
	Danemark	6.0	Islande	7.5	Suisse	7.7
	Espagne	6.0	Italie	6.7	Royaume-Uni	6.1
	Suède	9.1	Portugal	5.6		
Autres	Australie	6.9	Etats-Unis	10.9	Nouvelle-Zélande	8.3
	Canada	8.7	Japon	6.7		

1. Les données ordonnées sont :

rang i	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$	5.6	6	6	6.1	6.7	6.7	6.8	6.9	6.9	7.1	7.5
rang i	12	13	14	15	16	17	18	19	20	21	22
$x_{(i)}$	7.5	7.7	8	8.1	8.3	8.3	8.3	8.5	8.7	9.1	10.9

Médiane : $n = 22$ étant pair, la médiane est la moyenne des deux valeurs centrales :

$$\text{rang}(\text{med}\, x) = \frac{n+1}{2} = \frac{23}{2} = 11.5 \quad \Rightarrow \quad \text{med}(x) = \frac{7.5+7.5}{2} = \boxed{7.5}$$

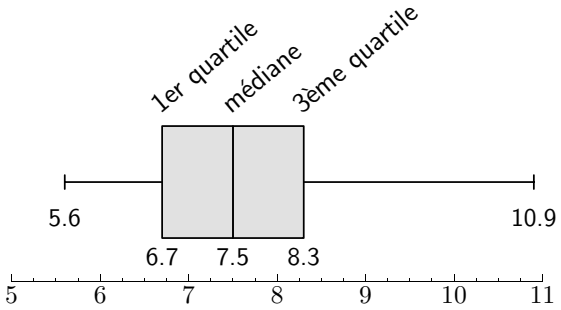
Quartiles : $n = 22$ étant pair, q_1 est la médiane des $n/2 = 11$ premières observations et q_3 est la médiane des $n/2 = 11$ dernières. (Pour n impair, q_1 serait la médiane des $(n+1)/2$ premières observations, et q_3 celle des $(n+1)/2$ dernières.)

$$\text{rang}(q_1) = \frac{11+1}{2} = 6 \quad \Rightarrow \quad \boxed{q_1 = 6.7}$$

$$\text{rang}(q_3) = \frac{n}{2} + \text{rang}(q_1) = 11 + 6 = 17 \quad \Rightarrow \quad \boxed{q_3 = 8.3}$$

(Pour n impair, on aurait $\text{rang}(q_3) = (n-1)/2 + \text{rang}(q_1)$.)

Boxplot



2. *Moyenne.* On a $n = 22$ dont $n_E = 17$ pays européens.

$$\sum_{i \in \text{Eur}} x_i = 124.2 \quad \Rightarrow \quad \bar{x}_E = \frac{1}{n_E} \sum_{i \in \text{Eur}} x_i = \frac{124.2}{17} = \boxed{7.3}$$

$$\sum_{i \notin \text{Eur}} x_i = 6.9 + 10.9 + 8.3 + 8.7 + 6.7 = 41.5$$

$$\sum_i x_i = \underbrace{\sum_{i \in \text{Eur}} x_i}_{124.2} + \underbrace{\sum_{i \notin \text{Eur}} x_i}_{41.5} = 165.7 \quad \Rightarrow \quad \bar{x} = \frac{1}{n} \sum_i x_i = \frac{165.7}{22} = \boxed{7.53}$$

Les pays de l'OCDE consacraient donc en moyenne 7.53% du PIB à la santé en 1986. Ce pourcentage est un peu moins fort si l'on considère les pays européens seulement.

3. *Variance.* On donne $\sum_{i\in\text{Eur}} (x_i - \bar{x})^2 = 17.42$. Il reste donc à calculer la somme des $(x_i - \bar{x})^2$ pour les 5 pays non européens.

$$\begin{aligned}\sum_{i\notin\text{Eur}} (x_i - \bar{x})^2 &= (6.9 - 7.53)^2 + (10.9 - 7.53)^2 + (8.3 - 7.53)^2 + \\ &\quad + (8.7 - 7.53)^2 + (6.7 - 7.53)^2 \quad = \quad 14.40 \\ \Rightarrow \quad \sum_i (x_i - \bar{x})^2 &= \underbrace{\sum_{i\in\text{Eur}} (x_i - \bar{x})^2}_{17.42} + \underbrace{\sum_{i\notin\text{Eur}} (x_i - \bar{x})^2}_{14.4} \quad = \quad 31.82\end{aligned}$$

d'où la variance s_x^2 et l'écart type s_x

$$\begin{aligned}\text{var}\, x = s_x^2 &= \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{31.82}{22} = \boxed{1.446} \\ s_x &= \sqrt{1.446} = \boxed{1.20}\end{aligned}$$

4. *Moyenne et variance approximées à partir des données groupées*

k	classe	centre x_k	frequence n_k	freq. rel. f_k	$f_k x_k$	$(x_k - \bar{x}_A)^2$	$f_k(x_k - \bar{x}_A)^2$
1	[0, 3[	1.5	0	0	0	37.663	0
2	[3, 6[	4.5	1	0.0455	0.205	9.841	0.447
3	[6, 9[	7.5	19	0.8636	6.477	0.019	0.016
4	[9,12]	10.5	2	0.0909	0.955	8.197	0.745
total			22	100%	$\underbrace{7.637}_{\bar{x}_A}$		$\underbrace{1.208}_{\text{var}_A\, x}$

La moyenne approximée est $\bar{x}_A = \sum_k f_k x_k = \boxed{7.637}$.

La variance approximée est $\text{var}_A\, x = \sum_k f_k (x_k - \bar{x}_A)^2 = \boxed{1.208}$, ce qui donne l'écart type approximé de 1.10.

En prenant les intervalles fermés à droite, les résultats deviennent

k	classe	centre x_k	frequence n_k	freq. rel. f_k	$f_k x_k$	$(x_k - \bar{x}_A)^2$	$f_k(x_k - \bar{x}_A)^2$
1	[0, 3]	1.5	0	0	0	34.386	0
2	]3, 6]	4.5	3	0.1364	0.614	8.202	1.119
3	]6, 9]	7.5	17	0.7727	5.795	0.018	0.014
4	]9,12]	10.5	2	0.0909	0.955	9.834	0.894
total			22	100%	$\underbrace{7.364}_{\bar{x}_A}$		$\underbrace{2.027}_{\text{var}_A\, x}$

La moyenne approximée est $\bar{x}_A = \sum_k f_k x_k = \boxed{7.364}$.

La variance approximée est $\text{var}_A\, x = \sum_k f_k (x_k - \bar{x}_A)^2 = \boxed{2.027}$, ce qui donne l'écart type approximé de 1.42.

Commentaire : Le choix des classes n'est pas pertinent, puisque 19 cas sur 22, soit 86% (17/22 = 77% dans le second cas), se retrouvent dans la même classe et que la première classe est vide.

Si la moyenne approximée est proche de la vraie moyenne, cela tient au fait que la vraie moyenne 7.53 est proche du milieu de la classe qui contient ces 19 cas.

Compte tenu de la concentration de données dans la troisième classe, on a de la chance que dans le premier cas la variance approximée ne soit que faiblement inférieure (1.208 contre 1.44) à la vraie variance.

On remarque qu'en déplaçant les 2 cas qui sont à la limite inférieure de la 3ème classe vers la 2ème, on double presque la variance. Ceci illustre la sensibilité des résultats approximés aux limites de classe.

Solution de l'exercice 2.7

Le tableau suivant donne les taux de participation à une votation fédérale pour cinq catégories (années) d'âges. Déterminez les valeurs manquantes par interpolation linéaire.

Age	Taux de participation
23	28%
30	?
32	40%
?	45%
44	55%

1. On prend les valeurs qui entourent ce que l'on cherche

23	28%
30	?
32	40%

entre 32 et 23 ans, il y a 9 ans de différence qui correspondent à 12% de différence $(40\% - 28\%)$

entre 23 et 30, il y a 7 ans de différence.

On fait le calcul suivant :

a) un écart de 9 ans $\Rightarrow$ 12% de différence

b) un écart de 1 ans $\Rightarrow 12/9 = 1.33\%$ de différence

c) un écart de 7 ans $\Rightarrow 7 \cdot 1.33\% = 9.31\%$ de différence

Finalement, la participation pour 30 ans est de 28% (pour 23 ans) + 9.31% (pour 7 ans) = 37.31%.

2. On prend les valeurs qui entourent ce que l'on cherche

32	40%
?	45%
44	55%

entre 55% et 40% il y a 15% de différence qui correspondent à 12 ans de différence $(44 - 32$ ans)

entre 45% et 40%, il y a 5% ans de différence.

On fait le calcul suivant :

a) un écart de 15% $\Rightarrow$ 12 ans de différence

b) un écart de 1% $\Rightarrow 12/15 = 0.8$ ans de différence

c) un écart de 5% $\Rightarrow 5 \cdot 0.8 = 4$ ans de différence

Finalement, l'âge correspondant à 45% est de 32 ans (pour 40%) + 4ans (pour 5%) = 36 ans.

Solution de l'exercice 2.8

classe k	n_k	f_k	F_k	$\bar{x}_k$	$f_k\bar{x}_k$
[0-1000[	15	0.15	0.15	500	75
[1000-3000[	20	0.20	0.35	2000	400
[3000-4000[	30	0.30	0.65	3500	1050
[4000-6000[	20	0.20	0.85	5000	1000
[6000-7000]	15	0.15	1	6500	975
	$n = 100$	1			$\bar{x}_A = 3500$

avec n_k effectif de la classe k $\quad f_k = \frac{n_k}{n}$ fréquence relative de k
F_k fréquences relatives cumulées $\quad \bar{x}_k$ milieu de la classe k

1. Calcul de la médiane

   (a) La *classe médiane* est la première classe dont la fréquence relative cumulée dépasse 0,5. Dans notre cas, la classe médiane est $]3000 - 4000]$.

   (b) Parmi les cas observés, 35% ont un salaire de moins de 3000 Frs, et 65% un salaire inférieur à 4000Frs. On veut le salaire médian, c'est-à-dire celui pour lequel 50% des cas ont un salaire inférieur. Comme on ne dispose que des données groupées, on approxime cette valeur par interpolation linéaire :

3000	0.35
x	0.50
4000	0.65

$$\frac{x-3000}{4000-3000} = \frac{x-3000}{1000} = \frac{0.50-0.35}{0.65-0.35} = \frac{0.15}{0.30}$$

$$x = 3000 + \frac{0.15}{0.30}1000 = \boxed{3500}$$

2. Calcul de l'approximation de la moyenne.

   Il s'agit de la moyenne des centres des classes pondérées par les fréquences relatives f_k :

$$\begin{aligned}\bar{x}_A &= \sum_{k=1}^{5} f_k\bar{x}_k \\ &= 0.15\cdot 500 + 0.20\cdot 2000 + 0.30\cdot 3500 + 0.20\cdot 5000 + 0.15\cdot 6500 \\ &= 75 + 400 + 1050 + 1000 + 975 = \boxed{3500}\end{aligned}$$

Remarque : Les médiane et moyenne approximées sont identiques car la distribution groupée est symétrique.

Solution de l'exercice 3.1

Le tableau suivant donne la répartition de 100 personnes entre 15 et 30 ans selon leur âge et leurs goûts musicaux. On a donc deux caractères : "âge" et "goût musical". Chaque caractère a deux modalités.

	goût musical		
âge	Rock	Techno	Total
15 à 20 ans	18	35	53
21 à 30 ans	25	22	47
Total	43	57	100

1. *Fréquences relatives caractérisant la distribution conjointe.*

   On doit calculer pour chaque case (i, j) du tableau la fréquence relative avec laquelle on observe simultanément la modalité ligne i et la modalité colonne j. Cette fréquence est donnée par $f_{ij} = n_{ij}/n$. Nous avons donc :

$$f_{11} = \frac{n_{11}}{n} = \frac{18}{100} = 0,18; \quad f_{12} = \frac{35}{100} = 0,35; \quad f_{21} = \frac{25}{100} = 0,25; \quad f_{22} = \frac{22}{100} = 0,22$$

   et le tableau de la distribution conjointe s'écrit

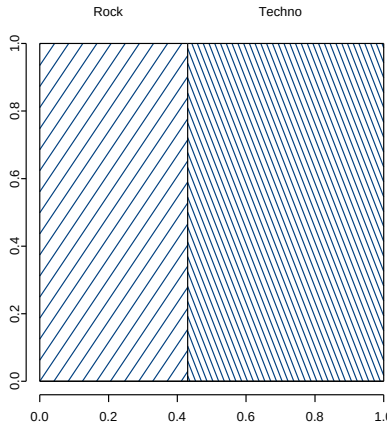
âge	Rock	Techno	Total
15 à 20 ans	0,18	0,35	0,53
21 à 30 ans	0,25	0,22	0,47
Total	0,43	0,57	1

2. Les distributions marginales sont celles qu'on trouve dans les *marges* (les totaux), à droite et en bas, du tableau de la distribution conjointe. A droite on a les total de chaque ligne de fréquences conjointes f_{ij}, et en bas, le total de chaque colonne.

   Pour la distribution marginale des goûts musicaux (colonnes), les fréquences marginales sont définies par $f_{\cdot j} = \frac{n_{\cdot j}}{n} = \sum_{i=1}^{2} f_{ij}$, pour $j = 1$ (Rock), 2 (Techno), soit

$$f_{\cdot 1} = \sum_{i=1}^{2} f_{i1} = 0,18 + 0,25 = 0,43 \qquad f_{\cdot 2} = \sum_{i=1}^{2} f_{i2} = 0,35 + 0,22 = 0,57$$

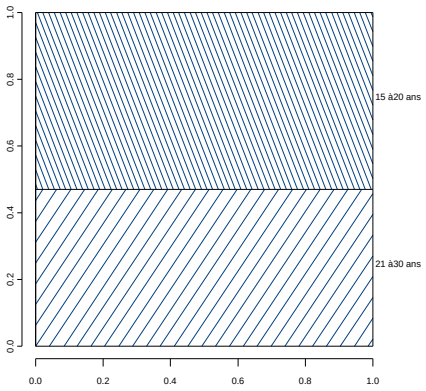
   La présentation graphique de cette distribution (en utilisant le carré unitaire) est



3. Les fréquences marginales de l'âge sont définies par $f_{i\cdot} = n_{i\cdot}/n = \sum_{j=1}^{2} f_{ij}$, pour $i = 1, 2$, ce qui donne

$$f_{1\cdot} = \sum_{j=1}^{2} f_{1j} = 0,18 + 0,35 = 0,53 \qquad f_{2\cdot} = \sum_{j=1}^{2} f_{2j} = 0,25 + 0,22 = 0,47$$

   La présentation graphique de cette distribution est



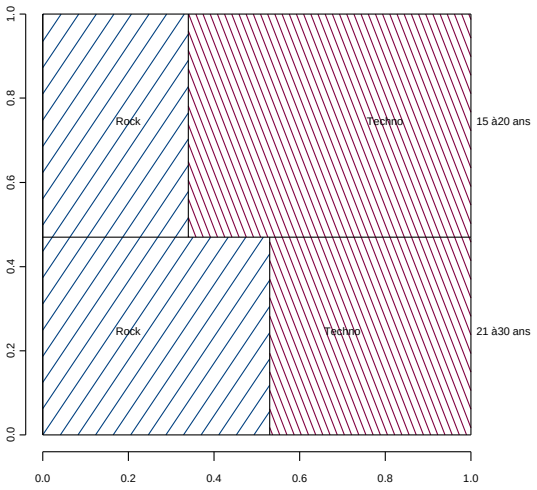
4. Les pourcentages lignes (distributions conditionnelles du goût musical selon l'âge) sont définis par $f_{ij|i} = n_{ij}/n_{i\cdot} = f_{ij}/f_{i\cdot}$, ce qui donne

$$f_{11|i=1} = \frac{f_{11}}{f_{1\cdot}} = \frac{0,18}{0,53} \approx 0,34 = 34\% \qquad f_{12|i=1} = \frac{f_{12}}{f_{1\cdot}} = \frac{0,35}{0,53} \approx 0,66 = 66\%$$
$$f_{21|i=2} = \frac{f_{21}}{f_{2\cdot}} = \frac{0,25}{0,47} \approx 0,53 = 53\% \qquad f_{22|i=2} = \frac{f_{22}}{f_{2\cdot}} = \frac{0,22}{0,47} \approx 0,47 = 47\%$$

et le tableau des distributions conditionnelles du goût selon l'âge est

âge	Rock	Techno	Total
15 à 20 ans	0,34	0,66	1
21 à 30 ans	0,53	0,47	1
En tout	0,43	0,57	1

On représente ces distributions sur le graphique de la question 3, en partitionnant chaque rectangle selon la distribution correspondante.



Les deux caractères sont ils indépendants ? Non, car les distributions conditionnelles sont clairement différentes.

5. Les pourcentages colonnes (distributions conditionnelles de l'âge selon le goût musical) sont

définis par $f_{ij|j} = n_{ij}/n_{\cdot j} = f_{ij}/f_{\cdot j}$, ce qui donne

$$f_{11|j=1} = \frac{f_{11}}{f_{\cdot 1}} = \frac{0,18}{0,43} \approx 0,42 = 42\% \qquad f_{12|j=2} = \frac{f_{12}}{f_{\cdot 2}} = \frac{0,35}{0,57} \approx 0,614 = 61,4\%$$
$$f_{21|j=1} = \frac{f_{21}}{f_{\cdot 1}} = \frac{0,25}{0,43} \approx 0,58 = 58\% \qquad f_{22|j=2} = \frac{f_{22}}{f_{\cdot 2}} = \frac{0,22}{0,57} \approx 0,386 = 38,6\%$$

et le tableau des distributions conditionnelles de l'âge selon le goût est donc

âge	Rock	Techno	En tout
15 à 20 ans	0,42	0,614	0,53
21 à 30 ans	0,58	0,386	0,47
Total	1	1	1

6. Effectifs attendus en cas d'indépendance ($e_{ij} = n_{i\cdot}n_{\cdot j}/n$)

âge	Rock	Techno	Total
15 à 20 ans	22.79	30.21	53
21 à 30 ans	20.21	26.79	47
Total	43	57	100

$$X^2 = \frac{(18-22.79)^2}{22.79} + \frac{(25-20.21)^2}{20.21} + \frac{(35-30.21)^2}{30.21} + \frac{(22-26.79)^2}{26.79} = \boxed{3.76}$$
$$v = \sqrt{3.76/100} = \boxed{0.194}$$

Association entre âge et goût musical plutôt faible (de l'ordre de 0.2) et non significative, car le khi-deux est plus petit que le seuil critique de 3.84.

Solution de l'exercice 3.5

Données de départ :

		Longueur du col			Total
		court 15-25km	moyen 25-35km	long 35-50km	
altitude	1400-2000	3	2	5	10
en mètres	2000-2300	0	4	5	9
	2300-2600	1	4	2	7
	Total	4	10	12	26

1. *Longueur et altitude moyennes approximées des cols*

   Altitude moyenne approximée $\bar{x}$: on calcule d'abord les centres de classes x_k et leurs effectifs n_k, puis on calcule l'approximation de la moyenne en appliquant l'expression : $\bar{x} = \frac{1}{n}\sum_{k=1}^{c}(n_k * x_k)$.

k	x_k	n_k	$n_k * x_k$
1400-2000	1700	10	17000
2000-2300	2150	9	19350
2300-2600	2450	7	17150
Total		26	53500

$$\bar{x} = \frac{10 * 1700 + 9 * 2150 + 7 * 2450}{26} = \frac{17000 + 19350 + 17150}{26} = \frac{53500}{26} = 2057.69$$

Longueur moyenne approximée $\bar{y}$: on procède de la même manière que pour l'altitude moyenne approximée.

k	y_k	n_k	$n_k * y_k$
15-25	20	4	80
25-35	30	10	300
35-50	42.5	12	510
Total		26	890

$$\bar{y} = \frac{4 * 20 + 10 * 30 + 12 * 42.5}{26} = \frac{80 + 300 + 510}{26} = \frac{890}{26} = 34.23$$

2. *Fréquences relatives caractérisant la distribution conjointe*

   On calcule les fréquences relatives $f_{ij} = \frac{n_{ij}}{n}$ à partir du tableau de contingence. Par exemple, $f_{11} = \frac{n_{11}}{n} = 3/26 = 0.115$.

		Longueur du col		
		court	moyen	long
		15-25km	25-35km	35-50km
altitude	1400-2000	0.115	0.077	0.192
en mètres	2000-2300	0	0.154	0.192
	2300-2600	0.038	0.154	0.077

3. *Distribution marginale selon la longueur du col*

   La distribution marginale de la longueur du col (en gras dans le tableau suivant) correspond aux fréquences $f_{.j}$, c'est-à-dire au total des fréquences relatives pour chaque longueur de col :

		Longueur du col			
		court	moyen	long	Total
		15-25km	25-35km	35-50km	
altitude	1400-2000	0.115	0.077	0.192	0.385
en mètres	2000-2300	0	0.154	0.192	0.346
	2300-2600	0.038	0.154	0.077	0.269
	Total	**0.154**	**0.385**	**0.462**	1

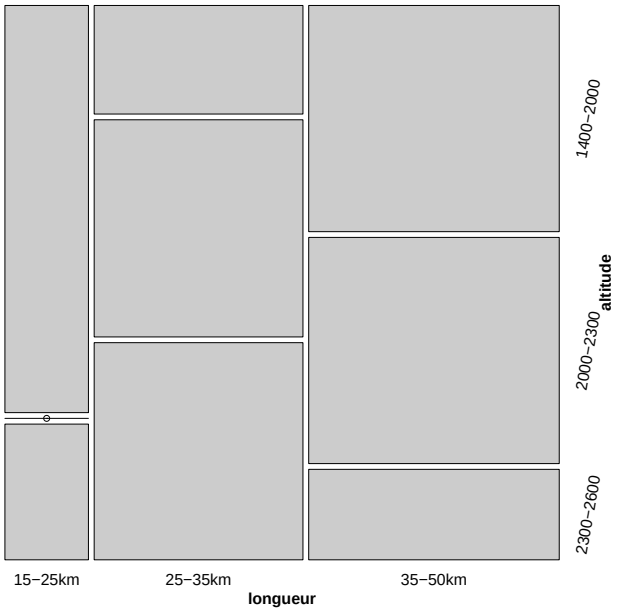
   Par ailleurs, la distribution marginale de l'altitude des cols a également été reportée dans la dernière colonne du tableau. Elle a été obtenue en sommant les fréquences relatives de chaque ligne du tableau.

4. *Distribution de l'altitude (distribution conditionnelle) au sein de chaque catégorie de longueur (court, moyen, long)*

   La distribution conditionnelle en fonction de la longueur, $f_{ij|j}$, correspond à la fréquence relative sachant la longueur du col j. La formule est la suivante : $f_{ij|j} = \frac{n_{ij}}{n_{.j}}$. On peut travailler de manière équivalente à partir des fréquences relatives ; dans ce cas $f_{ij|j} = \frac{f_{ij}}{f_{.j}}$. Par exemple, on trouve à partir des effectifs n_{ij} que $f_{12|2} = n_{12}/n_{.2} = 2/10 = 0.2$. De la même manière, $f_{12|2} = f_{12}/f_{.2} = 0.077/0.385 = 0.2$. On préférera utiliser les effectifs n_{ij} plutôt que les fréquences relatives pour atténuer les imprécisions dues aux arrondis.

		Longueur du col			
		court 15-25km	moyen 25-35km	long 35-50km	Total
altitude en mètres	1400-2000	0.75	0.2	0.417	0.385
	2000-2300	0	0.4	0.417	0.346
	2300-2600	0.25	0.4	0.167	0.269
	Total	1	1	1	1

On peut représenter cette distribution conditionnelle par un carré unitaire (mosaïque) :



5. *Commentaires*

La comparaison entre la distribution conditionnelle de l'altitude en fonction de la longueur du col et la distribution marginale de l'altitude nous montre qu'elles sont très différentes. En effet, la répartition des effectifs selon l'altitude diffère si l'on prend en compte la longueur du col. On peut donc en conclure que les deux variables ne sont pas indépendantes.

6. *Covariance approximée entre altitude et longueur des cols*

Grâce au point 1 de cet exercice, on connaît déjà les moyennes approximées $\bar{x}$ et $\bar{y}$, respectivement la moyenne de l'altitude et la moyenne de la longueur. On utilise la formule suivante pour calculer la covariance :

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \bar{x})(y_j - \bar{y})$$

Dans le cas de données groupées, on va prendre pour x_i et y_j le centre des classes et pour $\bar{x}$ et $\bar{y}$ les moyennes approximées calculées au premier point. Les termes $(x_i - \bar{x})(y_j - \bar{y})$ sont pondérés par la fréquence relative des couples (x_i, y_j) correspondants. Le tableau suivant contient les valeurs nécessaires au calcul de la covariance approximée :

		$y-\bar{y}$			
		-14.23	-4.23	8.27	Total
$x-\bar{x}$	-357.69	3	2	5	10
	92.31	0	4	5	9
	392.31	1	4	2	7
	Total	4	10	12	26

La covariance entre x et y se calcule de la manière suivante :

$$\begin{aligned} cov(x,y) &= \frac{1}{26}\Big(3(-357.66)(-14.23)+2(-357.66)(-4.23)+5(-357.66)(8.27)+ \\ &\quad +0(92.31)(-14.23)+4(92.31)(-4.23)+5(92.31)(8.27)+ \\ &\quad +1(392.31)(-14.23)+4(392.31)(-4.23)+2(392.31)(8.27)\Big)=1.11 \end{aligned}$$

Solution de l'exercice 3.6

$n = 30$ enfants choisis au hasard parmi les élèves inscrits en 6ème primaire dans une école privée genevoise

x_1 : âge en années
x_2 : moyenne annuelle en 5ème primaire
x_3 : nombre de frères et sœurs

Données :

$$\sum_{i=1}^{30}(x_{1i}-\bar{x}_1)^2=11.18 \qquad \sum_{i=1}^{30}(x_{2i}-\bar{x}_2)^2=145.87$$
$$\sum_{i=1}^{30}(x_{3i}-\bar{x}_3)^2=48.8 \qquad \sum_{i=1}^{30}(x_{1i}-\bar{x}_1)(x_{3i}-\bar{x}_3)=3.24$$

1. Variances et covariances de x_1 et x_3.

$$\begin{aligned} \text{var}(x_1) &= \tfrac{1}{30}\textstyle\sum_{i=1}^{30}(x_{1i}-\bar{x}_1)^2 &&= \tfrac{11.18}{30}=0.373 \\ \text{var}(x_3) &= \tfrac{1}{30}\textstyle\sum_{i=1}^{30}(x_{3i}-\bar{x}_3)^2 &&= \tfrac{48.8}{30}=1.627 \\ \text{cov}(x_1,x_3) &= \tfrac{1}{30}\textstyle\sum_{i=1}^{30}(x_{1i}-\bar{x}_1)(x_{3i}-\bar{x}_3) &&= \tfrac{3.24}{30}=0.108 \end{aligned}$$

2. Compléter le tableau C des corrélations ci-dessous et interpréter la valeur des coefficients.

	x_1	x_2	x_3
x_1	.	-0.11	.
x_2	.	.	.
x_3	.	-0.82	.

Sur la diagonale on a des 1 par définition (corrélation de chaque variable avec elle même). Les corrélations étant symétriques ($\text{corr}(x_i,x_j)=\text{corr}(x_j,x_i)$), il n'y a que la corrélation entre x_1 et x_3 à déterminer :

$$\begin{aligned} \text{corr}(x_1,x_3) &= \frac{\sum_i(x_{1i}-\bar{x}_1)(x_{3i}-\bar{x}_3)}{\sqrt{\big(\sum_i(x_{1i}-\bar{x}_1)^2\big)\big(\sum_i(x_{3i}-\bar{x}_3)^2\big)}} \\ &= \frac{3.24}{\sqrt{11.18\cdot 48.8}}=0.139 \end{aligned}$$

$$C=\begin{bmatrix} 1 & -0.11 & 0.14 \\ -0.11 & 1 & -0.82 \\ 0.14 & -0.82 & 1 \end{bmatrix}$$

Interprétation :

Corrélation négative importante entre moyenne annuelle x_2 et nombre x_3 de frères et sœurs. On a tendance à avoir une moins bonne moyenne lorsqu'on appartient à une grande fratrie.

Corrélation faible entre l'âge et les deux autres variables. On observe cependant une légère tendance à avoir de moins bonnes notes lorsqu'on est plus âgé en cinquième. On a aussi tendance à avoir plus de frères et sœurs lorsqu'on est plus âgé.

Solution de l'exercice 3.7

Soit les deux séries de données centrées et normées

$$x = \begin{pmatrix} -0.22 \\ -0.67 \\ 0.22 \\ 0.67 \end{pmatrix} \qquad y = \begin{pmatrix} 0.11 \\ 0.57 \\ 0.11 \\ -0.80 \end{pmatrix}$$

1. On a $\sum x_i = 0$ et $\sum y_i = 0$, et donc $\bar{x} = \bar{y} = 0$.
   De même, $\sum_i x_i^2 = 1$ et $\sum_i y_i^2 = 1$, d'où $\mathrm{var}(x) = 1/4$ et $\mathrm{var}(y) = 1/4$.

   La variance de données centrées normées est $1/n$ fois celle des données standardisées (centrées réduites).

2. Produit scalaire des séries de données centrées et normées.

$$\begin{aligned} x'y &= (-0.22) \cdot 0.11 + (-0.67) \cdot 0.57 + 0.22 \cdot 0.11 + 0.67 \cdot (-0.80) \\ &= \boxed{-0.9179} \end{aligned}$$

   Comme les données sont centrées, la covariance est $\mathrm{cov}(x, y) = -0.9179/4$, et par suite la corrélation

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} = \frac{-0.9179/4}{\sqrt{((1/4)(1/4))}} = -0.9179$$

   Leur produit scalaire des séries centrées normées donne directement la corrélation.

   On a une forte corrélation négative, les variables ont donc fortement tendance à évoluer en sens contraire.

Solution de l'exercice 3.8 (1ère partie)

Le tableau suivant est tiré d'un sondage sur la présence des femmes au Conseil d'Etat.

	Opinion (en pourcents)				
	pas nécessaire	utile sans plus	indispensable	ne sait pas	Total
Femmes	2.8	28.3	63.8	5.1	100
Hommes	4.9	44.3	45.9	4.9	100

On admet que l'échantillon sondé contient 60% de femmes et 40% d'hommes. Ceci est la distribution marginale du sexe.

1. Déterminer la distribution marginale des réponses, puis donner le tableau des distributions conditionnelles "lignes" que l'on aurait en cas d'indépendance entre le sexe et l'opinion.

   Pour obtenir la distribution marginale, on détermine tout d'abord la distribution conjointe en multipliant les proportion $f_{ij|i}$ données dans le tableau par la proportion totale $f_{i.}$ de cas dans la ligne concernée :

$$f_{ij} = f_{ij|i}\, f_{i.}$$

Comme on a ici des pourcents ($100 f_{ij|i}$), on doit faire par exemple pour le pourcent de sondés qui sont dans la cellule "femmes" et "pas nécessaire" :

$$2.8 \cdot 60/100 = 1.68$$

On obtient ainsi :

Distribution conjointe (pourcents)

	pas nécessaire	utile sans plus	indispensable	ne sait pas	Marginale
Femmes	1.68	16.98	38.28	3.06	60
Hommes	1.96	17.72	18.36	1.96	40
Marginale	3.64	34.7	56.64	5.02	100

La distribution marginale cherchée est alors le total des colonnes de ce tableau.

En cas d'indépendance, les distributions conditionnelles "ligne" en sont toutes égales et égales à la distribution marginale. Ainsi, le tableau demandé est :

Distributions conditionnelles ligne en cas d'indépendance (pourcents)

	pas nécessaire	utile sans plus	indispensable	ne sait pas	Total
Femmes	3.64	34.7	56.64	5.02	100
Hommes	3.64	34.7	56.64	5.02	100
Marginale	3.64	34.7	56.64	5.02	100

Remarque : La distribution marginale des réponses est la moyenne pondérée des deux distributions ligne, les pondérations étant 0.6 (60%) pour les femmes et 0.4 (40%) pour les hommes.

Solution de l'exercice 4.1

Données de départ :

		Revenu familial			
		Faible	Moyen	Elevé	Total
Occupation	Au foyer	8	26	6	40
	Ouvrier	16	40	14	70
	Cadre	6	62	12	80
	Professionnel	0	2	8	10
	Total	30	130	40	200

1. *Probabilité pour une personne choisie au hasard*
   (a) p(Au Foyer) = 40/200 = 0.2 (Le nombre d'individus au foyer divisé par le nombre total d'individus)
   (b) p(Ouvrier) = 70/200 = 0.35
   (c) p(Cadre) = 80/200 = 0.4
   (d) p(Professionnel) = 10/200 = 0.05
2. *Probabilité que le revenu de la personne choisie soit*
   (a) faible : p(faible) = 30/200 = 0.15 (Le nombre d'individus avec un revenu faible divisé par le nombre total d'individus)
   (b) moyen : p(moyen) = 130/200 = 0.65
   (c) élevé : p(élevé) = 40/200 = 0.2

3. *Probabilité que cette personne soit*

   (a) cadre avec un revenu élevé : $p(\text{Cadre, élevé}) = 12/200 = 0.06$

   (b) au foyer avec un revenu faible : $p(\text{Au Foyer, faible}) = 8/200 = 0.04$

   (c) professionnel avec un revenu moyen : $p(\text{Professionnel, moyen}) = 2/200 = 0.01$

4. *Probabilités conditionnelles*

   (a) Probabilité d'avoir un salaire élevé sachant qu'on est au foyer :
   $p(\text{élevé} \mid \text{au foyer}) = 6/40 = 0.15$ (nombre de personnes avec un revenu élevé divisé par le nombre de personnes au foyer)

   (b) Probabilité d'être cadre sachant qu'on a un revenu élevé :
   $p(\text{cadre} \mid \text{élevé}) = 12/40 = 0.3$ (nombre de cadres divisé par le nombre de personnes avec un revenu élevé)

   (c) Probabilité d'être ouvrier sachant qu'on a un revenu faible :
   $p(\text{ouvrier} \mid \text{faible}) = 16/30 = 0.53$ (nombre d'ouvriers divisé par le nombre de personnes avec un revenu faible)

5. *Indépendance des événements A et B*

   A est l'événement "la personne choisie est un ouvrier"

   B est l'événement "la personne choisie a un revenu faible".

   Que peut-on dire sur l'indépendance des événements A et B.

   On peut vérifier l'indépendance entre deux événements en comparant $p(A)$ et $p(A|B)$. Si ces deux probabilités sont égales, alors les deux événements sont indépendants. Dans cet exercice on a $p(A) = p(\text{Ouvrier}) = 0.35$, et $p(A|B) = P(\text{Ouvrier} \mid \text{Faible}) = 0.53$. Étant donné que $0.35 \neq 0.53$, on peut conclure que les événements A et B ne sont pas indépendants.

Solution de l'exercice 4.2

Petit rappel :

$$\begin{aligned} p(A,B) &= p(A \text{ et } B) \\ p(A|B) &= \frac{p(A,B)}{p(B)} \\ p(A \text{ ou } B) &= p(A) + p(B) - p(A,B) \end{aligned}$$

1. *Déduire les probabilités suivantes :*

   (a) $p(A,C) = p(C|A)p(A) = \frac{1}{5} * \frac{1}{2} = \frac{1}{10}$

   (b) $p(B,C) = p(B|C)p(C) = \frac{1}{2} * \frac{1}{4} = \frac{1}{8}$

   (c) $p(A,B,C) = p(A|B,C)p(B,C) = \frac{2}{3} * \frac{1}{8} = \frac{1}{12}$

   (d) $p(A|B) = \frac{p(A,B)}{p(B)} = \frac{\frac{1}{8}}{\frac{1}{3}} = \frac{3}{8}$

   (e) $p(C|B) = \frac{p(B,C)}{p(B)} = \frac{\frac{1}{8}}{\frac{1}{3}} = \frac{3}{8}$

   (f) $p\Big((BetC) \mid A\Big) = \frac{p(A,B,C)}{p(A)} = \frac{\frac{1}{12}}{\frac{1}{2}} = \frac{1}{6}$

   (g) $p(A \text{ ou } B) = p(A) + p(B) - p(A,B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{8} = \frac{17}{24}$

   (h) $p(A \text{ ou } C) = p(A) + p(C) - p(A,C) = \frac{1}{2} + \frac{1}{4} - \frac{1}{10} = \frac{13}{20}$

2. *Paires d'événements mutuellement exclusifs parmi A, B et C*

   Pour que deux événements soient mutuellement exclusifs, il faut que leur probabilité conjointe soit égale à 0. Si $p(A, B) = 0$, A et B sont mutuellement exclusifs car ils ne peuvent pas se produire en même temps. Dans cet exercice on constate que :

   - $p(A, B) = \frac{1}{8} \neq 0$
   - $p(A, C) = \frac{1}{10} \neq 0$
   - $p(B, C) = \frac{1}{8} \neq 0$

   Il n'y a donc pas de paire d'événements mutuellement exclusifs.

3. *Quelles paires d'événements ne sont pas indépendants ?*

   On peut vérifier l'indépendance de deux manières :

   - A et B sont indépendants si $p(A|B) = p(A)$
   - A et B sont indépendants si $p(A, B) = p(A)p(B)$

   Dans cet exercice, on dispose des valeurs nécessaires pour affirmer que

   - $p(A)p(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \neq p(A, B)$
   - $p(A)p(C) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \neq p(A, C)$
   - $p(B)p(C) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12} \neq p(B, C)$

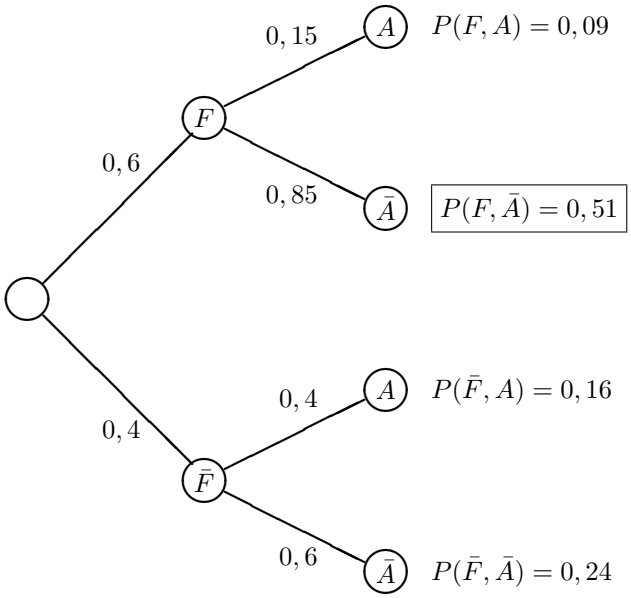
   On peut conclure qu'il n'y pas d'événements indépendants, puisqu'aucune des paires d'événements ne satisfait la condition d'indépendance $p(A, B) = p(A)p(B)$.

Solution de l'exercice 4.3

Épidémie dans une crèche :
$F =$ « fille », $\bar{F} =$ « garçon », $A =$ « absent », $\bar{A} =$ « présent ».

Données : 40 garçons, 60 filles $\Rightarrow p(F) = 0,6$, $p(\bar{F}) = 0,4$;
probabilité d'être absente si l'on est une fille $= p(A|F) = 0,15 \Rightarrow p(\bar{A}|F) = 0,85$;
probabilité d'être absent si l'on est un garçon $= p(A|\bar{F}) = 0,4 \Rightarrow p(\bar{A}|\bar{F}) = 0,6$.

1. Présentation arborescente :



2. $p(F, \bar{A})$: probabilité d'être une fille et présente.

$$p(F, \bar{A}) = p(F)p(\bar{A}|F) = 0.6 \cdot 0,85 = 0,51 = 51\%$$

$p(\bar{F}, \bar{A})$: probabilité d'être un garçon et présent.

$$p(\bar{F}, \bar{A}) = p(\bar{F})p(\bar{A}|\bar{F}) = 0,4 \cdot 0,6 = 0,24 = 24\%$$

$p(\bar{A})$: probabilité qu'un enfant soit présent.

$$p(\bar{A}) = p(F, \bar{A}) + p(\bar{F}, \bar{A}) = 0,51 + 0,24 = 0,75 = 75\%$$

3. Probabilité qu'un enfant choisi au hasard parmi les présents soit une fille. C'est la probabilité d'être une fille sachant que l'enfant est présent :

$$p(F|\bar{A}) = \frac{p(F, \bar{A})}{p(\bar{A})} = \frac{0,51}{0,75} = 0,68 = \boxed{68\,\%}$$

Solution de l'exercice 4.6

Calcul de la covariance.

a.

distribution donnée

$X \setminus Y$	-1	1	
-1	1/8	1/8	1/4
1	3/8	3/8	3/4
	1/2	1/2	1

distribution si indépendance

$X \setminus Y$	-1	1	
-1	1/8	1/8	1/4
1	3/8	3/8	3/4
	1/2	1/2	1

Les distributions sont les mêmes, il y a donc indépendance de X et Y et la covariance $\mathrm{Cov}(X, Y)$ est par conséquent nulle. On vérifie qu'on a en effet $\mathrm{E}(X) = 0,5$, $\mathrm{E}(Y) = 0$ et

$$\mathrm{Cov}(X, Y) \quad = \quad \frac{1}{8}(-1,5)(-1) + \frac{1}{8}(-1,5)1 + \frac{3}{8}0{,}5(-1) + \frac{3}{8}0{,}5 \cdot 1 = 0$$

b.

distribution donnée

$X \setminus Y$	-1	1	
-1	1/4	1/4	1/2
1	0	1/2	1/2
	1/4	3/4	1

distribution si indépendance

$X \setminus Y$	-1	1	
-1	1/8	3/8	1/2
1	1/8	3/8	1/2
	1/4	3/4	1

Comme $p(X = -1, Y = -1) = 1/4 \neq p(X = -1)p(Y = -1) = 1/8$, X et Y ne sont pas indépendants. On a $\mathrm{E}(X) = 0$, $\mathrm{E}(Y) = 0,5$ et la covariance $\mathrm{Cov}(X, Y)$ est

$$\mathrm{Cov}(X, Y) \quad = \quad \frac{1}{4}(-1)(-1{,}5) + \frac{1}{4}(-1)(0{,}5) + 0 \cdot 1(-1{,}5) + \frac{1}{2}1 \cdot (0{,}5) = \frac{1}{2} \neq 0 \ .$$

c.

distribution donnée

$X \setminus Y$	-1	0	1	
0	0	1/2	0	1/2
1	1/4	0	1/4	1/2
	1/4	1/2	1/4	1

distribution si indépendance

$X \setminus Y$	-1	0	1	
0	1/8	1/4	1/8	1/2
1	1/8	1/4	1/8	1/2
	1/4	1/2	1/4	1

Comme $p(X = 0, Y = -1) = 0 \neq p(X = 0)p(Y = -1) = 1/8$, X et Y ne sont pas indépendants. On a $\mathrm{E}(X) = 0{,}5$, $\mathrm{E}(Y) = 0$ et la covariance $\mathrm{Cov}(X, Y)$ est

$$\begin{aligned}\mathrm{Cov}(X, Y) &= 0(-0{,}5)(-1) + \frac{1}{2}(-0{,}5)0 + 0(-0{,}5)1 + \frac{1}{4}0{,}5(-1) + 0 \cdot 0{,}5 \cdot 0 + \frac{1}{4}0{,}5 \cdot 1 \\ &= 0 \ .\end{aligned}$$

On constate qu'une covariance nulle n'implique pas l'indépendance.

Solution de l'exercice 4.8

Soit les événements : L = "étranger originaire d'un pays limitrophe"; E = "étranger originaire d'Europe"; A = "étranger originaire d'Afrique".
Les informations données s'écrivent formellement : $p(L|E) = .552$, $p(L) = .444$, $p(A) = .033$.

1. Calculer la probabilité $p(A \cup L)$ qu'un étranger résidant à Genève soit originaire d'un pays africain ou d'un pays limitrophe de la Suisse.

   Aucun pays limitrophe n'étant en Afrique, on a $A \cap L = \emptyset$, et par suite $p(A \cap L) = 0$. Ainsi

$$p(A \cup L) \quad = \quad p(A) + p(L) - \underbrace{p(A \cap L)}_{0} = .444 + .033 = .477 = \boxed{47.7\%}$$

2. Calculer la probabilité $p(E)$ qu'un étranger résidant à Genève soit originaire d'Europe.

   Remarquons tout d'abord que comme $L \subset E$, on a $p(L \cap E) = p(L)$. Par ailleurs, par définition on a $p(\bar{L} \cap E) = p(\bar{L}|E)p(E)$. Ainsi :

$$\begin{aligned} p(E) &= p(L \cap E) + p(\bar{L} \cap E) \\ &= p(L) + p(\bar{L}|E)p(E) \end{aligned}$$

   En regroupant les termes en $p(E)$ et en notant que $p(L|E) = 1 - p(\bar{L}|E)$, il vient

$$\begin{aligned} \big(1 - p(\bar{L}|E)\big)p(E) &= p(L) \\ p(E) &= \frac{p(L)}{1 - p(\bar{L}|E)} \\ &= \frac{p(L)}{p(L|E)} = \frac{.444}{.552} = \boxed{80.4\%} \end{aligned}$$

Solution de l'exercice 4.11

Il s'agit ici d'analyser deux variables. On a un ami qui est parfois ivre (3 soirs sur 7) et chute parfois de son vélo (1 soir sur 3). De plus on sait qu'il est ivre et qu'il tombe la même soirée deux soirs sur sept.

1. Préciser les deux événements simples et leurs contraires. Donner le tableau des probabilités simples et conjointes.

   Les deux événements simples sont, ivre (I) et chute (C). Les contraires sont donnés par pas ivre ($\bar{I}$) et pas chute ($\bar{C}$). Les informations fournies peuvent être formalisées comme suit

$$\begin{aligned} p(I) = 3/7 = 0.42 \quad &\Leftrightarrow \quad p(\bar{I}) = 4/7 = 0.58 \\ p(C) = 1/3 = 0.33 \quad &\Leftrightarrow \quad p(\bar{C}) = 2/3 = 0.67 \\ p(I \cap C) = 2/7 = 0.285 \end{aligned}$$

   Ces informations peuvent être résumées dans le tableau de probabilités suivant. On peut ensuite utiliser ces informations pour compléter le tableau. Par exemple, la probabilité associée à l'événement ivre (I) et non chute ($\bar{C}$), c'est-à-dire $p(I \cap \bar{C})$ est calculée en utilisant $p(I) - p(I \cap C) = 3/7 - 2/7 = 1/7 = 0.135$. En fait, on utilise simplement le fait que les fréquences marginales sont les sommes des fréquences relatives correspondantes.

	chute (C)	pas chute ($\bar{C}$)	Somme
Ivre (I)	2/7	1/7	3/7
pas Ivre ($\bar{I}$)	1/21	11/21	4/7
Somme	1/3	2/3	1

2. On sait que notre ami va être ivre. Quelle est la probabilité qu'il chute à vélo ? Il s'agit ici d'une probabilité conditionnelle. Quelle est la probabilité qu'il chute à vélo sachant qu'il est ivre ?

$$p(C \mid I) = \frac{p(C \cap I)}{p(I)} = \frac{2/7}{3/7} = \frac{2}{3} = 66.7\%$$

3. La consommation d'alcool influence-t-elle son habileté à vélo ?

   Pour répondre à cette question, il s'agit en fait de voir s'il y a ou pas indépendance entre les deux variables. En cas d'indépendance, la probabilité $p(C|I)$ de chuter en ayant bu de l'alcool serait égale à la probabilité marginale $p(C)$. Ici on a :

$$p(C \mid I) = \frac{2}{3} \quad > \quad p(C) = \frac{1}{3}$$

   On voit donc que le fait d'être ivre double la probabilité de chuter par rapport à la situation marginale. Il y a donc clairement une influence.

   La probabilité de chuter de mon ami sous l'effet de l'alcool est égale à 8 fois la probabilité de chuter sans être ivre ($p(C|I) = 2/3$ contre $p(C|\bar{I}) = 1/12$).

Solution de l'exercice 5.1

1. La distribution de la variable aléatoire X représentant l'âge d'un individu choisi au hasard se construit comme suit :

x	15	20	35
$p(X = x)$	1/4	2/4	1/4

   Il a trois réalisations possibles x (15, 20 et 35) de la variable aléatoire X. L'âge 15 apparaît une fois sur quatre, l'âge 20, deux fois sur quatre et l'âge 35, une fois sur quatre.

   L'espérance μ de X est donnée par

$$\mu = \mathrm{E}(X) = \frac{1}{4}15 + \frac{2}{4}20 + \frac{1}{4}35 = 22.5$$

   et la variance σ^2 de X par

$$\sigma^2 = \mathrm{Var}(X) = (\frac{1}{4}15^2 + \frac{2}{4}20^2 + \frac{1}{4}35^2) - (22.5)^2 = 56.25$$

2. On considère un échantillon de taille 2 avec remise. On nous demande de calculer :

   a) Tous les échantillons de taille 2 possibles ainsi que leurs probabilités. Il y a $4^2 = 16$ échantillons possibles. Ils sont énumérés dans les deux premières colonnes du tableau ci-dessous. Tous les échantillons sont équiprobables et ont donc une probabilité de 1/16 d'être choisi.

tirage	échantillon	probabilité	moyenne	variance	minimum
(1,1)	(35,35)	1/16	35	0	35
(1,2)	(35,20)	1/16	27.5	56.25	20
(1,3)	(35,15)	1/16	25	100	15
(1,4)	(35,20)	1/16	27.5	56.25	20
(2,1)	(20,35)	1/16	27.5	56.25	20
(2,2)	(20,20)	1/16	20	0	20
(2,3)	(20,15)	1/16	17.5	6.25	15
(2,4)	(20,20)	1/16	20	0	20
(3,1)	(15,35)	1/16	25	100	15
(3,2)	(15,20)	1/16	17.5	6.25	15
(3,3)	(15,15)	1/16	15	0	15
(3,4)	(15,20)	1/16	17.5	6.25	15
(4,1)	(20,35)	1/16	27.5	56.25	20
(4,2)	(20,20)	1/16	20	0	20
(4,3)	(20,15)	1/16	17.5	6.25	15
(4,4)	(20,20)	1/16	20	0	20

b) La moyenne $\bar{x}_r$, la variance s_r^2 et le minimum $x_{\min}$ de chaque échantillon sont donnés dans les colonnes 4, 5 et 6. Par exemple, pour l'échantillon (1,1),

$$\begin{aligned} \bar{x}_r &= \frac{1}{2}\sum_{i=1}^{2} x_i = (35+35)/2 = 35 \\ s_r^2 &= \frac{1}{2}\sum_{i=1}^{2}(x_i-\bar{x}_r)^2 = \frac{1}{2}[(35-35)^2+(35-35)^2] = 0 \\ x_{\min} &= 35 \end{aligned}$$

c) Donner le tableau de probabilités de la moyenne $\bar{X}_r$, de la variance S_r^2 et du minimum $X_{\min}$. Il s'agit de la même démarche qu'au point 1 mais cette fois la variable aléatoire est soit $\bar{X}_r$, soit S_r^2 ou $X_{\min}$. Les réalisations possibles de $\bar{X}_r$ sont (cf. tableau) $\{35, 27.5, 25, 20, 17.5, 15\}$, celles de S_r^2 sont $\{0, 56.25, 100, 6.25\}$ et celles de $X_{\min}$ sont $\{35, 20, 15\}$. On peut donc construire les trois tableaux de probabilités suivant :

$\bar{x}_r$	15	17.5	20	25	27.5	35
$p(\bar{x}_r)$	1/16	4/16	4/16	2/16	4/16	1/16

s_r^2	0	6.25	56.25	100
$p(s_r^2)$	6/16	4/16	4/16	2/16

$x_{\min}$	15	20	35
$p(x_{\min})$	7/16	8/16	1/16

d) Calculons $\mathrm{E}(\bar{X}_r)$, $\mathrm{Var}(\bar{X}_r)$, $\mathrm{E}(S_r^2)$ et $\mathrm{E}(X_{\min})$,

$$\begin{aligned} \mathrm{E}(\bar{X}_r) &= (\frac{1}{16}15 + \frac{4}{16}17.5 + \frac{4}{16}20 + \frac{2}{16}25 + \frac{4}{16}27.5 + \frac{1}{16}35) = 22.5 \\ \mathrm{Var}(\bar{X}_r) &= (\frac{1}{16}15^2 + \frac{4}{16}17.5^2 + \frac{4}{16}20^2 + \frac{2}{16}25^2 + \frac{4}{16}27.5^2 + \frac{1}{16}35^2) - 22.5^2 = 28.125 \\ \mathrm{E}(S_r^2) &= (\frac{6}{16}0 + \frac{4}{16}6.25 + \frac{4}{16}56.25 + \frac{2}{16}100) = 28.125 \\ \mathrm{E}(X_{\min}) &= (\frac{7}{16}15 + \frac{8}{16}20 + \frac{1}{16}35) = 18.75 \end{aligned}$$

3. On travaille maintenant avec un échantillon sans remise. On nous demande :

   a) Tous les échantillons de taille 2 possibles ainsi que leur moyenne $\bar{x}_s$. On procède comme dans le cas avec remise, mais on exclut les échantillons obtenus en tirant deux fois le même cas. Les seuls échantillons possibles sont donc :

tirage	échantillon	moyenne
(1,2)	(35,20)	27.5
(1,3)	(35,15)	25
(1,4)	(35,20)	27.5
(2,1)	(20,35)	27.5
(2,3)	(20,15)	17.5
(2,4)	(20,20)	20
(3,1)	(15,35)	25
(3,2)	(15,20)	17.5
(3,4)	(15,20)	17.5
(4,1)	(20,35)	27.5
(4,2)	(20,20)	20
(4,3)	(20,15)	17.5

   b) Donner le tableau de probabilités de la moyenne $\bar{X}_s$. Les réalisations possibles de $\bar{X}_s$ sont $\{17.5, 20, 25, 27.5\}$, et le tableau des probabilités est,

$\bar{x}_s$	17.5	20	25	27.5
$p(\bar{x}_s)$	4/12	2/12	2/12	4/12

   c) Calculons $\mathrm{E}(\bar{X}_s)$, $\mathrm{Var}(\bar{X}_s)$

$$\begin{aligned} \mathrm{E}(\bar{X}_s) &= \frac{4}{12}17.5+\frac{2}{12}20+\frac{2}{12}25+\frac{4}{12}27.5 = 22.5 \\ \mathrm{Var}(\bar{X}_s) &= (\frac{4}{12}17.5^2+\frac{2}{12}20^2+\frac{2}{12}25^2+\frac{4}{12}27.5^2) - 22.5^2 = 18.75 \end{aligned}$$

   On voit que la moyenne ne change pas et que la moyenne des moyennes d'échantillon est égale à la moyenne de la population $\mu = \mathrm{E}(\bar{X}_r) = \mathrm{E}(\bar{X}_s) = 22.5$. Les moyennes des échantillons obtenus sans remise sont moins dispersées que dans le cas avec remise ($\mathrm{Var}\,\bar{X}_s = 18.75$ au lieu de $\mathrm{Var}\,X_r = 28.125$). Sans remise on ne peut pas, par exemple, obtenir la moyenne maximale 35 que l'on obtient avec remise en tirant deux fois le cas 1.

Solution de l'exercice 5.2

On s'intéresse à une grandeur X qui est distribuée comme suit dans une population de taille $m = 10$

x	-4	2	4
$p(X = x)$	1/10	5/10	4/10

La probabilité associée à la réalisation $x = 4$ est de 4/10 car on sait que la somme des probabilités doit être égale à 1.

Tirage d'un échantillon de taille 7 avec remises.

1. Calcul de la variance $\mathrm{Var}(\bar{X})$. On commence par calculer l'espérance μ de X

$$\mu = \mathrm{E}(X) = \frac{1}{10}(-4) + \frac{5}{10}2 + \frac{4}{10}4 = 2.2$$

puis on calcule la variance σ^2 de X

$$\sigma^2 = \mathrm{Var}(X) = (\frac{1}{10}(-4)^2 + \frac{5}{10}2^2 + \frac{4}{10}4^2) - 2.2^2 = 5.16$$

Enfin, pour des tirages i.i.d.(μ, σ^2), ce qui est le cas lorsqu'on procède avec remises, on a $\mathrm{Var}(\bar{X}) = \sigma^2/n$. D'où

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \frac{1}{n}\,\sigma^2 \\ &= \frac{1}{7}\,5.16 = 0.737 \end{aligned}$$

2. Calcul de $\mathrm{E}(S^2)$ où S^2 est la variance de l'échantillon avec remises. Pour des tirages i.i.d.(μ, σ^2), on a $\mathrm{E}(S^2) = \sigma^2(n-1)/n$. Ainsi

$$\begin{aligned} \mathrm{E}(S^2) &= \frac{n-1}{n}\,\sigma^2 \\ &= \frac{7-1}{7}\,5.16 = 4.42 \end{aligned}$$

Tirage d'un échantillon de taille 7 sans remise.

3. Calculer $\mathrm{E}(\bar{X}_s)$. On sait que $\mathrm{E}(\bar{X}_s) = \mathrm{E}(X) = \mu$. Donc $\mathrm{E}(\bar{X}_s) = 2.2$.
4. Calculer $\mathrm{Var}(\bar{X}_s)$. Dans le cas sans remise on a

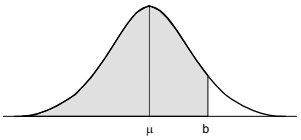
$$\begin{aligned} \mathrm{Var}(\bar{X}_s) &= \Big(\frac{m-n}{m-1}\Big)\,\frac{\sigma^2}{n} \\ &= \Big(\frac{10-7}{10-1}\Big)\,\frac{5.16}{7} = \frac{1}{3}\,0.737 = 0.246 \end{aligned}$$

Commentaire : On voit que la variance d'échantillon S^2 sous estime la variance puisque son espérance mathématique $\mathrm{E}(S^2) = 4.42$ est inférieure à $\sigma^2 = 5.16$. On relève également que le tirage sans remise est ici clairement plus efficace pour estimer la moyenne, puisque la variance $\mathrm{Var}(\bar{X}_s) = 0.246$ est inférieure à la variance avec remises qui vaut $\mathrm{Var}(\bar{X}) = 0.737$. La différence importante relevée s'explique ici par le fort taux d'échantillonnage de 7/10.

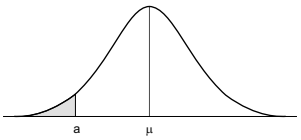
Solution de l'exercice 6.1

$Z \sim N(0,1)$, donc $\mu = 0$, $\sigma^2 = 1$ et $\sigma = 1$.

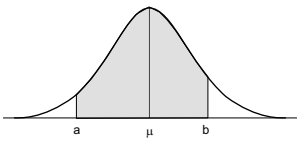
1. (a) $p(Z < 2.14) = ?$
   Se lit dans la table : ligne 2.1, colonne .04
   $\Rightarrow p(Z < 2.14) = .9838 = \boxed{98.38\%}$



   (b) $p(Z < -2.14) = ?$
   Par symétrie : $p(Z < -2.14) = p(Z > 2.14)$
   Par complémentarité : $p(Z > 2.14) = 1 - p(Z < 2.14)$
   $\Rightarrow p(Z < -2.14) = 1 - .9838 = .0162 = \boxed{1.62\%}$



   (c) $p(-2.14 < Z < 2.14) = ?$
   Probabilité à gauche de 2.14 moins celle à gauche de -2.14
   $p(-2.14 < Z < 2.14) = p(Z < 2.14) - p(Z < -2.14)$
   $\Rightarrow p(-2.14 < Z < 2.14) = .9838 - .0162 = .9676 = \boxed{96.76\%}$



2. (a) $p(Z < a) = .8686$
Dans la table, $p = .8686$ est à l'intersection de la ligne 1.1, et de la colonne .02
$p(Z < a) = 86.86\% \Leftrightarrow a = \boxed{1.12}$.

(b) $p(Z < a) = .9719$
Dans la table, $p = .9719$ est à l'intersection de la ligne 1.9 et de la colonne .01
$\Rightarrow p(Z < a) = 97.19\% \Leftrightarrow a = \boxed{1.91}$.

(c) $p(Z < a) = .2912$ ($p < .5 \Leftrightarrow a < 0$)
La table ne donne que des probabilités supérieures à 0.5.
Par complémentarité : $1 - p(Z < a) = p(Z > a) = 1 - .2912 = .7088$.
Par symétrie : $p(Z > a) = p(Z < -a) = .7088$.
Cette probabilité est à l'intersection de la ligne 0.5 et de la colonne .05.
$-a = .55 \Leftrightarrow a = \boxed{-.55}$.

(d) $p(Z < a) = .7$
Dans la table, on a $z_1 = 0.52$ pour $p_1 = .6985$ et $z_2 = 0.53$ pour $p_2 = .7019$
On approxime a par interpolation linéaire :

$$\frac{a - z_1}{z_2 - z_1} = \frac{a - .52}{.01} \quad = \quad \frac{p - p_1}{p_2 - p_1} = \frac{.0015}{.0034}$$

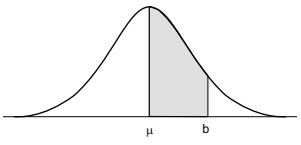
d'où $a = .52 + (.01)(.0015/.0034) = .5244 \Rightarrow p(Z < a) = 70\% \Leftrightarrow a \simeq \boxed{0.524}$.

$X \sim N(2, 1)$, donc $\mu = 2$, $\sigma^2 = 1$ et $\sigma = 1 \quad \Rightarrow Z = \dfrac{X - \mu}{\sigma} = X - 2 \sim N(0, 1)$.

3. (a) $p(X < 3.98) = p(Z < 3.98 - 2) = p(Z < 1.98) = \boxed{97.61\%}$
Dans la table, à l'intersection de la ligne 1.9 et de la colonne .08, on lit en effet 0.9761.

(b) $p(X < 1.5) = p(Z < 1.5 - 2) = p(Z < -.5) = 1 - p(Z < .5)$
$= 1 - .6915 = .3085 = \boxed{30.85\%}$
La valeur .6915 se lit dans la table, ligne 0.5, colonne .00.

(c) $p(X \in [2, 3]) = p(X < 3) - p(X < 2)$
$p(X < 3) = p(Z < 1) = .8413$ et, comme $\mu = 2$ est aussi la médiane, $p(X < 2) = .5$.
$\Rightarrow p(X \in [2, 3]) = .8413 - .5 = .3413 = \boxed{34.13\%}$



4. $p(X < a) = p(Z < z_p) = p$, avec $z_p = \dfrac{a - \mu}{\sigma}$ et donc $a = \sigma z_p + \mu = z_p + 2$

(a) $p(X < a) = p(Z < z_p) = .7454$, valeur qui se trouve à l'intersection de la ligne .6 et de la colonne .06, d'où
$z_p = .66$ et $a = z_p + 2 = \boxed{2.66}$

(b) $p(2 - a < X < 2 + a) = p(-a < Z < a) = .7458$
$p(Z < -a) = 1 - p(Z < a)$, d'où
$p(-a < Z < a) = p(Z < a) - p(Z < -a) = 2p(Z < a) - 1 = .7458$. On en déduit
$p(Z < a) = (.7458 + 1)/2 = 1.7458/2 = .8729$, valeur qui est dans la table à la ligne 1.1 et la colonne .04 $\Rightarrow a = \boxed{1.14}$

$Y \sim N(2, 9)$, donc $\mu = 2$ et $\sigma = \sqrt{9} = 3 \quad \Rightarrow Z = \dfrac{X - \mu}{\sigma} = \dfrac{X - 2}{3} \sim N(0, 1)$.

5. (a) $p(Y < 4.25) = p(Z < \frac{4.25-2}{3}) = p(Z < .75) = \boxed{77.34\%}$
   Dans la table, à l'intersection de la ligne 0.7 et de la colonne .05, on lit en effet 0.7734.

   (b) $p(Y < 1.5) = p(Z < \frac{1.5-2}{3}) = p(Z < -.17) = 1 - .5675 = \boxed{43.25\%}$.
   On a en effet, $p(Z < -.17) = 1 - p(Z < .17)$, et la table donne (ligne .1, colonne .07) $p(Z < .17) = .5675$.

6. $p(Y < a) = p(Z < z_p) = p$, avec $z_p = \dfrac{a-\mu}{\sigma}$ et donc $a = \sigma z_p + \mu = 3z_p + 2$

   (a) $p(Y < a) = p(Z < z_p) = .64$
   $\Rightarrow z_p = .36$ (.6406 se trouve ligne .3, colonne .06)
   $\Rightarrow a = 3 \cdot 0.36 + 2 = \boxed{3.08}$.

   (b) $p(a < Y < 2.5) = p(Y < 2.5) - p(Y < a) = 0.3788$
   $\Rightarrow p(Y < a) = p(Y < 2.5) - 0.3788$.
   Comme $p(Y < 2.5) = p(Z < \frac{0.5}{3} = .17) = .5675$, on a donc
   $p(Y < a) = p(Z < z_p) = .5675 - .3788 = .1887$.
   Cette probabilité étant plus petite que 0.5, on cherche $-z_p$ tel que $p(Z < -z_p) = 1 - .1887 = .8113$. Dans la table, la valeur la plus proche est .8106 à la ligne .8 et la colonne .08, d'où $-z_p = .88$ et $z_p = -.88$.
   Finalement : $a = 3(-0.88) + 2 = \boxed{-.64}$.

   (c) $p(2 - a < Y < 2 + a) = p(\frac{-a}{3} < Z < \frac{a}{3}) = 0.9$.
   On a $p(-\frac{a}{3} < Z < \frac{a}{3}) = 0.9$, et en développant comme en 4.b,
   $p(Z < \frac{a}{3}) = (1 + 0.9)/2 = 0.95$.
   On a donc $\frac{a}{3} = z_{.95}$. Dans la table, cette valeur est à la ligne 1.6 entre .9495 (colonne .04) et .9505 (colonne .05), d'où $\frac{a}{3} = 1.645$.
   Finalement $a = 3 \cdot 1.645 = \boxed{4.935}$.

Solution de l'exercice 6.2

On a : X : Le nombre de spectateurs
$X \sim N(24000, 16000000) \Rightarrow \mu = 24000$
$\sigma = 4000$

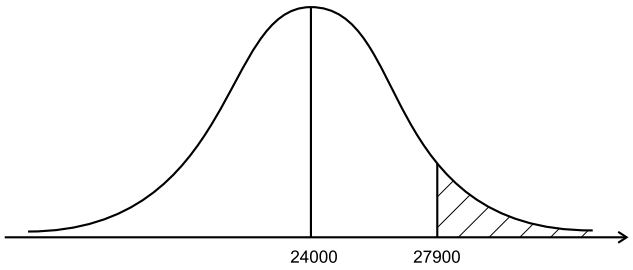
On cherche la probabilité que, lors d'un match choisi au hasard, il y ait au moins 90% de l'assistance maximum de l'année précédente (assistance maximum = 31000). Formellement, cette probabilité est :

$$p(X > 0.9 \cdot 31000) = p(X > 27900)$$

En centrant et en réduisant, il vient :

$$\begin{aligned} p(X > 27900) &= p(Z > \frac{27900 - \mu}{\sigma}) \\ &= p(Z > \frac{27900 - 24000}{4000}) \\ &= p(Z > 0.98) \\ &= 1 - p(Z < 0.98) \\ &= 1 - 0.8365 = 0.1635 = \boxed{16.35\%} \end{aligned}$$

La valeur $p(Z < 0.98) = 0.8365$ se lit dans la table de la loi normale. On peut aussi l'obtenir dans Excel avec la formule =LOI.NORMALE.STANDARD(0.98).



Solution de l'exercice 6.3

X poids d'un bonbon avec $X \sim N(\mu = 68\,\mathrm{g}, \sigma = 3\,\mathrm{g})$. En tout $m = 600$ bonbons.

1. Combien ont plus de 72 grammes ?

   On détermine la proportion $p(X > 72)$ de bonbons de plus de 72 grammes, puis en multipliant cette proportion par le nombre m de bonbons on obtient le nombre cherché

$$\begin{aligned} p(X > 72) &= p(Z > \frac{72-68}{3} = 1.33) = 1 - .9082 = .0918 \\ n_{>72} &= 600 \cdot 0.0918 \simeq \boxed{55} \end{aligned}$$

2. Combien ont moins de 64 grammes ?

$$\begin{aligned} p(X < 64) &= p(Z < \frac{64-68}{3} = -1.33) = p(Z > 1.33) = .0918 \\ n_{<64} &= 600 \cdot 0.0918 \simeq \boxed{55} \end{aligned}$$

3. Combien entre 65 et 71 grammes ?

$$\begin{aligned} p(65 < X < 71) &= p(X < 71) - p(X < 65) \\ &= p(Z < \frac{3}{3} = 1) - p(Z < \frac{-3}{3} = -1) \\ &= 2p(Z < 1) - 1 = 2 \cdot 0.8413 - 1 = 0.6826 \\ n_{[65,71]} &= 600 \cdot 0.6826 \simeq \boxed{409} \end{aligned}$$

4. Combien égal à 68 grammes ?

   X continu $\Rightarrow p(X = 68) = 0$. Mais, si on mesure au gramme près, on cherche le nombre de bonbons avec un poids entre 67.5 et 68.5 grammes :

$$\begin{aligned} p(67.5 < X < 68.5) &= p(\frac{-0.5}{3} < Z < \frac{0.5}{3}) = p(-0.167 < Z < 0.167) \\ &= 2p(Z < 0.167) - 1 = 2 \cdot 0.566 - 1 \simeq 0.132 \\ n_{=68} &= 600 \cdot 0.132 = \boxed{79} \end{aligned}$$

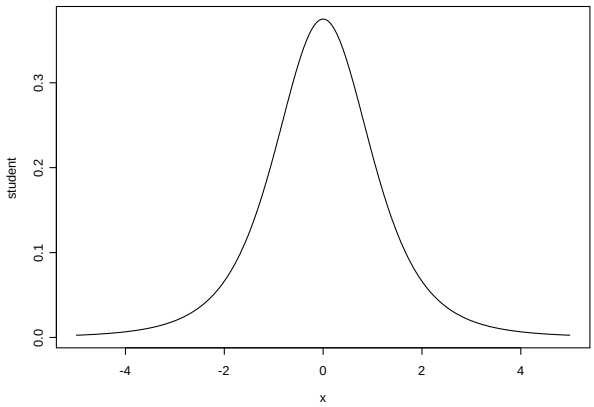
Solution de l'exercice 6.4

Loi de Student (St_n) à n degrés de liberté

$$T \sim St_n \quad \mathrm{E}(T_n) = 0, \mathrm{Var}(T_n) = \frac{n}{n-2}$$

Remarques : $\rightarrow St_n$ symétrique autour de 0, ressemble à la loi $N(0,1)$.
$\rightarrow St_n$ tend vers la loi $N(0,1)$ lorsque n devient grand.
$\rightarrow \mathrm{Var}(T_n)$ est non définie (variance infinie) pour $n \leq 2$.

Forme :



La table de la loi de Student se présente de façon inverse à celle de la table de la loi normale : probabilités en marge, seuils à l'intérieur.
Pour la probabilité : $p(T_n < a) = p$, avec $a > 0 \Leftrightarrow p > 0.5$ on lit dans la table :

→ 1ère colonne (identificateur des lignes) : les degrés de liberté n
→ 1re ligne (identificateur des colonnes : la probabilité $p > 0.5$
→ A l'intérieur de la table : le seuil $a > 0$ pour la probabilité p et n degrés de liberté tel que $p(T_n < a) = p$.

Pour $a < 0$ ($p < 0.5$), on exploite la symétrie et la complémentarité à 1 :

$$p(T_n < a) = p \quad \Leftrightarrow \quad p(T_n < -a) = 1 - p$$

Pour $a < 0$ ($p < 0.5$), la table donne $-a$ (respectivement $1 - p$) d'où il est aisé de déduire a (respectivement p).

i) Pour $n = 15$ degrés de liberté, on a

1. $P(T_{15} < a) = 0.9 \Rightarrow a = 1.341$
2. $P(T_{15} < -a) = 0.9 \Rightarrow -a = 1.341 \Rightarrow a = -1.341$
3. $P(T_{15} < a) = 0.05 \Rightarrow 1 - P(T_{15} < -a) = 0.05$
   $\Rightarrow P(T_{15} < -a) = 1 - 0.05 = 0.95 \Rightarrow -a = 1.753 \Rightarrow a = -1.753$
4. $P(-a < T_{15} < a) = 0.99 \Rightarrow P(T_{15} < a) - P(T_{15} < -a) = 0.99$
   $\Rightarrow P(T_{15} < a) - (1 - P(T_{15} < a)) = 0.99$
   $\Rightarrow 2\,P(T_{15} < a) - 1 = 0.99$
   $\Rightarrow P(T_{15} < a) = \frac{1.99}{2} = 0.995 \Rightarrow a = 2.947$

ii) Pour $n = 10$ degrés de liberté, on a

1. $P(T_{10} < a) = 0.9 \Rightarrow a = 1.372$
2. $P(T_{10} < -a) = 0.9 \Rightarrow -a = 1.372 \Rightarrow a = -1.372$
3. $P(T_{10} < a) = 0.05 \Rightarrow 1 - P(T_{10} < -a) = 0.05$
   $\Rightarrow P(T_{10} < -a) = 1 - 0.05 = 0.95 \Rightarrow -a = 1.812 \Rightarrow a = -1.812$

4. $P(-a < T_{10} < a) = 0.99 \Rightarrow P(T_{10} < a) - P(T_{10} < -a) = 0.99$
   $\Rightarrow P(T_{10} < a) - (1 - P(T_{10} < a)) = 0.99$
   $\Rightarrow 2\,P(T_{10} < a) - 1 = 0.99$
   $\Rightarrow P(T_{10} < a) = \dfrac{1.99}{2} = 0.995 \Rightarrow a = 3.169$

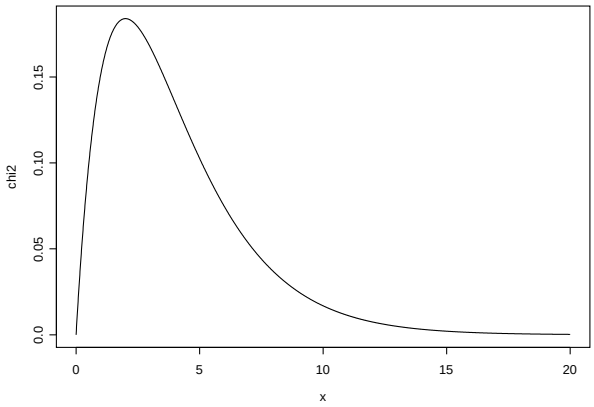
Solution de l'exercice 6.5

Loi du khi-2 (χ^2) à n degrés de liberté

$$Q_n \sim \chi^2_n \quad \mathrm{E}(Q_n) = n, \mathrm{Var}(Q_n) = 2n$$

Remarques : → Q_n est toujours positive
→ Q_n n'est pas symétrique
(on ne peut donc pas exploiter la symétrie)

Forme :



La table du χ^2 se présente comme celle de la loi de Student : probabilités en marge, seuils à l'intérieur.
Pour la probabilité : $p(Q_n < a) = p$, on lit dans la table :

→ 1ère colonne (identificateur des lignes) : les degrés de liberté n
→ 1ère ligne (identificateur des colonnes : la probabilité p
→ A l'intérieur de la table : le seuil a pour la probabilité p et n degrés de liberté tel que $p(Q_n < a) = p$.

Pour $n = 10$ degrés de liberté, on trouve :

1. $p(Q_{10} < a) = 0.95 \Rightarrow \boxed{a = 18.31}$
2. $p(Q_{10} < b) = 0.1 \Rightarrow \boxed{b = 4.87}$
3. $p(b < Q_{10} < a) = p(Q_{10} < a) - p(Q_{10} < b) = 0.95 - 0.1 = 0.85$
   $\boxed{p(4.87 < Q_{10} < 18.31) = 85\%}$

Solution de l'exercice 7.4

X : temps de lecture, avec $X \sim N(\mu, \sigma^2)$
Échantillon de 8 observations ($n = 8$)

330 547 461 380 412 356 502 484

1. Temps moyen de lecture :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{8} x_i = \boxed{434}$$

$\bar{X}$ est un estimateur non-biaisé de $\mathrm{E}(\bar{X}) = \mu$ car les X_i sont ici i.i.d.(μ, σ^2).

Autres estimateurs possibles :

(a) $\widehat{\mu}_1 = \operatorname{med} X$
Données ordonnées : 330 356 380 412 461 484 502 547

$$\begin{aligned} \operatorname{rang}(\operatorname{med} x) &= \frac{n+1}{2} = \frac{8+1}{2} = 4.5 \\ \Rightarrow \operatorname{med} x &= \frac{412+461}{2} = \boxed{436.5} \end{aligned}$$

La médiane est la moyenne de la 4ème et de la 5ème valeur.

(b) $\hat{\mu}_2 = \bar{X}_{.25}$ moyenne tronquée du 25% des valeurs les plus extrêmes (on enlève 330 et 547). La moyenne des 6 valeurs restantes est

$$\hat{\mu}_2 = \bar{x}_{.25} = \frac{1}{6}(356 + 380 + 412 + 461 + 484 + 502) = \boxed{432.5}$$

(c) $\widehat{\mu}_3$= moyenne des valeurs extrêmes $\min_i(X_i)$ et $\max_i(X_i)$

$$\hat{\mu}_3 = \frac{330 + 547}{2} = \boxed{438,5}$$

La moyenne est l'estimateur le plus efficace si la normalité est effectivement vérifiée. La médiane et la moyenne tronquée sont robustes. La moyenne tronquée est un bon compromis entre efficacité et robustesse. La moyenne des extrêmes $\hat{\mu}_3$ est à éviter, car trop sensible aux valeurs extrêmes.

2. Intervalle de confiance pour μ avec σ^2 connu ($\sigma^2 = 5850$), et un degré de confiance $1 - \alpha = 0.9$. L'intervalle est donné par

$$\mu = \bar{x} \pm \sigma_{\bar{X}} z_{1-\frac{\alpha}{2}}$$

où $z_{1-\alpha/2}$ est le seuil critique de la loi normale $N(0, 1)$.

Dans notre cas : $\bar{x} = 434$, $\sigma_{\bar{X}} = \sqrt{\sigma^2/n} = \sqrt{5850/8} = 27.04$ et $z_{1-\frac{\alpha}{2}} = z_{0.95} = 1.645$. D'où l'intervalle estimé :

$$\mu = 434 \pm \underbrace{1.645 \cdot 27.04}_{44.48} \quad \Leftrightarrow \quad \big[389.52\ ;\ 478.48\big]$$

A priori, avec notre façon de construire l'intervalle (avec l'estimateur par intervalle utilisé), on a une probabilité de 0.9 d'obtenir un intervalle qui contient μ.

3. Estimation de la variance σ^2.

On détermine successivement ns^2 la somme des carrés des écarts à la moyenne, la variance d'échantillon s^2 et l'estimation non biaisée $\hat{\sigma}^2$:

$$\begin{aligned} ns^2 &= \sum_{i=1}^{8}(x_i - \bar{x})^2 \\ &= (330-434)^2 + \cdots + (484-434)^2 = 40'922 \\ s^2 &= \frac{40'922}{8} = 5'115.25 \\ \hat{\sigma}^2 &= \frac{ns^2}{n-1} = \frac{40'922}{7} = \boxed{5'846} \end{aligned}$$

Notons que la valeur $\sigma^2 = 5'850$ supposée au point précédent est proche de cette estimation.

4. Intervalle pour σ^2 avec un degré de confiance $1-\alpha = 0.95$. L'intervalle est donnée par

$$\sigma^2 \in \left[\frac{ns^2}{q_{1-\frac{\alpha}{2}}^{(n-1)}} \; ; \; \frac{ns^2}{q_{\frac{\alpha}{2}}^{(n-1)}} \right]$$

où $q_{\frac{\alpha}{2}}^{(n-1)}$ et $q_{1-\frac{\alpha}{2}}^{(n-1)}$ sont les seuils critiques de la loi du khi-2 à $n-1$ degrés de liberté.
Dans notre cas, $n = 8$ et l'on a $n-1 = 7$ degrés de liberté. Le degré de confiance fixé étant $1-\alpha = 0.95$, on a $\alpha/2 = 0.025$ et $1-\alpha/2 = 0.975$. Dans la table du χ^2, on trouve les seuils critiques

$$q_{0.975}^{(7)} = 16.01 \quad \text{et} \quad q_{0.025}^{(7)} = 1.69$$

d'où l'intervalle estimé pour σ^2

$$\left[\frac{40'922}{16.01} \; ; \; \frac{40'922}{1.69} \right] \quad = \quad \big[2'556.03 \; ; \; 24'214.20\big]$$

Cet intervalle correspond à l'intervalle $\big[\sqrt{2556.03}; \sqrt{24214.2}\big] = \big[50.6 \; ; \; 155.6\big]$ pour l'écart type σ.

Remarque : L'intervalle est large, ce qui n'est pas surprenant vu la faible taille de l'échantillon. L'hypothèse faite précédemment d'une variance de 5850 (écart type de 76.5) est dans l'intervalle. L'échantillon ne permet pas de rejeter cette hypothèse.

5. Intervalle de confiance pour μ avec σ^2 inconnu et degré de confiance $= 1-\alpha = 0.9$
L'intervalle est donné par

$$\mu = \bar{x} \pm \hat{\sigma}_{\bar{X}} t_{1-\frac{\alpha}{2}}^{(n-1)}$$

où $t_{1-\alpha/2}^{(n-1)}$ est le seuil critique de la loi de Student à $n-1$ degrés de liberté.
Dans notre cas : $\bar{x} = 434$, $\hat{\sigma}_{\bar{X}} = \sqrt{\hat{\sigma}^2/n} = \sqrt{5846/8} = 27.03$ $(= \sqrt{s^2/(n-1)} = \sqrt{5115.25/7})$ et $t_{1-\frac{\alpha}{2}}^{(7)} = t_{0.95}^{(7)} = 1.895$. D'où l'intervalle estimé :

$$\mu = 434 \pm \underbrace{1.895 \cdot 27.03}_{51.22} \quad \Leftrightarrow \quad \big[382.78 \; ; \; 485.22\big]$$

L'intervalle est ici plus grand que celui trouvé au point 2) (lorsque σ^2 est connu). On a ici $\hat{\sigma}_{\bar{X}} \simeq \sigma_{\bar{X}}$. Comme σ^2 est inconnu, on a plus d'incertitude ce qui se traduit par $z_{1-\frac{\alpha}{2}} \leq t_{1-\frac{\alpha}{2}}^{(n-1)}$.
Lorsque n est grand (n>120), on a $t_{1-\frac{\alpha}{2}}^{(n-1)} \simeq z_{1-\frac{\alpha}{2}}$.

Solution de l'exercice 7.5

Données :

$n = 100$: taille de l'échantillon (nombre de passagers de première classe interrogés)

$\bar{x} = 76400$ km : moyenne de l'échantillon, distance moyenne annuelle parcourue par les voyageurs interrogés.

$s = 5250$ km : écart type de l'échantillon

On admet que $\sigma^2 = s^2$, c'est-à-dire que la variance des distances parcourues par l'ensemble des passagers de première (la population) est égale à la variance observée au sein de l'échantillon (les 100 personnes interrogées).

Intervalle de confiance pour μ avec σ^2 connu

Comme n est grand, $\bar{X} \sim N\big(\mu,\ \dfrac{\sigma^2}{n}\big)$ en vertu du théorème central limite.
En terme de probabilité on a :

$$p\big(\bar{X} - \sigma_{\bar{X}} \cdot z_{1-\frac{\alpha}{2}} < \mu < \bar{X} + \sigma_{\bar{X}} \cdot z_{1-\frac{\alpha}{2}}\big) = 1 - \alpha$$

avec $1 - \alpha = 0.95$.
L'intervalle de confiance, c'est-à-dire la réalisation de l'intervalle aléatoire ci-dessus pour l'échantillon observé (données), est :

$$\mu \quad = \quad \bar{x} \pm \sigma_{\bar{X}}\, z_{1-\frac{\alpha}{2}}$$

Dans notre cas, on a : $\boxed{\bar{x} = 76400}$ (donnée) et l'écart type $\sigma_{\bar{X}}$ de l'estimateur $\bar{X}$ s'obtient comme suit :

$$\sigma_{\bar{X}} \quad = \quad \sqrt{\frac{\sigma^2}{n}} \ = \ \frac{\sigma}{\sqrt{n}} \ = \ \frac{5250}{\sqrt{100}} = \boxed{525}$$

Pour le degré de confiance $1 - \alpha = 0.95$, on a $\alpha = 0.05$ et $1 - \frac{\alpha}{2} = 0.975$. Dans la table de la loi normale on trouve le seuil critique $z_{1-\frac{\alpha}{2}} = \boxed{z_{0.975} = 1.96}$ (tel que $p(Z < 1.96) = 0.975$).

Ainsi, l'intervalle de confiance à 95% est :

$$\begin{aligned} \mu &= 76400 \ \pm \ 525 \cdot 1.96 \\ &= 76400 \ \pm \ 1029 \\ \mu &\in \big[\,75371\ ;\ 77429\,\big] \end{aligned}$$

Solution de l'exercice 7.6

X : diamètre d'une pizza choisie au hasard supposé distribué normalement et tirages i.i.d. :

$$X \sim N(\mu, \sigma^2) \ \Rightarrow \ \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Données :

$n = 15$ pizzas, taille de l'échantillon

$\bar{x} = 30$ cm, diamètre moyen des pizzas observées (moyennne de l'échantillon) $\sigma^2 = 9$: variance supposée de la population

1. Intervalle de confiance pour μ avec σ^2 connu, à 95%

$\boxed{\bar{x} = 30}$ (donnée)

$\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n} = \frac{9}{15} = 0.6 \quad \Rightarrow \boxed{\sigma_{\bar{X}} = 0.775}$

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ et $1 - \frac{\alpha}{2} = 0.975$. On utilise donc le seuil critique $z_{1-\frac{\alpha}{2}} = \boxed{z_{0.975} = 1.96}$

L'intervalle de confiance pour la taille moyenne des pizza μ est alors :

$$\begin{aligned} \mu &= 30 \pm 0.775 \cdot 1.96 \\ &= 30 \pm 1.52 \\ \mu &\in \big[28.48\,;\,31.52\big] \end{aligned}$$

Données supplémentaires :

On observe la même moyenne avec 11 données supplémentaires :
$\Rightarrow \quad \bar{x} = 30$ et $n_2 = 15 + 11 = 26$.

2.a) Intervalle de confiance pour μ avec σ^2 connu, à 95%

$\boxed{\bar{x} = 30}$ (donnée)

$\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n_2} = \frac{9}{26} = 0.346 \quad \Rightarrow \boxed{\sigma_{\bar{X}} = 0.588}$

$\boxed{z_{0.975} = 1.96}$

L'intervalle de confiance pour la taille moyenne des pizza μ est alors :

$$\begin{aligned} \mu &= 30 \pm 0.588 \cdot 1.96 \\ &= 30 \pm 1.15 \\ \mu &\in \big[28.85\,;\,31.15\big] \end{aligned}$$

Remarque : La longueur de l'intervalle diminue lorsqu'on augmente n. On améliore la précision en utilisant plus d'information.

2.b) Intervalle de confiance pour μ avec σ^2 connu, à 90%

Pour un degré de confiance de $1 - \alpha = 90\%$, on a $1 - \alpha/2 = 95\%$ et l'on utilise alors le seuil critique

$\boxed{z_{1-\frac{\alpha}{2}} = z_{0.95} = 1.645}$ au lieu du 1.96 utilisé précédemment.

L'intervalle pour μ devient alors :

$$\begin{aligned} \mu &= 30 \pm 0.588 \cdot 1.645 \\ &= 30 \pm 0.97 \\ \mu &\in \big[29.03\,;\,30.97\big] \end{aligned}$$

Remarque : La longueur de l'intervalle diminue lorsqu'on réduit la confiance. On gagne en précision, mais au prix d'une moindre confiance.

Solution de l'exercice 7.7

On veut interroger n personnes par tirages au hasard sans remises parmi la population genevoise afin d'estimer la proportion p de personnes favorables à l'ouverture des commerces jusqu'à 21 heures.
X_i : prend la valeur 1 si le i-ème individu choisi est favorable et 0 sinon.

1. Les X_i, $i = 1, \ldots, n$ peuvent être considérés comme indépendants si n est petit par rapport à la population genevoise (env. 300000 adultes). Dans ce cas les X_i sont i.i.d selon le tableau de probabilités

$$\begin{array}{c|cc} x & 0 & 1 \\ \hline p(X_i = x) & 1-p & p \end{array}$$

   On a $\mathrm{E}(X_i) = p$ et $\mathrm{Var}(X_i) = p(1-p)$ (pour la démonstration voir cours).

Données :

$n = 500$ (faible taux de sondage $\frac{500}{300000} = 0.17\%$ justifie l'indépendance des X_i).

300 avis favorables ($\sum_{i=1}^{500} x_i = 300$).

2. Estimateur absolument correct $\hat{P} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. L'estimation correspondante est

$$\boxed{\hat{p} = \frac{300}{500} = 0.6}$$

3. Hypothèse : $\hat{P} \sim N(\mu_{\hat{P}}, \sigma^2_{\hat{P}})$
   Comme $\hat{P} = \bar{X}$, cette hypothèse est justifiée par le théorème central limite.

4. On a $\mathrm{Var}(\hat{P}) = \mathrm{Var}(X_i)/n = p(1-p)/n$. On admet que $\mathrm{Var}(X_i)$ est égal à la variance observée dans l'échantillon $\hat{p}(1-\hat{p}) = 0.6 \cdot 0.4 = 0.24$. L'écart type de $\hat{P}$ est alors

$$\boxed{\sigma_{\hat{P}} = \sqrt{\frac{\mathrm{Var}(X_i)}{n}} = \sqrt{\frac{0.24}{500}} = 0.0219}$$

   Pour un degré de confiance de 90%, on utilise le seuil critique $\boxed{z_{0.95} = 1.645}$. L'intervalle de confiance pour p est alors

$$\begin{aligned} p &= 0.6 \pm 1.645 \cdot 0.0219 \\ &= 0.6 \pm 0.036 \end{aligned}$$

   Remarque : pour le degré de confiance retenu de 90% la marge d'erreur est ici de 3.6%, c'est-à-dire que l'on estime que la proportion de personnes favorables est de 60% plus ou moins 3.6%.

Solution de l'exercice 8.4

Votation du 7 mars 1993 : 73,73% de oui à la réouverture des casinos.
Échantillon des 10 cantons sans frontière avec l'étranger :

Lucerne	76	Glaris	74
Uri	76	Zoug	74
Schwytz	75	Fribourg	75
Obwald	75	Appenzell R.-Ext.	77
Nidwald	78	Appenzell R.-Int.	77

$\sum_{i=1}^{10} x_i^2 = 57'321$

X pourcentage de oui par canton. On admet $X \sim N(\mu, \sigma^2)$.
Le taux de oui de ces cantons sans frontière est-il significativement plus grand que la moyenne suisse ?
$\Rightarrow$ tester

$H_0 : \mu = \mu_0 = 73.73$ contre
$H_1 : \mu = \mu_1 > 73.73$

Calcul préliminaire pour les 10 observations :

$\bar{x} = (76 + 76 + \cdots + 77)/10 = \boxed{75.7}$

$s_x^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = \frac{1}{10}\,57321 - (75.7)^2 = \boxed{1.61}$

1. Cas où σ^2 est inconnu.

   (a) Avec σ^2 inconnu, on ne peut pas utiliser directement la loi normale. Comme on admet la normalité des X_i, on peut utiliser la statistique $T_0 = (\bar{X} - \mu_0)/\hat{\sigma}_{\bar{X}}$ qui suit la loi de Student à $n - 1 = 9$ degrés de liberté sous H_0 et ne dépend pas de σ^2.

   (b) $\alpha = 0.05$ (donné.)

   (c) Région critique.

   - $H1 : \mu_1 > \mu 0 \Rightarrow$ test unilatéral à droite : $\boxed{R = \{t_0 \mid t_0 > r\}}$.
   - Le seuil est $r = t_{1-\alpha}^{(n-1)} = t_{.95}^{(9)} = \boxed{1.833}$ que l'on trouve dans la table de Student à l'intersection de la ligne 9 et de la colonne $p = 0.95$.

   (d) Valeur calculée de la statistique : on calcule tout d'abord l'erreur standard de $\bar{X}$

   $\hat{\sigma}_{\bar{X}} = \sqrt{\dfrac{s^2}{n-1}} = \sqrt{\dfrac{1.61}{9}} = 0.423$, ce qui donne pour la statistique t :

   $$t_{\text{calc}} = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{X}}} = \frac{75.7 - 73.73}{0.423} = \boxed{4.66}$$

   (e) Comme $t_{\text{calc}} = 4.66 > r = 1.833$, on a $t_{\text{calc}} \in R \Rightarrow$ on rejette H_0.
   En d'autres termes, la moyenne est significativement supérieure à celle de l'ensemble de la Suisse.

2. Cas où l'on connaît $\sigma^2 = 1.8$.

   (a) Comme on admet la normalité des X_i et que σ^2 est connu, on peut utiliser la statistique $Z_0 = (\bar{X} - \mu_0)/\sigma_{\bar{X}}$ qui suit la loi $N(0, 1)$ sous H_0.

   (b) $\alpha = 0.05$ (donné.)

   (c) Région critique.

   - $H1 : \mu_1 > \mu 0 \Rightarrow$ test unilatéral à droite : $\boxed{R = \{z_0 \mid z_0 > r\}}$.
   - Le seuil est $r = z_{1-\alpha} = z_{.95} = \boxed{1.645}$

   (d) Valeur calculée de la statistique : on calcule tout d'abord l'écart type de $\bar{X}$

   $\sigma_{\bar{X}} = \sqrt{\dfrac{\sigma^2}{n}} = \sqrt{\dfrac{1.8}{10}} = 0.424$, ce qui donne pour la statistique z :

   $$z_{\text{calc}} = \frac{\bar{x} - \mu_0}{\sigma_{\bar{X}}} = \frac{75.7 - 73.73}{0.424} = \boxed{4.64}$$

   (e) Comme $z_{\text{calc}} = 4.64 > r = 1.645$, on a $z_{\text{calc}} \in R \Rightarrow$ on rejette H_0.
   Comme précédemment, on conclut que la moyenne est significativement supérieure à la moyenne suisse.

3. Test de $H_0 : \sigma^2 = \sigma_0^2 = 1.8$

   (a) Pour le test de la variance, on utilise la statistique $Q_0 = \dfrac{nS^2}{\sigma_0^2}$ qui suit une loi χ^2_{n-1} sous H_0.

(b) $\alpha = 0.05$ (donné.)

(c) Région critique.

- Comme la variance estimée $\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{10}{9} 1.61 \simeq 1.789$ est inférieure à la valeur testée, un test unilatéral à gauche est indiqué : $\boxed{R = \{q_0 \mid q_0 < r\}}$.
  (On pourrait éventuellement faire un test bilatéral, voir ci-dessous.)
- Dans la table du khi-2, ligne 9 colonne $p = 0.05$, on trouve le seuil critique $r = \chi^2_{.05,(9)} = \boxed{3.33}$

(d) Valeur calculée de la statistique :
$q_{\text{calc}} = \dfrac{ns^2}{\sigma_0^2} = \dfrac{10 \cdot 1.61}{1.8} = \boxed{8.94}$.

(e) Comme $q_{\text{calc}} = 8.94 > \chi^2_{.05,(9)} = 3.33$, on n'est pas dans la région critique $\Rightarrow$ on ne rejette pas H_0. En d'autres termes l'échantillon ne remet pas en cause l'hypothèse faite sur la valeur de σ^2.

Test bilatéral : La région critique devient $R = \{q_0 \mid q_0 < r_1 \text{ ou } q_0 > r_2\}$, avec les seuils critiques : $r_1 = \chi^2_{\frac{\alpha}{2},(n-1)} = \chi^2_{.025,(9)} = 2.7$ et $r_2 = \chi^2_{1-\frac{\alpha}{2},(n-1)} = \chi^2_{.975,(9)} = 19.02$.
Comme $q_{\text{calc}} = 8.94 \in [2.7; 19.02]$, on n'est pas dans la région de rejet et l'on accepte H_0.

Variantes :
Il existe plusieurs variantes équivalentes des tests précédents.

- Exprimer les seuils en terme de moyenne pour le test de la moyenne et en terme de variance pour le test de la variance :

  1. $r_{\bar{x}} = \mu_0 + t^{(9)}_{1-\alpha} \hat{\sigma}_{\bar{X}} = 73.73 + 1.833 \cdot 0.423 = \boxed{74.51}$
     Comme $\bar{x} = 75.7 > r_{\bar{x}} = 74.51 \Rightarrow$ rejet de H_0.
  2. $r_{\bar{x}} = \mu_0 + z_{1-\alpha} \sigma_{\bar{X}} = 73.73 + 1.645 \cdot 0.424 = \boxed{74.43}$
     Comme $\bar{x} = 75.7 > r_{\bar{x}} = 74.43 \Rightarrow$ rejet de H_0.
  3. $r_{s^2} = \dfrac{\chi^2_{0.05,(9)}\, \sigma_0^2}{n} = \dfrac{3.33 \cdot 1.8}{10} = \boxed{.599}$
     La variance d'échantillon $s^2 = 1.61$ n'étant pas inférieure à ce seuil, on accepte H_0.

- Tests avec les p-valeurs (degrés de signification) :
  rejet de H_0 lorsqu'elles sont inférieures à α.

  Pour les trois tests précédents, les degrés de signification sont

  1. $p(T_0 > 4.66 | H_0) = .0006 < \alpha = .05 \Rightarrow$ rejet de H_0.
  2. $p(Z_0 > 4.64 | H_0) = .000002 < \alpha = .05 \Rightarrow$ rejet de H_0.
  3. $p(Q_0 < 8.94 | H_0) = .56 > \alpha = .05 \Rightarrow$ non rejet de H_0.

  Le sens des inégalités est dicté par la forme de la région critique.

  Les probabilités ont été obtenues respectivement avec les formules Excel (anglais) :
  =TDIST$(t, d, 1)$, =1-NORMSDIST(z) et =1-CHIDIST(q, d), où t, z et q sont les valeurs calculées des statistiques et d les degrés de liberté, soit 9 dans cet exercice.