

# Exercices

## Exercices du Chapitre 1

1.1 Le nombre d'étamines de 100 fleurs, triés dans l'ordre croissant, sont donnés dans le tableau suivant.

31	138	203	259	304	422	518	646	865	1483
73	139	205	262	338	423	520	648	888	1567
83	144	215	262	339	425	541	666	1095	1591
89	144	216	264	344	437	542	702	1195	1627
97	158	218	271	344	439	569	717	1205	1827
103	173	222	273	351	453	575	732	1362	2110
103	182	224	275	364	461	581	740	1406	2566
113	186	226	276	366	470	603	745	1412	2881
115	187	243	279	403	485	609	803	1469	3277
123	203	253	293	410	499	646	824	1470	5481

Les données suivantes sont les logarithmes naturels de ces nombres.

3.434	4.927	5.313	5.557	5.717	6.045	6.250	6.471	6.763	7.302
4.290	4.934	5.323	5.568	5.823	6.047	6.254	6.474	6.789	7.357
4.419	4.970	5.371	5.568	5.826	6.052	6.293	6.501	6.999	7.372
4.489	4.970	5.375	5.576	5.841	6.080	6.295	6.554	7.086	7.394
4.575	5.063	5.384	5.602	5.841	6.084	6.344	6.575	7.094	7.510
4.635	5.153	5.403	5.609	5.861	6.116	6.354	6.596	7.217	7.654
4.635	5.204	5.412	5.617	5.897	6.133	6.365	6.607	7.249	7.850
4.727	5.226	5.421	5.620	5.903	6.153	6.402	6.613	7.253	7.966
4.745	5.231	5.493	5.631	5.999	6.184	6.412	6.688	7.292	8.095
4.812	5.313	5.533	5.680	6.016	6.213	6.471	6.714	7.293	8.609

- En partageant les données en 11 intervalles (qui s'étendent de 0 à 5500 pour la première série et de 3.3 à 8.8 pour la seconde), tracer les histogrammes de ces deux séries de données.
- Construire les graphes des fonctions de distribution cumulatives.
- Commenter les résultats.

1.2 Un maraîcher est très fier de la quantité de fruits fournis par sa nouvelle sorte de cerisier. Pour tenter de comprendre quelle est la quantité "normale" de fruits pour un arbre de ce type, il mesure cette quantité sur chacun de ses 78 cerisiers et trouve les résultats suivants:

Quantité de cerises (Kg)	Nombre d'arbres	Quantité de cerises (Kg)	Nombre d'arbres
100	3	148	10
108	3	156	9
116	7	164	8
124	7	172	6
132	9	180	4
140	10	188	2

Faites deux histogrammes de ces valeurs; le premier avec 12 classes (de 96 à 192), le second avec 6 classes. Que remarquez-vous et quelle conclusion pouvez-vous en tirer ?

1.3 On cherche à savoir si le déficit alimentaire protéique est associé à la myopie. Les données du tableau représentent des mesures de réfraction sur l'oeil droit de 20 singes nourris pendant 32 mois (en moyenne) avec une diète à faible contenu protéique et de 17 singes nourris pendant 28 mois (en moyenne) avec une diète à haut contenu protéique.

Niveau protéique faible	Niveau protéique élevé
1.27	-6.00
-4.98	0.25
-0.50	1.25
1.25	-2.00
-0.25	3.14
0.75	2.00
-2.75	0.75
0.75	1.75
1.00	0.00
9.00	0.75
2.25	0.75
0.53	0.25
1.25	1.25
-1.50	1.25
-5.00	1.00
0.75	0.50
1.50	-2.25
0.50	
1.75	
1.50	

- (a) Représenter graphiquement ces deux séries de mesures par des points sur deux axes parallèles.
- (b) Est-ce que selon vous la représentation graphique soutient l'hypothèse que le déficit alimentaire protéique est associé à la myopie ?

## Exercices du Chapitre 2

2.1 Considérez les données de l'exercice 1.3.

- Calculez les moyennes, les médianes, les écarts types et les écarts interquartiles des deux séries de mesures.
- Est-ce que selon vous ces calculs soutiennent l'hypothèse que le déficit alimentaire protéique est associé à la myopie ?
- Une valeur aberrante (9.00) a été détectée dans la première série; après vérification on a constaté qu'il s'agissait d'une erreur. En éliminant cette valeur, recalculez la moyenne, la médiane, l'écart type et l'écart interquartile de l'échantillon modifié. Comparez les nouveaux résultats avec les premiers.

2.2 Considérez les données (nombres d'étamines de 100 fleurs) de l'exercice 1.1 ainsi que les données transformées en prenant leur logarithme naturel. Éliminez une valeur quelconque dans les deux ensembles (par exemple, la valeur la plus petite).

- Calculez les deux médianes et les deux moyennes.
- Comment se manifestent les similarités et les différences entre moyenne et médiane dans les calculs que vous venez d'effectuer en (a) ?
- Y a-t-il une relation entre la médiane des données non transformées et celle des données transformées ?
- Y a-t-il une relation entre le premier quartile des données non transformées et le premier quartile des données transformées ? Si oui, pouvez-vous étendre la propriété que vous venez de découvrir à d'autres quantiles ?
- Pouvez-vous étendre cette propriété des quantiles à d'autres transformations, par exemple, le carré et le sinus ?
- Est-ce que la moyenne a la même propriété ?
- Répétez (a)–(d) avec toutes les données.

2.3 Construire les boxplots des deux ensembles de valeurs de l'exercice 1.1. Quels changements remarque-t-on après la transformation logarithmique ?

2.4 Soient  $X$  et  $Y$  deux variables observées sur les mêmes unités,  $a$ ,  $b$  et  $c$  des nombres fixes (constantes). Soit  $m(X)$  la moyenne de  $x_1, \dots, x_n$  et  $m(Y)$  la moyenne de  $y_1, \dots, y_n$ . Démontrer les propriétés suivantes de la moyenne.

- Si  $x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0$  alors  $m(X) \geq 0$ .
- $m(aX) = a m(X)$ .
- $m(X + a) = m(X) + a$ .
- $m(X + Y) = m(X) + m(Y)$ . Donc  $m(aX + bY + c) = a m(X) + b m(Y) + c$ .
- En général  $m(XY) \neq m(X)m(Y)$ .

2.5 Soient  $X, Y$  deux variables observées sur les mêmes unités,  $a, b$  et  $c$  des constantes. Soit  $s^2(X)$  la variance de  $x_1, \dots, x_n$  et  $s^2(Y)$  la variance de  $y_1, \dots, y_n$ . Démontrer les propriétés suivantes.

- $s^2(c) = 0$ .
- $s^2(aX + b) = a^2 s^2(X)$ .
- $s(aX + b) = a s(X)$ .
- En général  $s^2(X + Y) \neq s^2(X) + s^2(Y)$ .
- La somme des écarts  $x_i - m(X)$  est toujours nulle.
- $s^2(X) = m(X^2) - m(X)^2$ .

## Exercices du Chapitre 3

3.1 On mesure la température du corps d'une vache tous les matins sur 30 jours consécutifs à l'aide d'un thermomètre implanté à l'intérieur de la vache. Cet appareil envoie des pulsations radio à un récepteur situé à proximité. Plus la température est élevée, plus les pulsations sont rapides. On enregistre chaque matin le nombre de pulsations dans un intervalle de 5 minutes. La température de la vache permet de prévoir les périodes de fertilité, qui sont généralement associées à de pics de température.

Jour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pulsations	60	70	54	56	70	66	53	95	70	69	56	70	70	60	60
Jour	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Pulsations	60	50	50	48	59	50	60	70	54	46	57	57	51	51	59

- Représenter graphiquement les pulsations en fonction du temps (jours). Que peut-on dire à partir de ce nuage de points ?
- Procéder à un "lissage" en prenant la médiane de 4 pulsations consécutives (quand c'est possible; 1 ou 2 sinon). Associer un temps (le milieu des 4 jours utilisés) à chacune des ces "pulsations lissées" et représenter les pulsations lissées en fonction du temps.
- Procéder à un nouveau lissage des pulsations lissées obtenues en (b), en prenant cette fois la médiane de deux données consécutives. Représenter ces pulsations lissées deux fois en fonction du temps.
- Que remarquez vous en lissant les données une e deux fois ?

3.2 Dans le cadre d'une enquête visant à comparer, selon certains critères, différents aliments vendus dans les fast-foods, nous avons retenu les informations se trouvant dans le tableau ci-dessous.

Aliment	Poids (g)	Prix (Fr)
Sandwitch végétarien	150	3.90
Sandwitch parisien au jambon	92	3.40
Big Mac	193	5.70
Hamburger	90	2.90
Hot Dog	135	3.80
Fallafel	241	6.00
Quiche Lorraine	169	3.30
Pizza tomates et jambon	165	3.30

Y-a-t-il une relation entre les variables Poids et Prix ? Pour répondre à cette question, faire un graphique, calculer le coefficient de corrélation des deux variables et l'équation de la droite de régression.

3.3 Afin de convaincre des clients potentiels, une société de marketing met en évidence la relation entre le chiffre d'affaires ( $Y$ ) d'une entreprise et son budget publicité ( $X$ ). Le tableau ci-dessous présente ces données (en millions de Fr) pour 11 entreprises issues du même domaine d'activité.

Publicité	Chiffre d'affaires
0.0	2
0.2	3
0.6	7
3.0	15
3.5	19
4.0	23
5.2	32
4.9	37
6.3	54
7.1	58
6.9	61

- (a) Représenter le nuage de points  $(X, Y)$ .
- (b) Calculer
- les variances  $s^2(X)$  et  $s^2(Y)$ ,
  - le coefficient de covariance  $v(X, Y)$ ,
  - le coefficient de corrélation  $r(X, Y)$ ,
  - la droite de régression  $Y = \hat{a} + \hat{b}X$  (et la représenter graphiquement),
  - le vecteur des réponses calculées  $\hat{Y}$ ,
  - le vecteur  $E$  des résidus (et représenter graphiquement les résidus),
  - l'écart type de l'erreur,
  - le coefficient de détermination  $R^2$ .
- (c) Vérifier numériquement que la somme des résidus est nulle.
- (d) Vérifier numériquement que  $s^2(Y) = s^2(\hat{Y}) + s^2(E) = \hat{b}^2 s^2(X) + s^2(E)$ .
- (e) Que devient la droite de régression (et les mesures associées) si l'on ajoute une entreprise ayant un budget pub de 9 millions et un chiffre d'affaires de 15 millions ?

## Exercices du Chapitre 4

4.1 Une étude sur 1000 personnes assurées auprès d'une certaine caisse maladie révèle les résultats suivants:

- 525 sont âgées de plus de 30 ans;
- 470 sont des personnes mariées;
- 312 sont des femmes;
- 147 sont des personnes mariées et âgées de plus de 30 ans;
- 86 sont des femmes mariées;
- 42 sont des femmes âgées de plus de 30 ans;
- 25 sont des femmes mariées et âgées de plus de 30 ans;

Montrer que ces chiffres ne peuvent pas être exacts.

4.2 Une famille planifie la naissance de 3 enfants. On considère les événements

$A$ : le nombre de filles sera 2 ou 3,

$B$ : il y aura des enfants des deux sexes.

- (a) Calculer les probabilités  $P(A)$  et  $P(B)$ ;
- (b) Est-ce que  $A$  et  $B$  sont indépendants ?

4.3 On choisit une personne au hasard dans une population. On sait que la probabilité que cette personne soit un fumeur est 0.47, qu'elle soit un alcoolique est 0.28 et qu'elle soit affligée de ces deux vices est 0.18. Déterminer la probabilité qu'une personne choisie ne soit ni un fumeur ni un alcoolique. Selon vous, ces deux caractères sont-ils indépendants ou non ? D'après les données, a-t-on cette indépendance ?

4.4 On sait qu'un chasseur A a 2 chances sur 3 de tuer un lièvre lorsqu'il tire, un autre chasseur B a 3 chances sur 4. Ils tirent ensemble sur le même lièvre. Quelle est la probabilité que l'animal soit tué ?

4.5 Quelle est la probabilité que, parmi une assemblée de 10 personnes, il y en ait deux au moins qui soient nées le même jour ?

4.6 On observe pendant 1000 jours les cours des actions de deux sociétés A et B. Les évolutions journalières sont résumées dans le tableau suivant, où le signe + indique que le cours a augmenté, le signe – indique que le cours a diminué et le zéro (0) indique que le cours est resté stable.

		Action B		
		+	0	–
Action A	+	312	34	54
	0	60	110	80
	–	28	56	266

Par exemple, le nombre 312 indique que les deux cours ont augmenté simultanément pendant 312 jours. Quelle est la probabilité que le cours de B augmente sachant que celui de A reste stable ?

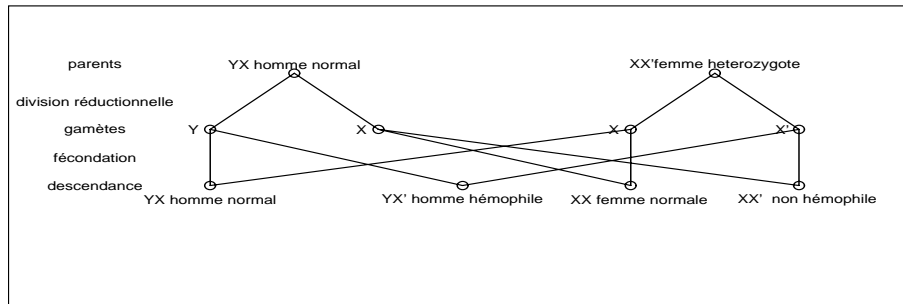
4.7 Une compagnie pétrolière exécute un forage dans la Mer du Nord et un autre en Méditerranée. La probabilité de trouver du pétrole en Mer du Nord est de 0.8 tandis qu'en Méditerranée elle est de 0.6. Quelle est la probabilité qu'un seul des deux forages conduise à la découverte de pétrole ?

4.8 Dans une ville, le 60% des chats sont gris et le 40% sont noirs. Le 60% des chats noirs et le 40% des chats gris vivent dans la rue. Si on attrape un chat au hasard dans la rue, quelle est la probabilité qu'il soit noir ?

4.9 Un baromètre est utilisé pour prévoir le temps. Il arrive toutefois que sa prévision soit fausse. On a constaté que dans 20 cas sur 200 jours de pluie il prévoyait du beau temps, tandis que dans 20 cas sur 100 jours de beau temps il prévoyait de la pluie. Le prospectus d'une localité touristique indique que les jours de pluie représentent le 10% du nombre total de jours en une année. Si le baromètre prévoit de la pluie, quelle est la probabilité qu'il pleuve ?

## Exercices du Chapitre 5

5.1 La transmission du gène hémophile ( $X'$ ) se fait selon le schéma suivant:



Supposons que la probabilité de naissance de chaque sexe est de 0.5. Supposons en outre qu'un homme normal épouse une femme hétérozygote et que ce couple souhaite avoir quatre enfants. Quelle est la probabilité pour

- que les quatre soient normaux génotypiquement (i.e. ni hémophiles ni hétérozygotes)?
- que deux au moins soient normaux génotypiquement ?
- qu'ils aient un fils hémophile au moins ?
- qu'ils aient deux fils hémophiles et deux filles hétérozygotes ?
- qu'aucun de leurs enfants ne soit hémophile ?

5.2 Supposons que l'on porte le gène "yeux bleus" avec une probabilité de 0.5. Si la reine est porteuse, chaque prince aura aussi une chance sur deux d'avoir les yeux bleus. De plus, seule la reine peut transmettre le gène.

- Sachant que la reine a eu trois fils aux yeux verts ("non-bleus") quelle est la probabilité qu'elle soit porteuse du gène ?
- S'il naît un quatrième prince et en supposant les hypothèses de la partie (a) (c'est-à-dire, la reine a déjà eu 3 fils aux yeux verts), avec quelle probabilité aura-t-il les yeux bleus ?

5.3 Une société spécialisée dans l'achat et la vente de matières premières pense qu'il y a 60% de chance que le prix du café augmente. Avant de conclure un important contrat d'achat, elle regarde les prévisions d'un journal spécialisé. La qualité de ses prévisions peut être jugée en regardant les résultats dans le passé. Dans 80% des cas le journal a correctement prévu l'évolution lorsqu'il y a eu une augmentation des prix, tandis qu'une diminution des prix n'a été correctement prévue que dans 55% des cas. Il n'y a jamais eu de cas où les prix n'aient pas varié. Supposons maintenant que le journal prévoit une augmentation.

- Comment la société peut tenir compte de cette information dans son estimation initiale de la probabilité que le prix du café augmente ?

La direction de cette société n'est pas encore convaincue qu'il soit opportun de conclure le contrat. Elle s'adresse donc à un brillant expert qui, dans le passé, a prévu correctement les augmentations dans 99% des cas et les diminutions dans 88% des cas. Supposons que l'expert prévoit une augmentation.

- Est-ce que la société peut améliorer son estimation de la probabilité que le prix du café augmente en tenant compte – outre la prévision du journal spécialisé – de cette nouvelle information fournie par l'expert ? Comment ?

## Exercices du Chapitre 6

6.1 Dans une classe de 10 élèves comprenant 4 garçons et 6 filles, on choisit 3 élèves au hasard. Calculer la probabilité que ce groupe contienne 0, 1, 2, 3 garçons.

6.2 Un épicier reçoit un lot de pommes dont 25% sont avariées. Il charge un employé de préparer des emballages de cinq pommes chacun. Celui-ci, négligent, ne se donne pas la peine de jeter les fruits avariés. Chaque client qui trouve, dans l'emballage qu'il achète, deux fruits ou plus qui sont avariés, revient au magasin se plaindre.

- Soit  $X$  le nombre de pommes avariées dans un emballage. Déterminer la distribution de probabilité de  $X$ , i.e. trouver  $P(X = 0), P(X = 1), \dots, P(X = 5)$ .
- Quelle est la probabilité qu'un client donné se plaigne auprès de son épicier ?
- Si l'épicier a 100 clients qui achètent des pommes ce jour-là, combien de plaintes devra-t-il attendre ?

6.3 Soit  $X$  une variable aléatoire dont la fonction de densité est:

$$f(x) = \begin{cases} c(1 - x^2) & \text{si } -1 < x < 1, \\ 0 & \text{sinon.} \end{cases}$$

- Calculer la valeur de  $c$ .
- Quelle est la fonction de distribution cumulative de  $X$  ?

6.4 Considérons la variable aléatoire  $X$  de densité

$$f(x) = \begin{cases} 2x & \text{si } 0 < x < 1, \\ 0 & \text{sinon.} \end{cases}$$

On veut étudier la variable aléatoire  $Y = \exp(-X)$ .

- Déterminer sa fonction de distribution cumulative  $F_Y(y) = P(Y \leq y)$ ,
- Déterminer sa densité de probabilité  $f_Y$ .

Facultatifs:

- Représenter graphiquement  $F_Y$  et  $f_Y$ .
- Calculer  $E(X), V(X), E(Y), V(Y)$ .

6.5 On jette deux dés homogènes et on note par  $(i, j)$  les chiffres obtenus. On considère les variables aléatoires  $X = \max(i, j)$  et  $Y = \text{Nombre de 5 ou 6}$ .

- Trouver la distribution de probabilité de  $X$  et  $Y$ .
- On pose  $Z = XY$ ; trouver la distribution de probabilité de  $Z$ .
- Calculer  $E(X), E(Y)$  et  $E(Z)$ .

Facultatif (non requis pour l'examen):

6.6 Pour étudier la performance d'un moteur de fusée, on considère les deux caractéristiques suivantes:  $X$  la poussée et  $Y$  le mélange de carburant. On suppose que  $(X, Y)$  est une variable aléatoire continue et que sa fonction de densité conjointe est donnée par

$$f_{XY}(x, y) = \begin{cases} 2(x + y - 2xy) & \text{si } 0 < x < 1, 0 < y < 1, \\ 0 & \text{sinon.} \end{cases}$$

Déterminer:

- (a) la fonction de distribution cumulative conjointe  $F_{XY}$ ,
- (b) les densités marginales de  $X$  et  $Y$ , notées  $f_X$  et  $f_Y$ ,
- (c) les densités conditionnelles de  $X$  sachant que  $Y = y$  et de  $Y$  sachant que  $X = x$ , notées respectivement  $f_{X|Y=y}$  et  $f_{Y|X=x}$ .
- (d)  $X$  et  $Y$  sont-elles indépendantes ? justifier la réponse.

## Exercices du Chapitre 7

7.1 Considérons un groupe de 4 individus choisis parmi une population de hollandais âgés de 60 à 75 ans. Le nombre de personnes dans cet échantillon, souffrant d'hypertension, est une variable aléatoire binômiale de paramètres  $n = 4$  et  $p = 0.15$ .

- Quelles sont les probabilités que aucun, un, deux, trois puis quatre des personnes souffrent d'hypertension?
- Représenter la distribution de probabilité de la variable.
- Représenter sa fonction de distribution cumulative.

7.2. Un ivrogne se promène dans l'unique rue de son village. A chaque portail il s'arrête et il essaie de l'ouvrir; puis il repart dans le même sens où dans le sens contraire avec probabilité  $1/2$ . L'homme commence par essayer le portail se trouvant en face du bistro, situé au milieu du village, puis part vers la gauche ou vers la droite. Quelle est la probabilité que le cinquième portail soit le même que le premier ? (Dans la théorie du calcul des probabilités, on dit que la promenade de cet ivrogne est une "marche aléatoire").

7.3 On considère deux variables aléatoires  $X_1$  et  $X_2$  indépendantes, telles que  $X_1 \sim \mathcal{B}(n_1, p)$  et  $X_2 \sim \mathcal{B}(n_2, p)$ . Quelle est la distribution de probabilité de la variable  $Y = X_1 + X_2$  ?

7.4 Dans un processus d'embouteillage, la quantité (en cl) de boisson mise en bouteille est une variable aléatoire normale de moyenne  $\mu$  et de variance  $\sigma^2 = 4$ . Une bouteille contenant moins d'un litre (100cl) de cette boisson n'est pas acceptable pour la vente. Calculer la probabilité qu'une bouteille ne soit pas acceptée pour  $\mu = 101$ ,  $\mu = 102$  et  $\mu = 103$ .

7.5 Supposons que la quantité de potassium contenue dans un verre de coca-cola soit représentée par une variable aléatoire  $X \sim \mathcal{N}(7\text{mg}, (0.4\text{mg})^2)$ . Supposons que vous buviez trois verres de coca et soit  $T$  la quantité totale (en mg) de potassium que vous recevez de ces trois verres. Déterminer la moyenne et l'écart-type de  $T$ .

Facultatif (non requis pour l'examen):

7.6 La population d'un certain pays est composée à 40% de Pygmées et à 60% de Watousis. La taille des Pygmées (en centimètres) a une distribution gaussienne  $N(120, 20^2)$  et la taille des Watousis une distribution  $N(200, 40^2)$ . Soit  $X$  la taille d'un individu choisi au hasard dans cette population. Calculer la probabilité que  $X$  soit comprise entre 120 et 200 centimètres. Calculer la moyenne et la variance de  $X$ .

## Exercices du Chapitre 8

8.1 On lance une pièce de monnaie douze fois, la probabilité d'obtenir pile étant inconnue et égale à  $p$ . Supposons que l'expérience a fourni le résultat suivant:

$$\{F, P, P, F, F, P, F, F, P, F, P, P\},$$

où  $F$  est l'abréviation pour "face" et  $P$  celle pour "pile". Donner une expression pour la fonction de vraisemblance en supposant que le résultat d'un jet est une épreuve de Bernoulli de paramètre  $p$ . En déduire une estimation de  $p$  par maximum de vraisemblance.

8.2 Les oiseaux d'un certain type prennent leur envol après avoir effectué quelques sauts sur le sol. On suppose que ce nombre  $X$  de saut peut être modélisé par une distribution géométrique:

$$P(X = x) = p(1 - p)^x, \quad x \geq 0.$$

Notons que la distribution géométrique correspond au nombre d'essais jusqu'au premier succès si l'on procède à des répétitions indépendantes d'épreuves de Bernoulli avec probabilité de succès  $p$ . Pour  $n = 130$  oiseaux de ce type, on a relevé les données suivantes:

nombre $x$ de sauts	1	2	3	4	5	6	7	8	9	10	11	12
fréquence de $x$	48	31	20	9	6	5	4	2	1	1	2	1

(a) Montrer que l'estimateur du maximum de vraisemblance de  $p$  est donné par

$$\hat{p} = \frac{1}{(\sum_{i=1}^n X_i)/n + 1}.$$

(b) Calculer la valeur de  $\hat{p}$  obtenue avec les données.

8.3 Démontrer la formule de la fonction de vraisemblance du modèle de Gauss donnée dans le cours. En déduire les estimateurs du maximum de vraisemblance pour les paramètres  $\mu$  et  $\sigma^2$ .

8.4 Une variable aléatoire  $X$  a pour densité  $f(x) = (\beta + 1)x^\beta$  où  $0 < x < 1$  et  $\beta > -1$ .

- (a) Donner l'estimateur du maximum de vraisemblance basé sur un échantillon de taille  $n$ .
- (b) Calculer l'estimation du maximum de vraisemblance de  $\beta$  basée sur l'échantillon suivant: 0.3, 0.8, 0.27, 0.35, 0.62, 0.55.

8.5 Considérons à nouveau les nombres d'étamines de 100 fleurs de l'exercice 1, Chapitre 1. Peut-on dire que le modèle de Gauss est une bonne description de ces données ? Peut-on dire que le modèle de Gauss est une bonne description du logarithme du nombre d'étamines ?

## Exercices du Chapitre 9

9.1 Supposons que la quantité de potassium contenue dans un verre de coca-cola soit représentée par une variable aléatoire  $X \sim \mathcal{N}(7\text{mg}, (0.4\text{mg})^2)$ . Supposons que vous buviez trois verres de coca et soit  $T$  la quantité totale (en mg) de potassium que vous recevez de ces trois verres. Calculer la probabilité que la quantité de potassium que vous avez ingurgité grâce à ces trois verres dépasse 15 mg.

9.2 Un dé homogène est jeté 1200 fois et soit  $X$  le nombre de fois que la face 6 apparaît. Calculer approximativement  $P(180 \leq X \leq 220)$ .

9.3 On suppose que le nombre de globules blancs par unité de volume d'une solution diluée de sang (ces globules étant comptés au moyen d'un microscope) suit une distribution de Poisson de moyenne 100. Calculer la probabilité d'en observer 90 au moins lors de la prochaine expérience.

9.4 La durée de vie de chaque ampoule d'un certain lot de  $N$  ampoules a une distribution de moyenne et d'écart type égaux à 1000 heures. Les durées de vie d'ampoules différentes sont indépendantes et identiquement distribuées. Soit  $T$  la durée de vie totale des  $N$  ampoules du lot (c'est-à-dire la somme des durées de vie des  $N$  ampoules). Calculer approximativement:

- (a)  $P(T > 115000 \text{ heures})$  pour  $N = 100$ ;
- (b) La plus petite valeur de  $N$  telle que  $P(T > 50000) \geq 0.95$ ;

### Facultatif

- (c) La plus grande valeur  $t$  telle que  $P(T > t) \geq 0.95$  si  $N = 100$ .

### Les exercices suivants sont facultatifs

9.5 Seize nombres aléatoires sont extraits d'une distribution discrète qui donne probabilité  $1/10$  à chacune des valeurs suivantes: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Calculer par approximation normale la probabilité que la moyenne arithmétique de ces seize nombres soit comprise entre 4 et 6.

9.6 Soit  $Y$  une variable aléatoire avec distribution uniforme sur l'intervalle  $(0, 1)$ .

- (a) Montrer que la fonction de distribution cumulative de la variable aléatoire  $X = F^{-1}(Y)$  est la fonction  $F$ .
- (b) Soit  $x_1, \dots, x_n$  un échantillon de taille  $n$  provenant d'une distribution  $F$  et soit  $F_n(x) = (\text{nombre de } x_i \leq x)/n$  la fonction de distribution cumulative empirique. Montrer que pour simuler un échantillon de taille  $n$  d'une variable aléatoire distribuée selon  $F_n$ , il suffit de tirer avec remise  $n$  éléments de l'ensemble  $\{x_1, \dots, x_n\}$ .

9.7 Soient  $X_1, \dots, X_n$ ,  $n$  variables aléatoires indépendantes et identiquement distribuées selon une distribution de Poisson de paramètre  $\lambda$ . Soit  $\bar{X} = (X_1 + \dots + X_n)/n$  leur moyenne arithmétique.

- (a) Montrer que  $\bar{X}$  est un estimateur sans biais de  $\lambda$
- (b) Montrer que  $\bar{X}^2$  n'est pas un estimateur sans biais de  $\lambda^2$

## Exercices du Chapitre 10

10.1 Dans un test à choix multiple on pose à un candidat 20 questions. Chaque question a quatre réponses possibles dont une seule est juste. On suppose que si le candidat n'a pas de connaissances, il choisit au hasard une des quatre réponses. Si le candidat répond correctement à 9 au moins des 20 questions, on considère qu'il possède des connaissances.

- Calculer la probabilité de considérer le candidat comme ayant des connaissances alors qu'il ne fait que deviner les réponses.
- Formuler la règle de décision sous forme de test statistique
- L'erreur mentionné sous (a) est-elle de type I ou de type II ?

10.2 Dans beaucoup d'espèces, la probabilité  $p$  de mourir durant l'année est indépendante de l'âge. La distribution des âges à la mort suit alors une distribution géométrique pour laquelle  $P(X = x) = p(1-p)^x$  ( $X$  étant l'âge au moment du décès). Nous nous intéressons à une population où la probabilité de disparaître au cours d'une année est supposée égale à 0.1. Lors d'un échantillonnage aléatoire de deux individus dans cette population, nous obtenons deux individus âgés de plus de 20 ans. Ce résultat permet-il de soutenir l'hypothèse selon laquelle la probabilité de décès est de 0.1? Sinon, proposer des valeurs de  $p$  plausibles.

10.3 Considérons une variable aléatoire suivant une distribution normale  $N(\mu, 1)$ . On désire tester  $H_0 : \mu = 6$  contre  $H_1 : \mu = 7$  sur la base d'un échantillon de taille  $n = 4$ . On rejette  $H_0$  si  $(X_1 + X_2 + X_3 + X_4)/4 \geq 7$ . Calculer les probabilités de commettre des erreurs de première et deuxième espèces.

10.4 Considérons la population des hommes de 12 à 40 ans. On s'intéresse à leur taille moyenne  $\mu$  inconnue. Supposons donné l'écart type de leur tailles  $\sigma = 6$  centimètres. Au seuil  $\alpha = 1\%$ , tester l'hypothèse  $H_0 : \mu = 160$  cm contre  $H_1 : \mu \neq 160$  cm en sachant que l'on a calculé la moyenne  $\bar{x}$  des tailles de 31 individus d'un échantillon et que  $\bar{x} = 147.4$  cm a été trouvé.

10.5 Dans un Gymnase d'un canton suisse, des enquêtes ont montré que la proportion de jeunes fumeurs était de 40%. Le directeur de l'établissement désire savoir si cette proportion a diminué grâce à une campagne anti tabac. Pour cela, il fait une enquête auprès de 20 jeunes du Gymnase dont il ressort que 8 sont fumeurs. Peut-on rejeter au seuil de 5% l'hypothèse  $H_0$  : "la proportion de fumeurs est de 40%" contre l'alternative  $H_1$  : "la proportion de fumeurs est inférieur à 40%" ?

Indication: on prendra comme statistique le nombre  $K$  de fumeurs et on supposera que sous  $H_0$  elle suit une distribution binomiale  $\mathcal{B}(n = 20, p = 0.4)$  donnée ci-dessous:

$k:$	0	1	2	3	4	5	6	7	8	9	10
$P(K = k):$	3.65e-05	.00048	.003	.012	.035	.074	.12	.16	.18	.16	.11
$k:$	11	12	13	14	15	16	17	18	19	20	
$P(K = k):$	.07	.035	.014	.0048	.0013	.00027	4.23e-05	4.7e-06	3.3e-07	1.01e-08	

## Exercices du Chapitre 11

11.1 Parmi les “lois de Murphy” les plus célèbres il y a la suivante: “si on laisse tomber une tartine à la confiture, il est plus probable qu’elle tombe sur la face tartinée plutôt que sur l’autre”. Pour la vérifier, une expérience à été réalisée auprès de la University of Southwestern Louisiana, dans laquelle on a fait tomber 1000 tartines à la confiture de groseille (*Ribes vulgare*). 540 tartines sont tombées sur la face avec la confiture.

- (a) Formulez une hypothèse nulle et une hypothèse alternative dans le but de tester cette loi.
- (c) Peut-on accepter cette loi au niveau de 5% ?
- (a) Donnez un intervalle de confiance (bilatéral) de couverture 95% pour la probabilité que la tartine tombe sur la face avec la confiture.

11.2 Une autre loi de Murphy affirme que “la probabilité de succès est une fonction croissante du dégât que l’on peut causer”. Dans le but de la vérifier, à la University of Southwestern Louisiana on a fait tomber 1000 tartines à la confiture de groseilles des bois (*Ribes sylvestre*): 400 sur le terrain de basket et 600 sur un magnifique tapis de Perse. Dans le premier cas, 220 tartines sont tombées sur la face avec la confiture; dans le deuxième cas, 350.

- (a) Formulez une hypothèse nulle et une hypothèse alternative dans le but de tester cette loi.
- (b) Peut-on accepter cette loi au niveau de 5% ?

11.3 Selon le journal Z le 50% de ses lecteurs lit les annonces publicitaires. Une agence désire vérifier cette affirmation. Elle choisit au hasard 100 lecteurs et en trouve 40 qui lisent les annonces.

- (a) Tester l’affirmation du journal en prenant un seuil de signification de 1%.
- (b) Donner un intervalle de confiance de couverture 99% pour la proportion de lecteurs qui lisent les annonces.

11.4 On s’intéresse à l’association entre le mode de vie, “seul” ou “en famille”, et la présence ou l’absence d’une névrose. Dans un échantillon aléatoire d’individus d’une certaine population on a trouvé les fréquences ci-dessous:

Mode de vie	Névrose		Total
	Présente	Absente	
En famille	40	60	100
Seul	100	60	160
Total	140	120	200

Cette étude est évidemment transversale, car elle se base sur un seul échantillon. Peut-on rejeter, au seuil de 1%, l’hypothèse de non association entre le mode de vie et la présence d’une névrose ?

## Exercices du Chapitre 12

12.1 On désire tester la durée de vie moyenne d'un tube électronique avec un seuil de 0.01. L'hypothèse nulle est que cette durée moyenne est  $H_0 : \mu = 1600$  heures, tandis que l'alternative est  $H_1 : \mu \neq 1600$ . 16 observations de cette durée de vie nous donnent une moyenne de 1590 (heures) et un écart type  $s = 30$  (heures)

- Déterminer la statistique à utiliser pour tester  $H_0$  contre  $H_1$ .
- Donner l'intervalle de rejet.
- Effectuer le test.

12.2 Un test de mathématique a été donné à 25 garçons et à 16 filles d'un collège. Les garçons ont obtenu une note moyenne de 82 et un écart-type de la note de 8. Les filles ont obtenu une moyenne de 78 et un écart type de la note de 7. Les notes sont réparties normalement et les variances des deux sous-populations sont supposées égales.

Tester avec un seuil de  $\alpha = 0.05$  l'hypothèse que la note des filles est significativement inférieure à celle des garçons.

12.3 Comparer la température du sang d'un animal à une valeur théorique spécifiée,  $\mu_0 = 39.6$  (degrés), avec  $\alpha = 0.05$  en considérant les données suivantes:

37.9	38.4	38.8	38.8	39.0	39.0	39.2	39.2	39.2	39.3
39.4	39.7	39.7	39.9	40.0	40.0	40.2	40.4	40.6	41.0

12.4 Les contenus de nicotine de 5 cigarettes d'une certaine marque ont montré une moyenne de 21.2 milligrammes et un écart type estimé de 2.05 mg. Tester avec un seuil de 5% l'hypothèse que le contenu moyen de nicotine de cette marque de cigarette vaut 19.7 mg.

12.5 Neuf adultes ont accepté de tester l'efficacité d'un nouveau programme diététique. Leurs poids (en livres) ont été mesurés avant et après le programme. Ils sont présentés dans le tableau ci-dessous:

sujet:	1	2	3	4	5	6	7	8	9
Avant:	132	139	126	114	122	132	142	119	126
Après:	124	141	118	116	114	132	145	123	121

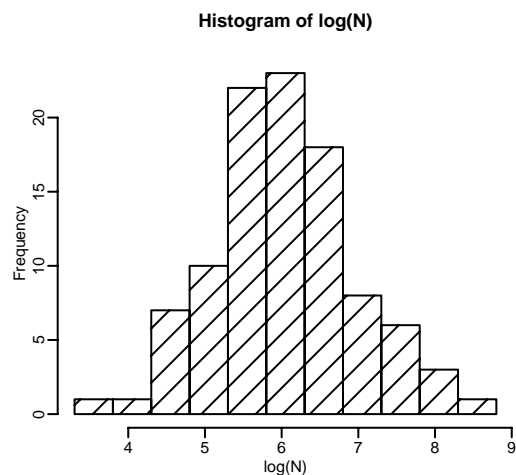
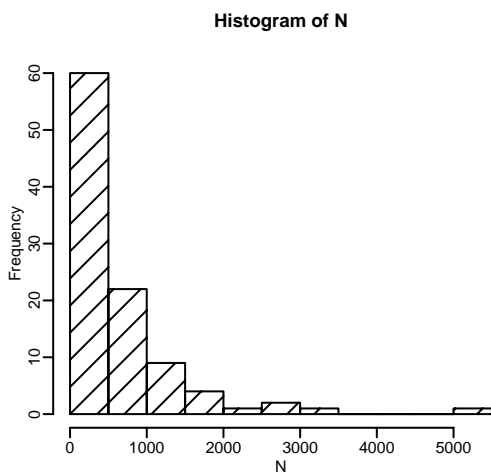
Au niveau  $\alpha = 1\%$ , le programme est-il efficace?

# Solutions

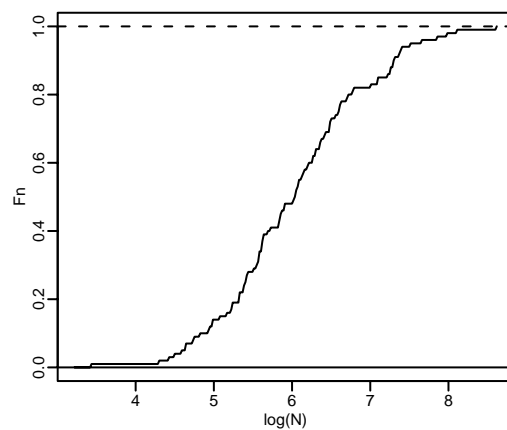
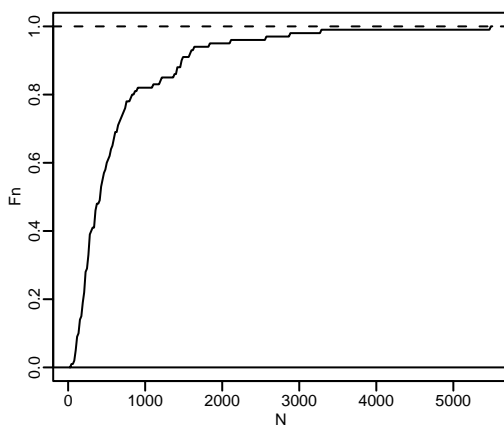
## Solutions des exercices du Chapitre 1

1.1 Pour réaliser l'histogramme de la variable  $N =$  nombre d'étamines, nous partageons l'intervalle  $[0, 5500]$  en 11 sous-intervalles et calculons les fréquences de  $N$  dans chaque sous-intervalle. Pour tracer l'histogramme de  $\log(N)$  nous partageons l'intervalle  $[2.6, 9.2]$  en 11 sous-intervalles etc. (Noter que  $3.3 < \log(31)$  et  $8.8 > \log(5481)$ ).

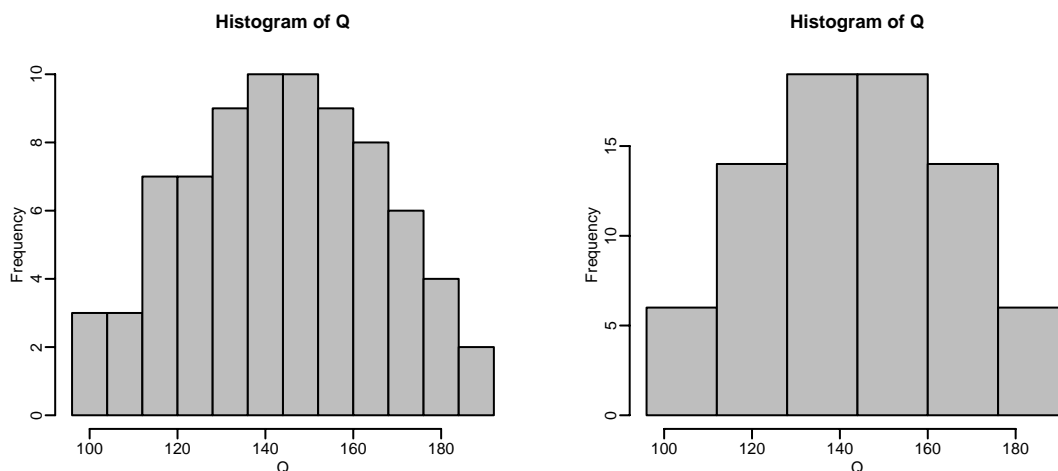
Intervalles pour $N$	Fréquences	Intervalles pour $\log(N)$	Fréquences
( 0, 500]	60	(3.3, 3.8]	1
( 500, 1000]	22	(3.8, 4.3]	1
(1000, 1500]	9	(4.3, 4.8]	7
(1500, 2000]	4	(4.8, 5.3]	10
(2000, 2500]	1	(5.3, 5.8]	22
(2500, 3000]	2	(5.8, 6.3]	23
(3000, 3500]	1	(6.3, 6.8]	18
(3500, 4000]	0	(6.8, 7.3]	8
(4000, 4500]	0	(7.3, 7.8]	6
(4500, 5000]	0	(7.8, 8.3]	3
(5000, 5500]	1	(8.3, 8.8]	1



Remarquons que l'histogramme de  $N$  est très asymétrique. Celui de  $\log(N)$  est presque symétrique.

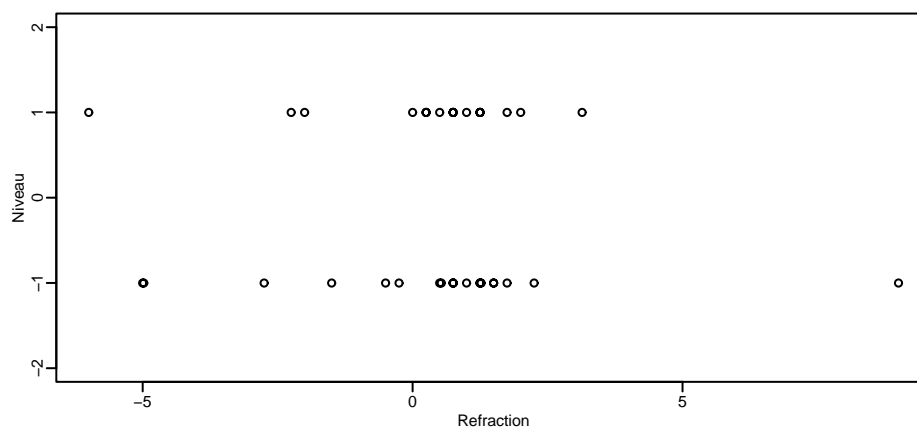


1.2 Soit  $Q$  la quantité de cerises. Les histogrammes sont donnés ci-dessous.



L'histogramme avec six classes ( $[96, 112), \dots, [176, 192)$ ) ne révèle pas la légère asymétrie visible dans l'histogramme à 12 classes. Il est donc important de choisir un nombre de classes suffisamment grand (mais pas trop: on pourrait perdre la vision succincte de la structure).

1.3 Dans la représentation graphique, le niveau faible est codé  $-1$  et le niveau élevé est codé  $+1$ . La représentation graphique ne soutient pas l'hypothèse que le déficit alimentaire protéique est associé à la myopie. On observe une mesure de réfraction atypique et très élevée (9.00) dans la première série (niveau faible).



## Solutions des exercices du Chapitre 2

2.1 Soit  $X$  le vecteur à  $m = 20$  composantes contenant les mesures à faible niveau protéique et  $Y$  le vecteur à  $n = 17$  composantes contenant les mesures à niveau protéique élevé.

(a) On pose  $\alpha = 0.25$  et  $\beta = 0.75$  et on obtient:

$$m(X) = 0.453, \text{ med}(X) = 0.750, s(X) = 2.805,$$

$$[[m\alpha]] = 5, [[m\beta]] = 15, q_\alpha(X) = -0.375, q_\beta(X) = 1.385, I_q(X) = 1.76;$$

$$m(Y) = 0.273, \text{ med}(Y) = 0.750, s(Y) = 2.007,$$

$$[[n\alpha]] = 4, [[n\beta]] = 12, q_\alpha(Y) = 0.00, q_\beta(Y) = 1.25, I_q(Y) = 1.25.$$

Le logiciel R utilise une version lissée de la fonction de distribution cumulative empirique dans le calcul des quantiles. Il donne les résultats suivants:

$$q_\alpha(X) = -0.3125, q_\beta(X) = 1.3275, I_q(X) = 1.64,$$

$$q_\alpha(Y) = -0.2500, q_\beta(Y) = 1.2500, I_q(Y) = 1.00.$$

(b)  $m(X) > m(Y)$ , mais les médianes sont identiques. On ne peut pas conclure que ces résultats soutiennent l'hypothèse.

(c) Après suppression de la donnée 9.00 dans  $X$ , on obtient:

$$m(X) = 0.00, \text{ med}(X) = 0.750, s(X) = 2.058,$$

$$[[m\alpha]] = 4, [[m\beta]] = 14, q_\alpha(X) = -1.5, q_\beta(X) = 1.25, I_q(X) = 2.75.$$

Avec R, on obtient  $I_q = 1.635$ . On remarque que  $m(X)$  et  $s(X)$  changent de façon importante, tandis que  $\text{med}(X)$  et  $I_q(X)$  (version lissée de R) sont presque insensibles à la suppression de la mesure extrême.

2.2 Soit  $N$  le nombre d'étamines. Après élimination de la plus petite valeur dans les deux ensembles de données on obtient:

(a)  $m(N) = 659.32$ ,  $\text{med}(N) = 422$ ,  $m(\log(N)) = 6.056$ ,  $\text{med}(\log(N)) = 6.045$ .

(b) La médiane représente le milieu de la distribution; la moyenne est influencée par les valeurs extrêmes (en particulier par les valeurs élevées de  $N$ ). La moyenne et la médiane de  $N$  sont très différentes car la distribution de  $N$  est très asymétrique. La médiane et la moyenne de  $\log(N)$  sont très proches, car la distribution de  $\log(N)$  est presque symétrique.

(c) On observe que  $\log(\text{med}(N)) = 6.045 = \text{med}(\log(N))$ .

Ceci s'explique de la façon suivante. Soient  $n_{[1]} \leq \dots \leq n_{[99]}$  les valeurs ordonnées de  $N$ . Donc  $\text{med}(N) = n_{[50]}$  et  $\log(\text{med}(N)) = \log(n_{[50]})$ . En outre, comme  $\log$  est une fonction monotone croissante, les valeurs ordonnées de  $\log(N)$  sont  $\log(n_{[1]}) \leq \dots \leq \log(n_{[99]})$ . Donc  $\text{med}(\log(N)) = \log(n_{[50]}) = \log(\text{med}(N))$ .

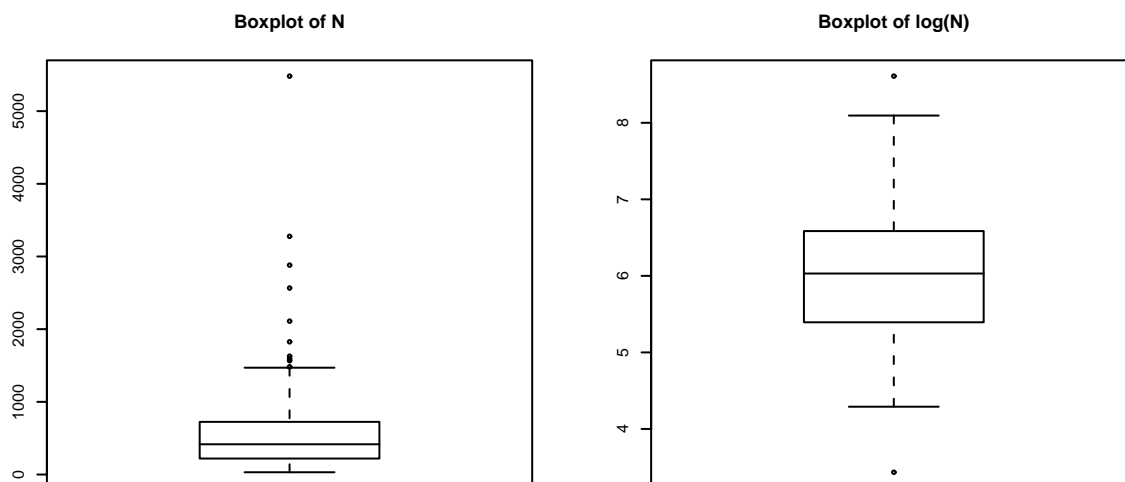
(g) Avec toutes les données (100), on obtient  $\text{med}(\log(N)) = 6.030581 \approx 6.030685 = \log(\text{med}(N))$ . Il n'y a pas une parfaite égalité car  $\text{med}(\log(N)) = [\log(n_{[50]}) + \log(n_{[51]})]/2 \neq \log[(n_{[50]} + n_{[51]})/2]$ .

(d) Évidemment, la propriété de la médiane que l'on vient d'établir s'étend à tout quantile (de façon approximative ou exacte selon la définition).

(e) On peut étendre les propriétés susmentionnées à d'autres transformations, pourvu qu'elle soient monotones croissantes. (Que se passe-t-il pour des transformations décroissantes ?) La propriété ne vaut pas pour des transformations non monotones: par exemple,  $\text{med}(\sin(N)) = -0.18472 \neq 0.85554 = \sin(\text{med}(N))$ .

(f) On a  $m(\log(N)) = 6.056395 < 6.419214 = \log(m(N))$ . La propriété ne vaut pas pour la moyenne.

2.3 Les boxplots sont donnés ci-dessous. La transformation logarithmique rend la distribution presque symétrique.



#### 2.4 Propriétés de la moyenne.

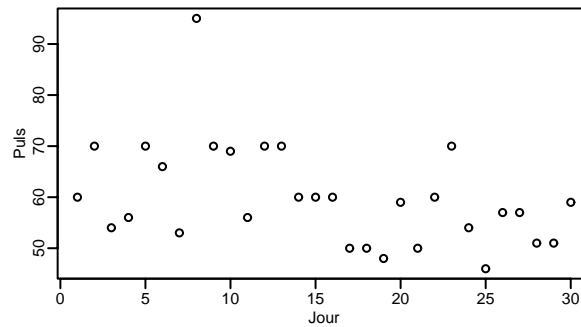
- (a) est évident.
- (b)  $m(aX) = (ax_1 + \dots + ax_n)/n = a(x_1 + \dots + x_n)/n = am(X)$ .
- (c)  $m(X+a) = (x_1+a + \dots + x_n+a)/n = (x_1 + \dots + x_n + na)/n = (x_1 + \dots + x_n)/n + na/n = m(X) + a$ .
- (d)  $m(X+Y) = (x_1+y_1 + \dots + x_n+y_n)/n = (x_1 + \dots + x_n + y_1 + \dots + y_n)/n = m(X) + m(Y)$ .  
En utilisant cette propriété, ainsi que (b) et (c), on obtient:  
 $m(aX + bY + c) = m(aX) + m(bY) + c = am(X) + bm(Y) + c$ .
- (e) Il suffit de donner un contreexemple. Soit  $X = (1, 9)$  et  $Y = (2, 0)$ . Donc  $m(X) = 5$ ,  $m(Y) = 1$  et  $m(X)m(Y) = 5$ . D'autre part,  $XY = (2, 0)$  et  $m(XY) = 1$ .

#### 2.5 Pour démontrer les propriétés de la variance et de l'écart type on utilise les définitions et les propriétés de la moyenne.

- (a) On a  $m(c) = c$ . Donc,  $s^2(c) = m([c - m(c)]^2) = m(0^2) = 0$ .
- (b)  $s^2(aX + b) = m([aX + b - m(aX + b)]^2) = m([aX + b - am(X) - b]^2)$   
 $= m([aX - am(X)]^2) = a^2m([X - m(X)]^2) = a^2s^2(X)$ .
- (c) En prenant la racine carrée de  $s^2(aX + b) = a^2s^2(X)$  on obtient  $s(aX + b) = as(X)$ .
- (d) Il suffit de donner un contreexemple. Soit  $X = (1, 2)$  et  $Y = (1, 2)$ . Donc  $s^2(X) = 0.5^2$  et  $s^2(Y) = 0.5^2$ . D'autre part,  $X + Y = (2, 4)$  et  $s^2(X + Y) = 1 \neq 0.5^2 + 0.5^2$ .
- (e)  $\sum(x_i - m(X)) = \sum x_i - nm(X) = nm(X) - nm(X) = 0$ .
- (f)  $s^2(X) = m([X - m(X)]^2) = m(X^2 - 2Xm(X) + m(X)^2)$   
 $= m(X^2) - 2m(X)m(X) + m(X)^2 = m(X^2) - m(X)^2$ .

## Solutions des exercices du Chapitre 3

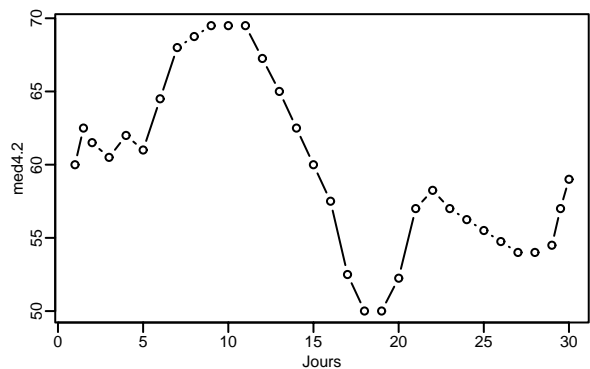
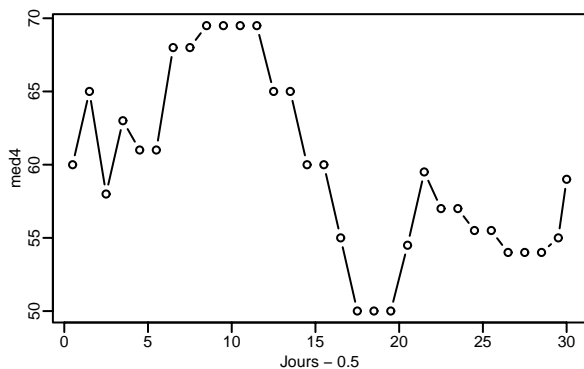
3.1 (a) Le graphique montre un pic au 8ème jour; il suggère uniquement une faible tendance à la baisse.



Les calculs des “pulsations lissées” selon (b) et (c) sont indiqués dans le tableau suivant. (Les nombres en italique sont des médianes sur une ou deux valeurs.)

Jour	Pulsation	Médiane de 4 pulsations	Médiane de 2 médianes
1.0	60	<i>60.00</i>	<i>60.00</i>
1.5		<i>65.00</i>	<i>62.50</i>
2.0	70		61.50
2.5		58.00	
3.0	54		60.50
3.5		63.00	
4.0	56		62.00
4.5		61.00	
5.0	70		61.00
5.5		61.00	
6.0	66		64.50
6.5		68.00	
7.0	53		68.00
7.5		68.00	
8.0	95		68.75
8.5		69.50	
9.0	70		69.50
9.5		69.50	
10.0	69		69.50
10.5		69.50	
11.0	56		69.50
11.5		69.50	
12.0	70		67.25
12.5		65.00	
13.0	70		65.00
13.5		65.00	
14.0	60		62.50
14.5		60.00	
15.0	60		60.00

Jour	Pulsation	Médiane de 4 pulsations	Médiane de 2 médianes
15.5		60.0	
16.0	60		57.50
16.5		55.0	
17.0	50		52.50
17.5		50.0	
18.0	50		50.00
18.5		50.0	
19.0	48		50.00
19.5		50.0	
20.0	59		52.25
20.5		54.5	
21.0	50		57.00
21.5		59.5	
22.0	60		58.25
22.5		57.0	
23.0	70		57.00
23.5		57.0	
24.0	54		56.25
24.5		55.5	
25.0	46		55.50
25.5		55.5	
26.0	57		54.75
26.5		54.0	
27.0	57		54.00
27.5		54.0	
28.0	51		54.00
28.5		54.0	
29.0	51		54.50
29.5		55.00	57.00
30.0	59	59.00	59.00

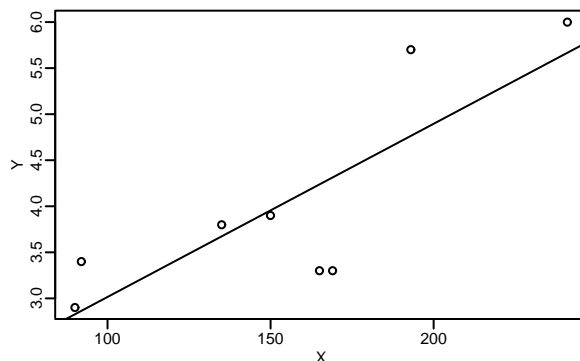


Les pulsations lissées selon (a) et (b) sont représentées dans les deux graphiques ci-dessus. Sur le graphique des “médianes de 4”, une courbe commence à apparaître. On observe des extréma; le pic du graphique (a) s’est transformé en un maximum aux jours 9 à 12. Sur le graphique des “médianes des 2 médianes” la courbe est très nette. Il y a un maximum

entre le 7ème et le 13ème jour, suivi d'un minimum au 19ème jour. Le "lissage" montre donc le cycle de la température de la vache. (Le deuxième pic est plus bas que le premier car la batterie du thermomètre se décharge).

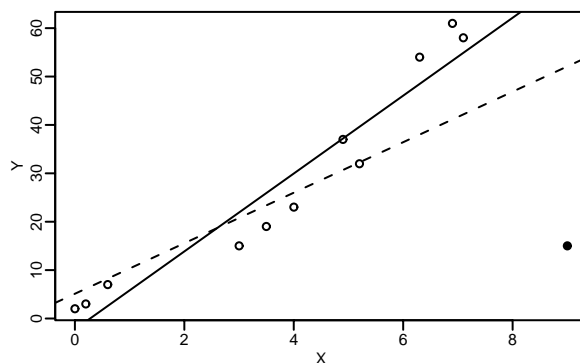
3.2 Soit  $X$  le poids et  $Y$  le prix. On obtient:

$$m(X) = 154.37, m(Y) = 4.03, s^2(X) = 2218.98, s^2(Y) = 1.18, v(X, Y) = 41.72, r(X, Y) = 0.814, b = 0.0188 \text{ et } a = 1.135.$$



3.3 On obtient:

$$m(X) = 3.79, m(Y) = 28.27, s^2(X) = 6.21, s^2(Y) = 434.38, v(X, Y) = 49.99, r(X, Y) = 0.962, b = 8.05 \text{ et } a = -2.24.$$



La droite  $Y = -2.24 + 8.05X$  est représentée dans le graphique par la ligne continue.

Les valeurs calculées  $\hat{Y}$  sont:

$$-2.237, -0.628, 2.592, 21.907, 25.931, 29.956, 39.613, 37.199, 48.466, 54.905, 53.295.$$

Les résidus  $E$  sont:

$$4.237, 3.628, 4.408, -6.907, -6.931, -6.956, -7.613, -0.199, 5.534, 3.0957.705.$$

On vérifie immédiatement que la somme  $\sum e_i$  des résidus est nulle. En outre, on obtient:  $s^2(\hat{Y}) = 402.36$ ,  $s^2(E) = 32.02$ , et donc  $s^2(Y) = 434.38 = s^2(\hat{Y}) + s^2(E)$ . On observe aussi que  $b^2 s^2(X) = 402.36$ .

Si on ajoute le point (9, 15) on obtient  $a = 5.11$  et  $b = 5.22$ . La droite  $Y = 5.11 + 5.22X$  est représentée dans le graphique par la ligne en traitillé. On observe donc que la droite de régression est influencée par l'adjonction (la suppression) d'un cas atypique (outlier).

## Solutions des Exercices du Chapitre 4

4.1 Soit  $A$  = “agées de plus de 30 ans”,  $B$  = “mariés”,  $C$  = “femmes”. On note  $\#(I)$ , le nombre d’éléments d’un ensemble  $I$ . On a alors:

$$\#(A) = 525, \#(B) = 470, \#(C) = 312,$$

$$\#(A \cap B) = 147, \#(B \cap C) = 86, \#(A \cap C) = 42, \#(A \cap B \cap C) = 25.$$

On pose:

$$S_1 = (A \cap B \cap C),$$

$$S_2 = (A \cap B) \setminus S_1,$$

$$S_3 = (B \cap C) \setminus S_1,$$

$$S_4 = (A \cap C) \setminus S_1,$$

$$S_5 = A \setminus (S_1 \cup S_2 \cup S_4),$$

$$S_6 = B \setminus (S_1 \cup S_2 \cup S_3),$$

$$S_7 = C \setminus (S_1 \cup S_3 \cup S_4).$$

Alors:

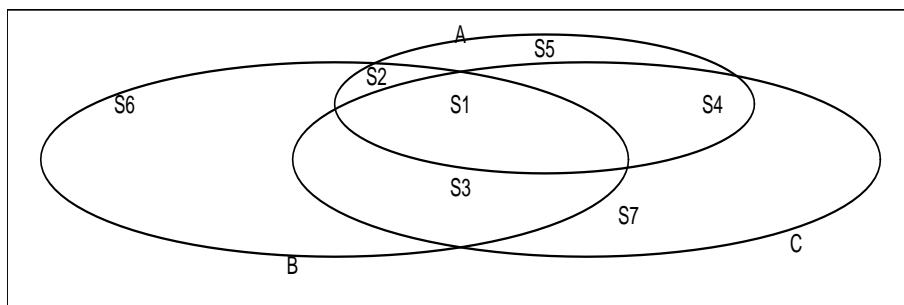
$$\#(S_1) = 25; \#(S_2) = 147 - 25 = 122; \#(S_3) = 86 - 25; \#(S_4) = 42 - 25 = 17;$$

$$\#(S_5) = 525 - (122 + 25 + 17) = 361;$$

$$\#(S_6) = 470 - (122 + 25 + 61) = 262;$$

$$\#(S_7) = 312 - (17 + 25 + 61) = 209.$$

Comme les  $S_i$  forment une partition de l’ensemble  $A \cup B \cup C$ , la somme des nombres d’éléments des  $S_i$  devrait donner 1000; or, le total est 1057.



4.2 On peut représenter les cas possibles par:

$$\{FFF, FFG, FGF, GFF, GGG, GGF, GFG, FGG\}.$$

On dénombre ainsi 8 cas possibles et on considère les événements:

$$A: \text{“le nombre de filles sera 2 ou 3”} = \{FFF, FFG, FGF, GFF\},$$

$$B: \text{“il y aura des enfants des deux sexes”} = \{FFG, FGF, GFF, GGF, GFG, FGG\}.$$

(a)  $P(A) = 4/8$  et  $P(B) = 6/8 = 3/4$ .

(b)  $A \cap B = \{FFG, FGF, GFF\}$ ,

$$P(A \cap B) = 3/8 \text{ et } P(A)P(B) = 3/8.$$

Les deux événements sont donc indépendants.

4.3 On note  $F$  l'évènement "la personne choisie fume" et  $A$  l'évènement "la personne choisie est alcoolique". On sait que  $P(F) = 0.47$ ,  $P(A) = 0.28$ ,  $P(F \cap A) = 0.18$ .

a) On démontre facilement que  $\overline{F \cap A} = \overline{F \cup A}$  (loi de Morgan), où  $\bar{A}$  dénote le complément de  $A$ ,  $\bar{F}$  le complément de  $F$ , etc. Donc,

$$P(\text{ni fumeur ni alcoolique}) = P(\overline{F \cap A}) = P(\overline{F \cup A}) = 1 - P(F \cup A) \\ = 1 - (P(F) + P(A) - P(F \cap A)) = 1 - 0.47 - 0.28 + 0.18 = 0.43.$$

b) On a:  $P(F \cap A) = 0.18 \neq 0.13 = P(F)P(A)$ . On en déduit que les deux vices ne sont pas indépendants.

4.4 On cherche la probabilité que le lièvre soit atteint par le chasseur A seulement ou par le chasseur B seulement ou par les deux. On sait que:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ . Donc  $P(\text{animal tué}) = 2/3 + 3/4 - (2/3 \cdot 3/4) = 0.91667$ .

4.5  $P(\text{au moins 2 personnes sont nées le même jour}) =$

$1 - P(\text{toutes les dates de naissance sont différentes}).$

$$P(\text{toutes les dates de naissance sont différentes}) = \frac{365 \times 364 \times \dots \times 356}{365^{10}} = 0.833.$$

Donc, la probabilité cherchée est 0.117.

4.6 Soit  $B_1$  l'évènement "B augmente" et  $A_0$  l'évènement "A est stable". Il s'agit de trouver  $P(B_1|A_0)$ . On a:

$$P(B_1|A_0) = \frac{P(B_1 \cap A_0)}{P(A_0)} = \frac{60/1000}{(60/1000 + 110/1000 + 80/1000)} = 0.24.$$

4.7 Soit  $A$  l'évènement "la compagnie trouve du pétrole en mer du Nord" et  $B$  l'évènement "la compagnie trouve du pétrole en méditerranée". Il est raisonnable de supposer que les deux événements sont indépendants:  $P(A \cap B) = P(A)P(B)$ . On en déduit:

$$P(A \cap \bar{B}) = P(\bar{B}|A)P(A) = (1 - P(B|A))P(A) = (1 - P(B))P(A) = P(A)P(\bar{B}). \text{ Donc,} \\ P(\text{un seul forage est fructueux}) = P(A \cap \bar{B}) + P(\bar{A} \cap B) = 0.8 \times 0.4 + 0.2 \times 0.6 = 0.44.$$

4.8 Soit  $N$  l'évènement "le chat est noir",  $G$  l'évènement "le chat est gris",  $R$  l'évènement "le chat vit dans la rue". A l'aide de la formule de Bayes et de la formule de la probabilité totale on obtient:

$$P(N|R) = \frac{P(R|N) \cdot P(N)}{P(R|N) \cdot P(N) + P(R|G) \cdot P(G)} = \frac{6/10 \cdot 4/10}{(6/10 \cdot 4/10 + 4/10 \cdot 6/10)} = 0.5.$$

4.9 On note  $A$  l'évènement "il pleut" et  $B$  "le baromètre prévoit de la pluie". On a:  $P(\bar{B}|A) = 0.1$  et  $P(B|\bar{A}) = 0.2$ ; donc  $P(B|A) = 0.9$ . A priori, on sait que la probabilité de pluie est de  $P(A) = 0.1$ ; donc, à l'aide de la formule de Bayes:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})} = \frac{9/10 \cdot 1/10}{(9/10 \cdot 1/10 + 2/10 \cdot 9/10)} = 0.3.$$

## Solution des Exercices du Chapitre 5

5.1 On a:  $P(\text{1 enfant normal}) = 2/4 = 1/2$ . On note:

$N$  : “enfant normal”,

$A$  : “enfant anormal”,

$E_1 = \{(NNNN)\}$  : “4 enfants normaux”,

$E_2 = \{(N N N A), (N N A N), (N A N N), (A N N N)\}$  : “3 enfants normaux et 1 anormal”,

$E_3 = \{(N N A A), (A A N N), (A N N A), (N A A N), (N A N A), (A N A N)\}$  : “2 enfants normaux et 2 anormaux”.

(a)  $P(E_1) = (1/2)^4 = 1/16$ ,

(b)  $P(E_2) = 4 \times 1/16$  et  $P(E_3) = 6 \times 1/16$ . Donc

$$P(\text{au moins enfants 2 normaux}) = P(E_1) + P(E_2) + P(E_3) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} = \frac{11}{16}.$$

(c) Soit  $H$ : “enfant hémophile”. On a:  $P(H) = 1/4$  et  $P(\overline{H}) = 3/4$ . En outre,  $P(\text{aucun enfant hémophile}) = (3/4)^4 = 81/256$ , donc:

$$P(\text{au moins 1 enfant hémophile}) = 1 - 81/256 = 175/256.$$

(d) Soit  $H$  : “fils hémophile” et  $G$  : “fille hétérozygote”.

Donc  $P(H) = P(YX') = 1/4$  et  $P(G) = P(XX') = 1/4$ .

Soit  $E_4$  : “2 fils hémophiles et 2 filles hétérozygotes”.

Donc  $E_4 = \{(HHGG), (GGHH), (GHHG), (HGGH), (HG HG), (GHGH)\}$  et

$$P(E_4) = 6 \times (1/4)^4 = 3/128.$$

(e)  $P(\text{aucun enfant hémophile}) = (3/4)^4 = 81/256$ .

5.2 Soit  $A$  : “la reine est porteuse du gène yeux bleus” et  $B_i$  : “le prince  $i$  a les yeux bleus”. On sait que  $P(A) = 0.5$  et  $P(B_i|A) = 0.5$ .

(a) On cherche  $P(A | \overline{B}_1 \cap \overline{B}_2 \cap \overline{B}_3)$ .

Posons  $F = \overline{B}_1 \cap \overline{B}_2 \cap \overline{B}_3$ . Par le Théorème de Bayes:

$$P(A|F) = \frac{P(F|A)P(A)}{P(F|A)P(A) + P(F|\overline{A})P(\overline{A})}.$$

Grâce à l'indépendance des  $B_i$ , on a:

$$P(F|A) = P(\overline{B}_1|A)P(\overline{B}_2|A)P(\overline{B}_3|A),$$

$$P(F|\overline{A}) = P(\overline{B}_1|\overline{A})P(\overline{B}_2|\overline{A})P(\overline{B}_3|\overline{A}).$$

Comme  $P(\overline{B}_i|\overline{A})=1$  et  $P(\overline{B}_i|A) = 1 - P(B_i|A) = 0.5$ , on obtient finalement:

$$P(A|F) = \frac{(1/2)^4}{(1/2)^4 + 1/2} = 1/9.$$

- (b) Soit  $B_4$  : “le quatrième prince a les yeux bleus”; on cherche  $P(B_4 | \overline{B}_1 \cap \overline{B}_2 \cap \overline{B}_3)$ . On rappelle que  $F$  : “les trois premiers princes ont les yeux non bleus”; on veut donc  $P(B_4|F)$ . On sait que:

$$P(\overline{B}_4|F) = \frac{P(\overline{B}_4 \cap F)}{P(F|A) \cdot P(A) + P(F|\overline{A}) \cdot P(\overline{A})}.$$

On pose  $G = \overline{B}_4 \cap F$  et on a  $P(G) = P(G|A)P(A) + P(G|\overline{A})P(\overline{A})$ . Grâce à l'indépendance des  $B_i$ , on a:

$$\begin{aligned} P(G|A) &= P(\overline{B}_4|A)P(\overline{B}_1|A)P(\overline{B}_2|A)P(\overline{B}_3|A), \\ P(G|\overline{A}) &= P(\overline{B}_4|\overline{A})P(\overline{B}_1|\overline{A})P(\overline{B}_2|\overline{A})P(\overline{B}_3|\overline{A}). \end{aligned}$$

Avec  $P(\overline{B}_i|\overline{A})=1$ , on a:

$$P(\overline{B}_4|F) = \frac{(1/2)^5 + 1 \cdot (1/2)}{(1/2)^4 + 1 \cdot (1/2)} = 17/18.$$

La probabilité cherchée est finalement:

$$P(B_4|F) = 1 - P(\overline{B}_4|F) = 1/18.$$

5.3 Soit  $A$  : “le prix du café augmente” et  $J$  : “le journal prévoit une augmentation du prix du café”. On a  $P(A) = 0.6$  donc  $P(\overline{A}) = 0.4$ ,  $P(J|A) = 0.8$ ,  $P(\overline{J}|\overline{A}) = 0.55$ , donc  $P(J|\overline{A}) = 0.45$ .

- (a) On évalue la probabilité que le prix du café augmente sachant que le journal a prévu une hausse soit:

$$P(A|J) = \frac{P(J|A)P(A)}{P(J|A)P(A) + P(J|\overline{A})P(\overline{A})} = \frac{0.8 \times 0.6}{0.8 \times 0.6 + 0.45 \times 0.4} = 0.727.$$

- (b) On note  $E$  l'évènement “l'expert prévoit une hausse du prix du café”. Avec l'énoncé  $P(E|A) = 0.99$  et  $P(\overline{E}|\overline{A}) = 0.88$ . La société peut améliorer son estimation de la probabilité de hausse en combinant les informations obtenues du journal et de l'expert, en supposant l'indépendance des deux sources d'information. On veut donc  $P(A|J \cap E)$ . Par le Théorème de Bayes:

$$P(A|J \cap E) = \frac{P(J \cap E|A)P(A)}{P(J \cap E|A)P(A) + P(J \cap E|\overline{A})P(\overline{A})}.$$

Avec l'indépendance conditionnelle des deux sources:

$$\begin{aligned} P(J \cap E|A) &= P(J|A)P(E|A) = 0.8 \times 0.99 = 0.792, \\ P(J \cap E|\overline{A}) &= P(J|\overline{A})P(E|\overline{A}) = 0.45 \times 0.12 = 0.054, \end{aligned}$$

d'où:

$$P(A|J \cap E) = \frac{0.792 \times 0.6}{0.792 \times 0.6 + 0.054 \times 0.4} = 0.956.$$

## Solutions des exercices du Chapitre 6

6.1 Rappel. Le nombre de façons de prendre  $k$  objets parmi  $n$  objets, sans remise et sans tenir compte de l'ordre est:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad \text{avec } 0 \leq k \leq n.$$

On utilise aussi le symbole  $C_k^n$  à la place de  $\binom{n}{k}$ .

Soit  $X$ , le nombre de garçons dans un échantillon (choisi au hasard) de 3 élèves pris parmi les 10 de la classe. D'une part, le nombre des échantillons possibles comptant 3 élèves est  $C_3^{10}$ . D'autre part, le nombre d'échantillons comprenant  $X$  garçons est  $C_X^4 C_{3-X}^6$  (avec  $0 \leq X \leq 3$ ). Donc, la probabilité d'avoir  $X = x$  garçons dans un échantillon de taille 3 est:

$$P(X = x) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}} = \frac{C_x^4 C_{3-x}^6}{C_3^{10}}.$$

Exemple:

$$\begin{aligned} P(X = 2) &= \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}} \\ &= \frac{(\text{nb de façons de tirer 2 garçons parmi 4}) \times (\text{nb de façons de tirer 1 fille parmi 6})}{\text{nombre de cas possibles}} \\ &= \frac{C_2^4 C_1^6}{C_3^{10}}. \end{aligned}$$

On obtient:

$$P(X = 0) = 0.166, \quad P(X = 1) = 0.5, \quad P(X = 2) = 0.3, \quad P(X = 3) = 0.033.$$

6.2  $X$  prend ses valeurs dans l'ensemble  $\{0, 1, 2, 3, 4, 5\}$ .

(a) En posant  $p = 0.25 =$  probabilité qu'une pomme soit avariée, on obtient:

$$\begin{aligned} P(X = 0) &= (1 - p)^5 = 0.2373047, \\ P(X = 1) &= 5p(1 - p)^4 = 0.3955078, \end{aligned}$$

Explication:

En considérant le tirage de chaque pomme comme indépendant des autres, on peut multiplier les probabilités relatives à chacune des pommes. Pour  $X = 0$ , nous avons donc un facteur  $(1 - p)$  pour chaque pomme, ce qui donne  $(1 - p)^5$ . Pour  $X = 1$ , c'est un peu plus compliqué. Appelons  $A$  l'évènement "tirer une pomme avariée". Pour obtenir un emballage avec une seule pomme avariée, on peut par exemple faire le tirage  $T_1 = \{A\bar{A}\bar{A}\bar{A}\bar{A}\}$ . La probabilité de  $T_1$  est  $P(T_1) = p(1 - p)^4$ . Mais  $T_1$  n'est pas la seule possibilité pour obtenir une seule pomme avariée: les 4 tirages suivants,  $T_2$  à  $T_5$ , donnent aussi ce résultat:

$$\begin{aligned} T_2 &= \{\bar{A}A\bar{A}\bar{A}\bar{A}\}, & T_3 &= \{\bar{A}\bar{A}A\bar{A}\bar{A}\}, \\ T_4 &= \{\bar{A}\bar{A}\bar{A}A\bar{A}\}, & T_5 &= \{\bar{A}\bar{A}\bar{A}\bar{A}A\}. \end{aligned}$$

Les probabilités de tous ces évènements sont les mêmes que celle de  $T_1$ . On a donc

$$P(X = 1) = P(T_1) + P(T_2) + P(T_3) + P(T_4) + P(T_5) = 5P(T_1) = 5p(1 - p)^4.$$

Pour les autres résultats possibles (i.e.  $P(X = x)$  avec  $x = 2, 3, 4, 5$ ), le principe est le même: il faut multiplier la probabilité de base  $p^x(1 - p)^{5-x}$  par le nombre de tirages différents qui donnent ce résultat. Et ce nombre n'est autre que  $C_x^5$ , car il s'agit du nombre de façons de placer  $x$  avariées parmi 5, donc de choisir  $x$  positions parmi 5. Les probabilités restantes sont donc:

$$P(X = 2) = C_2^5 p^2 (1 - p)^3 = 10p^2(1 - p)^3 = 0.2636719,$$

$$P(X = 3) = C_3^5 p^3 (1 - p)^2 = 10p^3(1 - p)^2 = 0.08789063,$$

$$P(X = 4) = C_4^5 p^4 (1 - p) = 5p^4(1 - p) = 0.01464844,$$

$$P(X = 5) = C_5^5 p^5 = 1p^5 = 0.0009765625.$$

Le nombre de pommes avariées suit une distribution que l'on nomme binomiale, et qui sera introduite au chapitre 7 du cours.

- (b) Le client se plaindra lorsque  $X \geq 2$ . On a  $P(X \geq 2) = 1 - P(X < 2)$  et  $P(X < 2) = P(X = 0) + P(X = 1)$ . Donc,  $P(X \geq 2) = 0.367$ .
- (c) Intuitivement, on se dit que si chaque client a une probabilité de 0.367 de se plaindre, alors sur 100 clients on peut s'attendre à ce qu'environ  $100 \times 0.367 = 36.7$  clients se plaignent. Ce nombre correspond à l'espérance mathématique du nombre de clients qui se plaignent. (Le nombre de clients qui se plaignent suit lui aussi une distribution binomiale, et vous verrez au chapitre 7 que le calcul ci-dessus correspond à une formule générale pour trouver l'espérance d'une variable binomiale.)

### 6.3

- (a) La fonction  $f$  étant une densité de probabilité, on doit avoir  $\int_{-\infty}^{+\infty} f(x)dx = 1$  et comme  $f$  est non nul seulement sur  $[-1, 1]$ , on doit avoir:

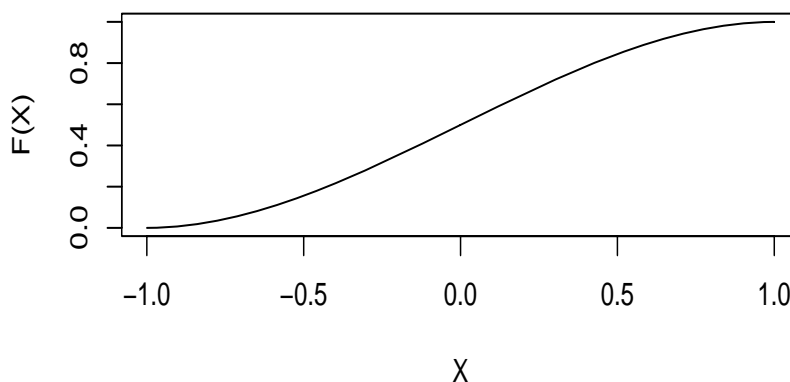
$$\int_{-1}^{+1} f(x)dx = 1, \text{ soit } \int_{-1}^{+1} c(1 - x^2)dx = 1.$$

Le calcul de l'intégrale donne  $c(x|_{-1}^1 - x^3/3|_{-1}^1) = 1$  soit  $c \times 4/3 = 1$ . On en déduit donc que  $c = 3/4$ .

- (b)  $F(x) = \int_{-\infty}^x f(x)dx = \int_{-1}^x f(x)dx = (3/4) \int_{-1}^x (1 - x^2)dx$ .

En évaluant l'intégrale, on trouve:

$$F(x) = \begin{cases} 0 & \text{si } x \leq -1, \\ (3/4)x - (1/4)x^3 + 1/2 & \text{si } -1 < x < 1, \\ 1 & \text{si } 1 \leq x. \end{cases}$$



## 6.4

$$(a) F_Y(y) = P(Y \leq y) = P(\exp(-X) \leq y) = P(-X \leq \ln y) = P(X \geq -\ln y) = \int_{-\ln y}^{+\infty} f_X(x) dx = \int_{-\ln y}^1 2x dx = 1 - (\ln y)^2, \text{ si } 0 < -\ln y < 1, \text{ i.e. si } 1/e < y < 1.$$

Dans le cas  $-\ln y \leq 0$ , soit  $1 \leq y$ , on a  $F_Y(y) = 1$ .

Pour  $1 \leq -\ln y$ , soit  $y \leq 1/e$ , on a  $F_Y(y) = 0$ .

On a donc

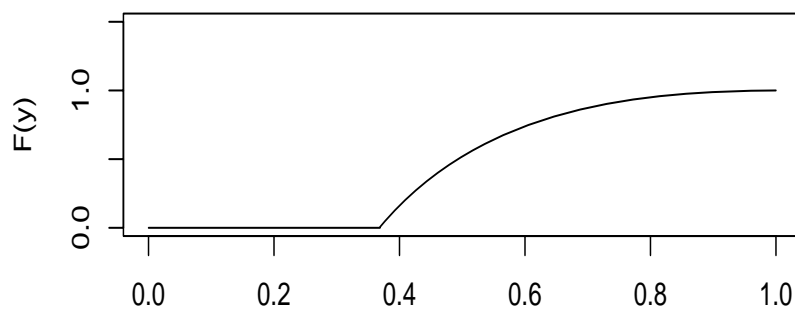
$$F_Y(y) = \begin{cases} 0 & \text{si } y \leq 1/e, \\ 1 - (\ln y)^2 & \text{si } 1/e < y < 1, \\ 1 & \text{si } y \geq 1. \end{cases}$$

Remarquons que  $0 < -\ln y < 1$  est équivalent à  $1/e < y < 1$  (dans la première expression, il suffit de prendre l'exponentielle de chaque membre, puis de prendre leur inverse (il faut alors changer le sens des inégalités), pour obtenir la deuxième).

$$(b) f_Y(y) = F'_Y(y), \text{ donc}$$

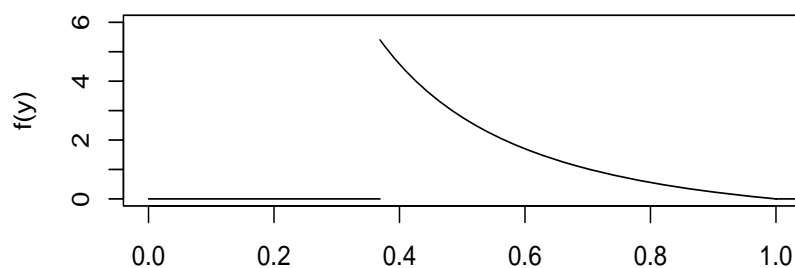
$$f_Y(y) = \begin{cases} (-2 \ln y)/y & \text{si } 1/e < y < 1, \\ 0 & \text{sinon.} \end{cases}$$

(c)



y

(Code dans R: `x0<-c(0,0.369)`  
`y0<-0*x0`  
`x1<-seq(0.369,1,by=0.001)`  
`y1<-1-(log(x1))^2`  
`plot(x0,y0,xlim=c(0,1),ylim=c(0,1.2),xlab="y",ylab="F(y)",`  
`type="l")`  
`lines(x1,y1)`)



y

```
(Code dans R: x3<-c(0,0.369)
y3<-c(0,0)
x4<-seq(0.369,1,by=0.001)
y4<--2*log(x4)/(x4)
x5<-c(1,1.05)
plot(x3,y3,xlim=c(0,1),ylim=c(0,6),xlab="y",ylab="f(y)",
type="l")
lines(x4,y4)
lines(x5,y3))
```

(d)  $E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx = \int_0^1 2x^2 dx = 2/3.$   
 $\sigma^2(X) = E(X^2) - E(X)^2 = \int_0^1 2x^3 dx - 4/9 = 1/2 - 4/9 = 1/18.$   
 $E(Y) = \int_{-\infty}^{+\infty} y \cdot f_Y(y) dy = \int_{1/e}^1 y \cdot ((-2/y) \ln y) dy = \int_{1/e}^1 (-2) \ln y dy = 2 - 4/e \simeq 0.524.$   
 $E(Y^2) = -2 \int_{1/e}^1 y \ln y dy$

En intégrant par parties, on trouve que

$$\int_{1/e}^1 x \ln x = \frac{1}{2} x^2 \ln x \Big|_{1/e}^1 - \int_{1/e}^1 \frac{1}{2} x^2 \frac{1}{x} = (3e^{-2} - 1)/4.$$

On obtient donc

$$E(Y^2) = (1 - 3e^{-2})/2 \approx 0.297, \text{ et donc}$$

$$\sigma^2(Y) = E(Y^2) - E(Y)^2 = 1/2 - 3e^{-2}/2 - (2 - 4e^{-1})^2 \approx 0.022.$$

On peut aussi utiliser R pour évaluer numériquement l'expression  $-2 \int_{1/e}^1 y \ln y dy.$

Le code est: `f<-function(y){-2*y*log(y)}`  
`integrate(f,exp(-1),1,subdivisions=10000)`

6.5 L'ensemble des cas possibles contient 36 évènements élémentaires équiprobables.

(a) Pour  $X = 1$ , le seul cas favorable est  $\{(1, 1)\}$ , donc  $P(X = 1) = 1/36$ . Pour  $X = 2$ , les cas favorables sont  $\{(1, 2), (2, 1), (2, 2)\}$ , donc  $P(X = 2) = 3/36$ . Pour  $X = 3$ , les cas favorables sont  $\{(1, 3), (3, 1), (2, 3), (3, 2), (3, 3)\}$ , donc  $P(X = 3) = 5/36$ . On procède de la même manière pour avoir  $P(X = 4) = 7/36$ ,  $P(X = 5) = 9/36$  et  $P(X = 6) = 11/36$ .

On obtient ainsi le tableau suivant pour la distribution marginale de  $X$ :

$X$	1	2	3	4	5	6
$P(X = x_i)$	1/36	3/36	5/36	7/36	9/36	11/36

A partir de là, une façon simple de trouver la distribution marginale de  $Y$  est de passer par la distribution conjointe de  $X$  et de  $Y$  (dont on a de toute façon besoin pour la question (b)). La table de distribution conjointe de  $X$  et de  $Y$  est la suivante:

$Y$	$X$	1	2	3	4	5	6
0		1/36	3/36	5/36	7/36	0	0
1		0	0	0	0	8/36	8/36
2		0	0	0	0	1/36	3/36

Pour la trouver, on a procédé comme suit:

On prend d'abord le cas  $X = 1$  (première colonne de la table) et on se demande dans combien de cas parmi ceux où  $X = 1$  on a  $Y = 0$ , dans combien de cas  $Y = 1$  et dans combien de cas  $Y = 2$ . Il n'y a qu'un seul cas où  $X = 1$ , c'est  $(1, 1)$ , et dans ce cas  $Y = 0$  (il n'y a pas de 5 ou de 6 dans le résultat  $(1, 1)$ ). On inscrit donc  $1/36$  en face de  $Y = 0$ , et zéro en face de  $Y = 1$  et  $Y = 2$ .

On prend ensuite le cas  $X = 2$ . Dans les trois cas où  $X = 2$  ( $\{(1, 2), (2, 1), (2, 2)\}$ ),  $Y = 0$ . On inscrit donc  $3/36$  en face de  $Y = 0$ , et zéro en face de  $Y = 1$  et  $Y = 2$ .

Et ainsi de suite.

Prenons encore le cas  $X = 5$ . Les cas où  $X = 5$  sont  $\{(1, 5), (2, 5), (3, 5), (4, 5), (5, 5), (5, 4), (5, 3), (5, 2), (5, 1)\}$ . Dans le cas  $(5, 5)$ ,  $Y = 2$ . Dans tous les autres cas,  $Y = 1$ . On inscrit donc  $1/36$  en face de  $Y = 2$ ,  $8/36$  en face de  $Y = 1$  et zéro en face de  $Y = 0$ .

On trouve ensuite la distribution de probabilité de  $Y$  en additionnant dans la table conjointe les probabilités se trouvant dans les lignes  $Y = 0$ ,  $Y = 1$  et  $Y = 2$ . On obtient:

$Y$	0	1	2
$P(Y = y_i)$	16/36	16/36	4/36

- (b) Pour trouver la distribution de  $Z = XY$ , on se base sur la table de distribution conjointe trouvée à la question précédente. On considère tous les cas dont la probabilité n'est pas nulle, et on calcule la valeur de  $Z$  correspondante.

Commençons par la première ligne. Dans ce cas,  $Y = 0$  et donc  $Z = XY = 0$ . La probabilité que  $Z = 0$  est donc égale à  $1/36 + 3/36 + 5/36 + 7/36$ .

Prenons la deuxième ligne. Dans ce cas,  $Y = 1$ . Les deux seuls cas qui ont une probabilité non nulle sont  $X = 5$  et  $X = 6$ . Dans le premier cas,  $Z = XY = 5$ , dans le deuxième  $Z = 6$ . Les probabilités de ces cas sont données dans la table et valent toutes deux  $8/36$ . On a donc  $P(X = 5) = P(X = 6) = 8/36$ .

En faisant de même pour la dernière ligne, on obtient la table suivante pour la distribution de  $Z$ :

$Z$	0	5	6	10	12
$P(Z = z_i)$	16/36	8/36	8/36	1/36	3/36

- (c)

$$E(X) = 1 \cdot \frac{1}{36} + 2 \cdot \frac{3}{36} + 3 \cdot \frac{5}{36} + 4 \cdot \frac{7}{36} + 5 \cdot \frac{9}{36} + 6 \cdot \frac{11}{36} = 4.47,$$

$$E(Y) = 0 \cdot \frac{16}{36} + 1 \cdot \frac{16}{36} + 2 \cdot \frac{4}{36} = 24/36 = 0.67,$$

$$E(Z) = 0 \cdot \frac{16}{36} + 5 \cdot \frac{8}{36} + 6 \cdot \frac{8}{36} + 10 \cdot \frac{1}{36} + 12 \cdot \frac{3}{36} = 3.72.$$

6.6 On suppose que  $0 \leq x \leq 1$  et  $0 \leq y \leq 1$ . Toutes les densités sont nulles en dehors de ce rectangle.

(a)

$$\begin{aligned}
 F_{XY} &= P(X \leq x, Y \leq y) = \int_0^x \int_0^y f_{XY}(u, v) du dv \\
 &= \int_0^x \int_0^y (2u + 2v - 4uv) du dv \\
 &= \int_0^x (2uv + v^2 - 2uv^2) \Big|_0^y du \\
 &= (u^2y + uy^2 - u^2y^2) \Big|_0^x = xy(x + y - xy).
 \end{aligned}$$

(b) Densité marginale de  $X$ :

$$\begin{aligned}
 f_X(x) &= \int_{-\infty}^{+\infty} f_{XY}(x, v) dv \\
 &= \int_0^1 (2x + 2v - 4xv) dv \\
 &= 2x + [v^2]_0^1 - 2x[v^2]_0^1 \\
 &= 1.
 \end{aligned}$$

Densité marginale de  $Y$ :

$$\begin{aligned}
 f_Y(y) &= \int_{-\infty}^{+\infty} f_{XY}(u, y) du \\
 &= \int_0^1 (2u + 2y - 4uy) du \\
 &= [u^2]_0^1 + 2y[u]_0^1 - y[2u^2]_0^1 \\
 &= 1.
 \end{aligned}$$

(c)  $f_{X|Y=y}(x|y) = f_{XY}(x, y)/f_Y(y) = (2x + 2y - 4xy)/1 = 2x + 2y - 4xy.$

$f_{Y|X=x}(y|x) = f_{XY}(x, y)/f_X(x) = (2x + 2y - 4xy)/1 = 2x + 2y - 4xy.$

(d)  $X$  et  $Y$  ne sont pas indépendants car  $f_{XY}(x, y) = 2x + 2y - 4xy \neq 1 \cdot 1 = f_X(x) \cdot f_Y(y).$

## Solutions des exercices du Chapitre 7

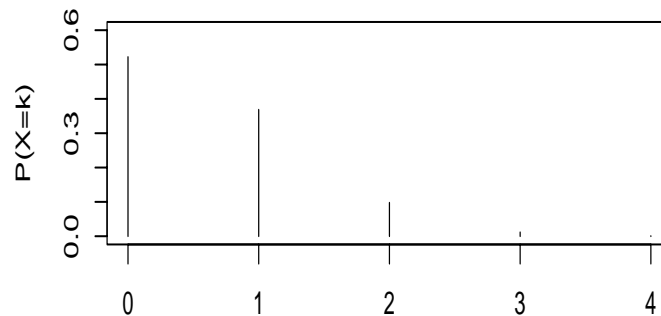
7.1 Soit  $X$  le nombre de personnes souffrant d'hypertension dans le groupe de 4 individus. On peut considérer  $X$  comme le résultat de quatre épreuves de Bernoulli,  $X$  suit donc une distribution binomiale et on écrit  $X \sim \mathcal{B}(n, p)$  avec  $n = 4$  et  $p = 0.15$ .

(a) Les probabilités cherchées sont

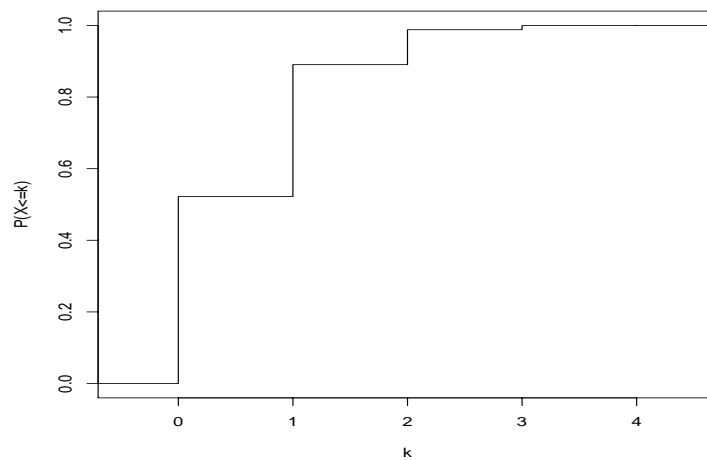
$$\begin{aligned} P(X = 0) &= \binom{4}{0} p^0 (1-p)^4, \\ P(X = 1) &= \binom{4}{1} p^1 (1-p)^3, \\ &\vdots \\ P(X = k) &= \binom{4}{k} p^k (1-p)^{4-k}. \end{aligned}$$

On trouve ainsi:  $P(X = 0) = 0.522$ ,  $P(X = 1) = 0.3684$ ,  $P(X = 2) = 0.0975$ ,  $P(X = 3) = 0.0114$  et  $P(X = 4) = 0.0005$

(b)



(c)



7.2 Supposons que la direction de la rue est Nord-Sud et qu'il y a au moins deux portails au Sud et deux au Nord du premier.

Première solution. Pour que l'ivrogne se retrouve au point de départ, il faut que la suite de ses mouvements (N = vers le Nord; S = vers le Sud) soit une des suivantes:

N	N	S	S
N	S	N	S
N	S	S	N
S	S	N	N
S	N	S	N
S	N	N	S

Comme chaque suite à probabilité  $(1/2)(1/2)(1/2)(1/2) = 1/16$ , la probabilité que l'ivrogne se retrouve au point de départ est  $6/16 = 3/8$ .

Deuxième solution. Soit  $X_i$  une variable aléatoire qui vaut 1 lorsque l'ivrogne part dans un sens après la  $i$ ème tentative et  $-1$  s'il part dans le sens contraire. On peut considérer la marche de l'ivrogne comme résultat de quatre épreuves. En posant  $S_4 = X_1 + X_2 + X_3 + X_4$ , cela revient à trouver  $P(S_4 = 0)$ . Les cas favorables sont

$\{(1, -1, 1, -1), (-1, 1, 1, -1), (1, 1, -1, -1), (-1, -1, 1, 1), (1, -1, -1, 1), (-1, 1, -1, 1)\}$ .

Avec  $P(X_i = 1) = P(X_i = -1) = 1/2$  et si on suppose l'indépendance des variables aléatoires  $X_i$ , on a  $P(S_4 = 0) = 6 \times (1/2)^4 = 0.375$ .

Troisième solution. Considérons quatre départs. Soit  $X$  la variable aléatoire qui indique, sur ces quatre départs, le nombre de fois où l'ivrogne est parti vers le Nord. Cette variable à une distribution binomiale  $\mathcal{B}(n = 4, p = 1/2)$ . Pour que l'ivrogne se retrouve au point de départ, il faut que  $X = 2$  et

$$P(X = 2) = \binom{4}{2} p^2 (1-p)^2 = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} (1/2)^4 = 3/8.$$

7.3 La variable aléatoire  $X_1$  est le nombre de succès dans  $n_1$  épreuves indépendantes de Bernoulli avec la même probabilité  $p$  de succès. La variable aléatoire  $X_2$  est le nombre de succès dans  $n_2$  épreuves indépendantes de Bernoulli avec la même probabilité  $p$  de succès. Donc,  $Y = X_1 + X_2$  est le nombre de succès dans les  $n_1 + n_2$  épreuves indépendantes de Bernoulli avec la même probabilité  $p$  de succès. On conclut que  $Y \sim \mathcal{B}(n_1 + n_2, p)$ .

7.4 Soit  $X$  la quantité mise en bouteille (en cl). Donc  $X \sim \mathcal{N}(\mu, \sigma^2 = 4)$ .

(a)  $\mu = 101$ ;

$P(X < 100) = P((X - 101)/2 < (100 - 101)/2) = P(Z < -1/2) = \Phi(-1/2)$  où  $Z = (X - 101)/2 \sim \mathcal{N}(0, 1)$ , et  $\Phi$  est la fonction de distribution cumulative de la distribution normale centrée réduite. Grâce à la symétrie de la densité normale,  $\Phi(-1/2) = 1 - \Phi(1/2)$  et, à l'aide d'une table ou du logiciel R, on trouve  $\Phi(1/2) = 0.6915$ . Donc,  $P(X < 100) = 1 - 0.6915 = 0.3081$ .

(b)  $\mu = 102$ ;

$P(X < 100) = P((X - 102)/2 < (100 - 102)/2) = \Phi(-1) = 1 - \Phi(1) = 0.1587$ .

(c)  $\mu = 103$ ;

$P(X < 100) = P((X - 103)/2 < (100 - 103)/2) = \Phi(-3/2) = 1 - \Phi(3/2) = 0.0668$ .

7.5 Soient  $X_1$ ,  $X_2$  et  $X_3$  les variables aléatoires qui représentent les quantités de potassium contenues dans les trois verres et soit  $T = X_1 + X_2 + X_3$ . On supposera que  $X_1$ ,  $X_2$  et  $X_3$  sont indépendantes. On obtient:

$$\mu(T) = \mu(X_1) + \mu(X_2) + \mu(X_3) = 21\text{mg},$$

$$\sigma^2(T) = \sigma^2(X_1) + \sigma^2(X_2) + \sigma^2(X_3) = 3 \cdot (0.4\text{mg})^2 = 0.48\text{mg}^2,$$

$$\sigma(T) = \sqrt{0.48\text{mg}^2} = 0.69282\text{mg}.$$

7.6 Soit  $A$  l'évènement " $120 \leq X \leq 200$ ",  $B$  l'évènement "l'individu est un Pygmée" et  $W$  "l'individu est un Watousi", nous avons  $P(B) = 0.4$  et  $P(W) = 0.6$ ; il s'agit de trouver  $P(A)$ . On utilise la formule de la probabilité totale:  $P(A) = P(A|B)P(B) + P(A|W)P(W)$ . On sait que:

$$P(A|B) = P(120 \leq X \leq 200) \quad \text{avec } X \sim \mathcal{N}(120, 20^2),$$

$$P(A|W) = P(120 \leq X \leq 200) \quad \text{avec } X \sim \mathcal{N}(200, 40^2).$$

Donc, à l'aide des tables ou de R,

$$P(A|B) = P((120 - 120)/20 \leq (X - 120)/20 \leq (200 - 120)/20) = 0.49997,$$

$$P(A|W) = P((120 - 200)/40 \leq (X - 200)/40 \leq (200 - 200)/40) = 0.47725,$$

d'où  $P(120 \leq X \leq 200) = 0.486337$ .

Plus généralement, soient  $Y \sim \mathcal{N}(120, 20^2)$  et  $Z \sim \mathcal{N}(200, 40^2)$  deux variables indépendantes et soient  $f_Y$  et  $f_Z$  leurs densités. Alors,  $X = Y$  avec probabilité 0.4 et  $X = Z$  avec probabilité 0.6. La fonction de distribution cumulative de  $X$  est donc

$$F_X(x) = P(X \leq x) = 0.4P(Y \leq x) + 0.6P(Z \leq x) = 0.4F_Y(x) + 0.6F_Z(x)$$

et la densité de  $X$

$$f_X(x) = F'_X(x) = 0.4f_Y(x) + 0.6f_Z(x).$$

Alors,

$$E(X) = 0.4E(Y) + 0.6E(Z) = 0.4 \times 120 + 0.6 \times 200 = 168.$$

En outre,

$$E(Y^2) = [E(Y)]^2 + \sigma^2(Y) = 120^2 + 20^2 = 14800,$$

$$E(Z^2) = [E(Z)]^2 + \sigma^2(Z) = 200^2 + 40^2 = 41600,$$

$$E(X^2) = 0.4E(Y^2) + 0.6E(Z^2) = 0.4 \times 14800 + 0.6 \times 41600 = 30880.$$

Donc

$$\sigma^2(X) = E(X^2) - [E(X)]^2 = 30880 - 168^2 = 2656.$$

Si vous êtes étonnés que  $\sigma^2(X)$  soit si grand, essayez les commandes R suivantes:

```
p <- rbinom(10000,1,0.40)
Y <- rnorm(10000,mean=120,sd=20)
Z <- rnorm(10000,mean=200,sd=40)
X <- p*Y+(1-p)*Z
mean(Y); mean(Z); mean(X)
var(Y); var(Z); var(X)
```

## Solutions des exercices du Chapitre 8

8.1 Soit  $X$  la variable aléatoire qui vaut 1 lorsqu'on a pile et 0 dans le cas contraire.  $X$  est une variable aléatoire de Bernoulli avec probabilité de succès égale à  $p$ . La fonction de vraisemblance est donc:

$$L(p) = \prod_{i=1}^{12} P(X = x_i) = p^6(1-p)^6.$$

On cherche la valeur de  $p$  qui maximise cette fonction, ou de manière équivalente, le maximum de  $\ln(L) = 6 \ln(p) + 6 \ln(1-p)$ . Le maximum est obtenu là où la dérivée est nulle, donc:

$$\frac{6}{p} - \frac{6}{(1-p)} = 0$$

d'où  $p = 1/2$ .

8.2 On a  $P(X = x_i) = p(1-p)^{x_i}$  pour  $i = 1, \dots, n$  où les  $x_i$  sont le nombres de sauts observés et  $n = 130$ .

(a) La fonction de vraisemblance est

$$L(p) = \prod_{i=1}^n P(X = x_i) = \prod_{i=1}^n p(1-p)^{x_i} = (p)^n \prod_{i=1}^n (1-p)^{x_i}$$

et le logarithme de cette fonction est

$$\log(L(p)) = n \log(p) + [\log(1-p)] \sum_{i=1}^n x_i.$$

Le maximum de cette fonction s'obtient là où la dérivée par rapport à  $p$  est nulle. L'estimateur du maximum de vraisemblance vérifie donc

$$\frac{n}{\hat{p}} - \frac{1}{1-\hat{p}} \sum_{i=1}^n x_i = 0,$$

d'où

$$\hat{p} = \frac{1}{(\sum_{i=1}^n X_i)/n + 1}.$$

(b) Avec les données on trouve  $\hat{p} = 0.263$ .

8.3 On considère un échantillon  $x_1, \dots, x_n$  de taille  $n$ . La fonction de vraisemblance s'écrit donc:

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n f(x_i) = \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp(-(x_i - \mu)^2/(2\sigma^2)), \\ &= [(2\pi\sigma^2)^{-1/2}]^n \exp(-\sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2)). \end{aligned}$$

Le logarithme de cette fonction est donné par:

$$\ln(L) = (-n/2) \ln(2\pi) - (n/2) \ln(\sigma^2) - \sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2).$$

Les dérivées partielles par rapport aux deux paramètres  $\mu$  et  $\sigma^2$  sont

$$\frac{\partial \ln(L)}{\partial \mu} = 2 \sum_{i=1}^n (x_i - \mu) / (2\sigma^2),$$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = -n / (2\sigma^2) + \sum_{i=1}^n (x_i - \mu)^2 / (2\sigma^4).$$

Le système d'équations

$$\frac{\partial \ln(L)}{\partial \mu} = 0,$$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = 0,$$

a comme solution

$$\hat{\mu} = \frac{1}{n} \sum x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \hat{\mu})^2.$$

8.4 On a  $f(x_i) = (\beta + 1)x_i^\beta$ ,  $i = 1, \dots, n$ .

(a) La vraisemblance est:

$$L(\beta) = \prod_{i=1}^n (\beta + 1)x_i^\beta = (\beta + 1)^n \prod_{i=1}^n x_i^\beta,$$

et son logarithme

$$\ln(L) = n \ln(\beta + 1) + (\beta) \sum_{i=1}^n \ln(x_i).$$

Le maximum s'obtient là où la dérivée s'annule, donc

$$\frac{n}{\hat{\beta} + 1} + \sum_{i=1}^n \ln(x_i) = 0,$$

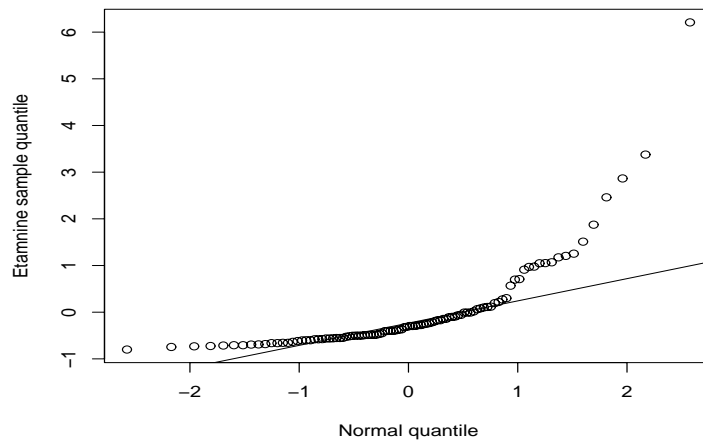
d'où:

$$\hat{\beta} = -1 - \frac{n}{\sum_{i=1}^n \ln(x_i)}.$$

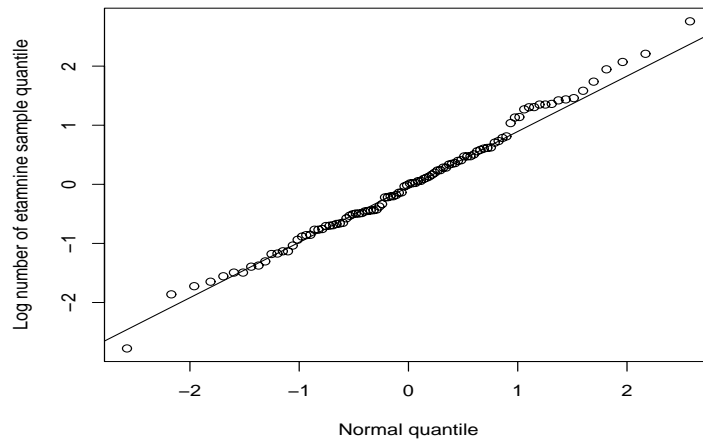
(b) Avec les données  $\hat{\beta} = 0.23$

8.5 La réponse est donné par les qq-plots: on peut donc approcher la distribution du logarithme du nombre d'étamines par la distribution normale. Ce n'est pas le cas pour les nombres d'étamines.

**Normal Q-Q Plot**



**Normal Q-Q Plot**



## Solutions des exercices du Chapitre 9

9.1 Soient  $X_1$ ,  $X_2$  et  $X_3$  les variables aléatoires qui représentent les quantités de potassium contenues dans les trois verres. On suppose que  $X_1$ ,  $X_2$  et  $X_3$  sont indépendantes et soit  $T = X_1 + X_2 + X_3$ . On trouve  $\mu(T) = 21\text{mg}$  et  $\sigma^2(T) = 0.48\text{mg}^2$  (voir solution de l'exercice 7.5 du Chapitre 7). Donc,  $P(T > 15\text{mg}) = P((T - 21\text{mg})/0.69282\text{mg} > (15 - 21)/0.69282) = P(Z > -8.66026) = 1$ , où  $Z \sim \mathcal{N}(0, 1)$ .

9.2 Soit  $X_i$  la variable aléatoire qui vaut 1 lorsque la face 6 apparaît lors du  $i$ -ème jet et 0 sinon.  $X$  n'est autre que la somme des  $X_i$ :

$$X = \sum_{i=1}^n X_i,$$

où  $n = 1200$  est le nombre de jets.

Comme les  $X_i$  sont indépendantes et identiquement distribuées (i.i.d.) et que  $n$  est grand, on peut utiliser **l'approximation normale**. On va donc considérer  $Z$ , la version centrée et réduite de  $X$ , obtenue en lui soustrayant son espérance  $\mu$  et en la divisant par son écart type  $\sigma$ :

$$Z = \frac{X - \mu}{\sigma},$$

et faire **l'approximation que  $Z$  suit une distribution  $\mathcal{N}(0, 1)$** .

On trouve  $\mu$  et  $\sigma$  en utilisant les propriétés de la distribution binomiale (p. 7.3 du polycopié). En effet, on a que

$$X \sim \mathcal{B}(n, p),$$

où  $n = 1200$  est le nombre de jets et  $p = 1/6$  est la probabilité d'obtenir un 6. On trouve

$$\mu = np = 200 \quad \text{et} \quad \sigma = \sqrt{np(1-p)} \cong 12.91.$$

On trouve la probabilité demandée en écrivant

$$\begin{aligned} P(180 \leq X \leq 220) &= P\left(\frac{180 - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{220 - \mu}{\sigma}\right) \\ &= P\left(\frac{180 - \mu}{\sigma} \leq Z \leq \frac{220 - \mu}{\sigma}\right) \\ &\cong P(-1.55 \leq Z \leq 1.55). \end{aligned}$$

En utilisant la table, on trouve

$$P(-1.55 \leq Z \leq 1.55) = 0.8788.$$

**N.B.:** Cette méthode est valable pour n'importe quelle distribution binomiale  $Y \sim \mathcal{B}(n, p)$  lorsque  $n$  est assez grand, car  $Y$  est toujours la somme de  $n$  variables i.i.d. suivant une distribution  $\mathcal{B}(1, p)$ . (Dans l'exemple ci-dessus, on a  $X_i \sim \mathcal{B}(1, 1/6)$ .)

9.3 Soit  $X$  le nombre de globules blancs par unité de volume. On peut considérer  $X$  comme la somme de 100 variables aléatoires i.i.d. avec distribution de Poisson de paramètre  $\lambda = 1$ . On peut donc utiliser l'approximation normale.

D'après les propriétés de la distribution de Poisson (p. 7.3. du polycopié),  $E(X) = 100$  et  $\sigma^2(X) = 100$ . On a donc

$$P(X \geq 90) = P((X - 100)/10 \geq (90 - 100)/10) = P(Z \geq -1) = P(Z < 1),$$

où  $Z = (X - 100)/10$  suit approximativement une distribution  $\mathcal{N}(0, 1)$ . La probabilité cherchée est donc de 0.8413 (table).

9.4 Comme  $T$  est la somme d'un grand nombre de variables i.i.d., on peut utiliser l'approximation normale.

Soit  $X_i$  la durée de vie de l'ampoule  $i$  et soient  $\mu = E(X_i) = 1000$  et  $\sigma^2 = \sigma^2(X_i) = 1000^2$ . Comme  $T = \sum_{i=1}^N X_i$ , on a que  $E(T) = N\mu$  et  $\sigma^2(T) = N\sigma^2$ .

(a)

$$P(T > 115000) = P\left(Z > \frac{115000 - N\mu}{\sqrt{N}\sigma}\right) = P(Z > 1.5) = 1 - P(Z \leq 1.5),$$

où  $Z = (T - N\mu)/(\sqrt{N}\sigma)$  suit approximativement une distribution  $\mathcal{N}(0, 1)$ . La probabilité cherchée est donc de 0.0668 (table).

(b) On veut  $N$  tel que  $P(T > 50000) \geq 0.95$ , ce qui est équivalent à  $P(T \leq 50000) \leq 0.05$ . Or,

$$P(T \leq 50000) = P\left(Z \leq \frac{50000 - N\mu}{\sqrt{N}\sigma}\right),$$

où  $Z$  est défini comme au point (a). En faisant l'approximation  $Z \sim \mathcal{N}(0, 1)$  et en définissant  $z_{0.05}$  comme le quantile 0.05 de la distribution  $\mathcal{N}(0, 1)$ , on obtient que

$$P\left(Z \leq \frac{50000 - N\mu}{\sqrt{N}\sigma}\right) \leq 0.05 \iff \frac{50000 - N\mu}{\sqrt{N}\sigma} \leq z_{0.05}$$

(faire un dessin pour s'en convaincre).

Nous allons donc résoudre l'équation suivante:

$$50000 - N\mu - \sqrt{N}\sigma z_{0.05} = 0.$$

En posant  $y = \sqrt{N}$ , la variable auxiliaire  $y$  vérifie l'équation du second degré:

$$y^2\mu + y\sigma z_{0.05} - 50000 = 0.$$

A l'aide de la table, on trouve que  $z_{0.05} = -1.64$ . En remplaçant les différents paramètres par leur valeur, on obtient l'équation

$$y^2 - 1.64y - 50 = 0$$

dont la solution positive vaut 7.94. Comme  $N = y^2$ , la solution pour  $N$  vaut 63.04. La valeur minimale de  $N$  qui satisfait  $P(T > 50000)$  est donc  $N_{min} = 64$ .

- (c) On veut  $P(T > t) \geq 0.95$  ou, ce qui est équivalent,  $P(T \leq t) \leq 0.05$  pour  $N = 100$ .  
On a :

$$P(T \leq t) = P\left(\frac{T - N\mu}{\sqrt{N}\sigma} \leq \frac{t - N\mu}{\sqrt{N}\sigma}\right) = P\left(Z \leq \frac{t - N\mu}{\sqrt{N}\sigma}\right),$$

où  $Z = (T - N\mu)/(\sqrt{N}\sigma) \sim \mathcal{N}(0, 1)$  (approximativement). Comme au point (b), on a l'équivalence

$$P\left(Z \leq \frac{t - N\mu}{\sqrt{N}\sigma}\right) \leq 0.05 \iff \frac{t - N\mu}{\sqrt{N}\sigma} \leq z_{0.05},$$

d'où

$$t \leq \sqrt{N}\sigma z_{0.05} + N\mu.$$

En remplaçant, on obtient  $t \leq 83600$ .

9.5 On dit que  $X$  suit une distribution uniforme discrète sur  $[a, b]$ , si les valeurs possibles de  $X$  sont  $\{a, a+1, \dots, b\}$  et  $P(X = x) = 1/(b-a+1)$ . On démontre que  $E(X) = (b+a)/2$  et  $\sigma^2(X) = (b-a)(b-a+2)/12$ . Dans le cas du problème traité, les 16 nombres aléatoires forment des variables aléatoires uniformes i.i.d. de moyenne  $E(X_i) = \mu = 9/2$  et variance  $\sigma^2(X_i) = \sigma^2 = 33/4$ . En utilisant le théorème limite centrale:

$$P(4 < \bar{X} < 6) = P\left(\frac{4 - \mu}{\sigma/\sqrt{n}} < Z < \frac{6 - \mu}{\sigma/\sqrt{n}}\right) = P(-0.6963 < Z < 2.0889) = 0.7385$$

où  $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0, 1)$ .

9.6 On rappelle que si  $Y \sim \mathcal{U}(0, 1)$  alors  $P(Y \leq a) = a$  pour  $0 \leq a \leq 1$ .

- (a) On suppose  $Y \sim \mathcal{U}(0, 1)$ , donc:

$$P(X \leq x) = P(F^{-1}(Y) < x) = P(Y \leq F(x)) = F(x).$$

- (b) Si  $F(x) = F_n(x)$  (la distribution cumulative empirique)

$$P(X \leq x) = F_n(x) = (\text{nombre de } x_i \leq x)/n.$$

En tirant avec remise  $n$  éléments de l'ensemble  $\{x_1, \dots, x_n\}$ , la probabilité de l'évènement "nombre de  $x_i \leq x$ " se calcule facilement en évaluant le nombre de valeurs inférieures à  $x$  et en la divisant par le nombre de cas possibles qui est  $n$ .

9.7

- (a) On sait que  $E(X_i) = \lambda$ , donc:

$$E(\bar{X}) = E((X_1 + \dots + X_n)/n) = (1/n)(E(X_1) + \dots + E(X_n)) = (1/n)(n\lambda) = \lambda.$$

- (b) On rappelle que pour une variable aléatoire  $X$ ,  $\sigma^2(X) = E(X^2) - [E(X)]^2$  et que pour une distribution de Poisson  $E(X) = \sigma^2(X) = \lambda$ .

Pour  $n = 2$ ,

$$E(\bar{X}^2) = E((1/4)(X_1^2 + 2X_1X_2 + X_2^2)) = (1/4)(E(X_1^2) + 2E(X_1)E(X_2) + E(X_2^2))$$

donc:

$$E(\bar{X}^2) = \lambda^2 + \lambda/2 \neq \lambda^2.$$

## Solutions des exercices du Chapitre 10

10.1 Définissons la statistique  $T$  comme le nombre de réponses correctes aux 20 questions. Donc  $T = \sum_{i=1}^{20} X_i$ , où  $X_i$  vaut 1 si la réponse à la question  $i$  est juste et 0 sinon.  $T$  suit une distribution binomiale  $\mathcal{B}(n = 20, p)$ . Considérons l'hypothèse nulle  $H_0$  : “le candidat devine les réponses”; en d'autres termes  $H_0 : p = 1/4$ . Considérons aussi l'alternative  $H_1$  : “le candidat a des connaissances”, c'est à dire  $H_1 : p > 1/4$ .

(a) La probabilité cherchée est donc:

$$\begin{aligned} P\left(\sum_{i=1}^{20} X_i \geq 9\right) &= 1 - P\left(\sum_{i=1}^{20} X_i \leq 8\right) \\ &= 1 - \sum_{k=0}^8 \binom{20}{k} (1/4)^k (3/4)^{20-k} = 1 - 0.9591 = 0.05. \end{aligned}$$

Pour calculer la quantité  $\sum_{k=0}^8 \binom{20}{k} (1/4)^k (3/4)^{20-k}$ , on peut utiliser R et taper: `pbinom(8,20,0.25)`.

(b) Un test de niveau 5% est défini par la règle: “rejeter  $H_0$  lorsque  $T > 9$ ”.

(c) Il s'agit d'une erreur de type I.

10.2 On dit qu'une variable aléatoire  $X$  suit une distribution géométrique de paramètre  $p$  si  $P(X = i) = p(1 - p)^i$ ,  $i = 0, 1, \dots$ . A l'aide de la formule de la série géométrique,

$$1 + \rho + \rho^2 + \dots + \rho^k = \frac{1 - \rho^{k+1}}{1 - \rho},$$

on trouve

$$P(X \leq k) = p \sum_{i=0}^k (1 - p)^i = p \frac{1 - (1 - p)^{k+1}}{p} = 1 - (1 - p)^{k+1},$$

et donc,

$$P(X > k) = (1 - p)^{k+1} = P(X \geq k + 1).$$

La variable  $X$  peut être considérée comme le nombre d'essais jusqu'au premier succès dans  $i$  épreuves de Bernoulli. Si  $X$  est l'âge de l'individu, la probabilité qu'il dépasse 20 ans est donnée par la probabilité qu'il ne meure pas jusqu'à l'âge de 20 ans, soit  $(1 - p)^{20}$ ,  $p$  étant la probabilité de disparaître au cours d'une année. Soit  $X_1$  l'âge du premier et  $X_2$  celui du deuxième individu échantillonné. Alors, si  $p = 0.1$ , la probabilité que les deux individus sont âgés de plus de 20 ans est

$$P(X_1 \geq 20 \cap X_2 \geq 20) = P(X_1 \geq 20)P(X_2 \geq 20) = (1 - p)^{40} = 0.01478,$$

Cette probabilité est très petite ( $< 5\%$ ). On peut donc rejeter l'hypothèse nulle  $H_0 : p = 0.1$ . Des valeurs de  $p$  plausibles sont, par exemple, les valeurs telles que  $(1 - p)^{40} \geq 0.05$ , c'est à dire  $p < 0.0722$ .

## 10.3

- (a) La probabilité d'erreur de type I est  $P_0(\bar{X} \geq 7)$ , où  $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$ . On sait que, sous  $H_0$ ,  $\bar{X}$  suit une distribution normale de moyenne  $\mu = 6$  et variance  $\sigma^2 = 1/4$ , donc:

$$P_0(\bar{X} \geq 7) = P_0((\bar{X} - 6)/\sigma \geq (7 - 6)/\sigma) = P(Z \geq 2) = 0.0227,$$

où  $Z = (\bar{X} - 6)/\sigma \sim \mathcal{N}(0, 1)$ .

- (b) La probabilité d'erreur de type II est  $P_1(\bar{X} \leq 7)$ . On sait que, sous  $H_1$ ,  $\bar{X}$  suit une distribution normale de moyenne  $\mu = 7$  et variance  $\sigma^2 = 1/4$ , donc:

$$P_1(\bar{X} \leq 7) = P_1((\bar{X} - 7)/\sigma \leq (7 - 7)/\sigma) = P(Z \leq 0) = 0.5,$$

où  $Z = (\bar{X} - 7)/\sigma \sim \mathcal{N}(0, 1)$ .

10.4 La valeur observée de la statistique de test est

$$z_0 = ((\bar{x} - \mu)/(\sigma/\sqrt{n})) = (147.4 - 160)/(6/\sqrt{31}) = -11.69.$$

Pour réaliser un test bilatéral au niveau 1%, nous considérons les percentiles  $z_{0.995} = 2.575$  et  $z_{0.005} = -2.575$ . Comme  $z_0 = -11.69$  n'est pas dans l'intervalle  $(-2.575, 2.575)$ , l'hypothèse nulle  $H_0 : \mu = 160\text{cm}$  est rejetée.

10.5 Calculons la p-value du test de  $H_0 : p = 0.4$  contre  $H_1 : p < 0.4$ :

$$P_0(K \leq 8) = \sum_{i=0}^8 P(K = i) \approx 0.585.$$

On s'intéresse à  $P(K \leq 8)$ , car des valeurs très petites de  $K$  suggèrent de rejeter  $H_0$  en faveur de  $H_1$ . Comme 0.585 n'est pas une petite probabilité (largement supérieure à 5%), nous acceptons  $H_0$ . (0.585 est la probabilité, si  $H_0$  est vraie, d'observer 8 ou encore moins de fumeurs dans un échantillon de 20.)

Au passage, il était évident que nous ne pouvions pas rejeter  $H_0$ , puisque la proportion de fumeurs observée correspond à exactement 40% ( $8 = 40\%$  de 20). On constate que le test confirme cela.

## Solutions des exercices du Chapitre 11

### 11.1

- (a) Soit  $p$  la probabilité que la tartine tombe sur la face avec la confiture.

$$H_0: p = 0.5,$$

$$H_1: p > 0.5 \text{ (loi de Murphy).}$$

Le test est unilatéral.

- (b)  $\hat{p} = 540/1000 = 0.54,$

$$z = (\hat{p} - p_0) / \sqrt{p_0 q_0 / n} = (0.54 - 0.50) / \sqrt{0.5 \times 0.5 / 1000}, = 2.52982.$$

$$z_{1-\alpha} = z_{0.95} = 1.645.$$

Il faut donc rejeter  $H_0$  et accepter cette loi de Murphy ( $H_1$ ) au niveau de 5%.

- (c)  $p_i = \hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} = 0.50911,$

$$p_s = \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} = 0.57089.$$

### 11.2

- (a) Soit

$p_1$  = probabilité que la tartine tombe sur la face tartinée sur le terrain de basket,

$p_2$  = probabilité que la tartine tombe sur la face tartinée sur le tapis de Perse.

$$H_0: p_1 = p_2,$$

$$H_1: p_1 < p_2 \text{ (loi de Murphy).}$$

Le test est unilatéral.

- (b)  $\hat{p}_1 = 220/400, \hat{p}_2 = 350/600,$

$$\hat{p} = (220 + 350) / 1000 = 570 / 1000,$$

$$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)} = -1.04307,$$

$$z_{1-\alpha} = z_{0.95} = 1.645$$

Comme  $-1.04307 > -1.645$ , il faut accepter  $H_0$  et rejeter cette loi de Murphy.

### 11.3

- (a) Soit  $p$  la proportion de lecteurs qui lisent les annonces publicitaires.

$$H_0: p = 0.5,$$

$$H_1: p \neq 0.5.$$

Le test est bilatéral.

$$\hat{p} = 40/100,$$

$$z = (\hat{p} - p_0) / \sqrt{p_0 q_0 / n} = (0.4 - 0.50) / \sqrt{0.5 \times 0.5 / 100} = -2,$$

$$z_{\alpha/2} = z_{0.005} = -2.57 \text{ et } z_{1-\alpha/2} = z_{0.995} = 2.57$$

Comme  $-2.57 < -2 < 2.57$ , il faut accepter  $H_0$ .

- (b)  $p_i = \hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} = 0.274$

$$p_s = \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} = 0.526.$$

11.4

On a

$H_0$  : La présence ou l'absence d'une névrose est indépendante du mode de vie

$H_1$  : La présence ou l'absence d'une névrose n'est pas indépendante du mode de vie.

On a donc un test bilatéral.

On calcule la statistique

$$z^2 = \frac{200 \cdot (40 \cdot 60 - 100 \cdot 60)^2}{140 \cdot 120 \cdot 100 \cdot 160} = 12.53.$$

Comme  $12.53 > 6.63 = \chi_{1,0.99}^2$  (voir table de la distribution  $\chi^2$  à 1 degré de liberté), l'hypothèse de non association  $H_0$  peut être rejetée au niveau 1%.

N.B.: Cette formule n'est utilisable que pour un test bilatéral. En effet, la statistique  $z^2$  est une mesure du **carré** de la différence de proportion de névrosés parmi les gens qui vivent seuls et ceux qui vivent en famille. Le signe de cette différence n'apparaît donc pas. Si on posait la question: "Est-ce que la proportion de névrosés est **plus grande** chez les gens qui vivent seuls que chez ceux qui vivent en famille?", il faudrait faire un test unilatéral, comme dans l'exercice 2. On devrait alors utiliser la statistique

$$z = \frac{(100/160 - 40/100)}{\sqrt{p \cdot (1 - p) \cdot (1/100 + 1/160)}}$$

avec  $p = (40 + 100)/(100 + 160)$ .

On obtient  $z = 3.54$ . Comme  $3.54 > 2.326 = z_{0.99}$  (voir table de la distribution de Gauss), on rejette là aussi l'hypothèse  $H_0$  au niveau 1% et on admet l'hypothèse alternative

$H'_1$  : La proportion de névrosés est supérieure chez les gens qui vivent seuls.

Question subsidiaire: on peut aussi tester  $H_0$  contre  $H_1$  (test **bilatéral**, premier cas ci-dessus) en utilisant  $z$  au lieu de  $z^2$ . Comment faut-il faire?

## Solution des exercices du Chapitre 12

12.1 (a) Nous utilisons la statistique du test de Student pour un seul échantillon:

$$T = (\bar{X} - \mu)/(S/\sqrt{n}).$$

(b) On rejette l'hypothèse  $H_0 : \mu = \mu_0$  ( $\mu_0$  une valeur donnée) si  $T \leq t_{n-1, \alpha/2}$  ou si  $T \geq t_{n-1, 1-\alpha/2}$ .

(c) Pour  $\alpha = 0.01$  et  $n = 16$  on a  $t_{15, 0.005} = -2.94$ ,  $t_{15, 0.995} = 2.94$  et avec les données on obtient  $T = (1590 - 1600)/(30/4) = -1.33$ . Comme  $-2.94 \leq T \leq 2.94$ , on accepte  $H_0$ .

12.2 Nous considérons le test de Student pour données non appariées. Nous avons  $m = 25$ ,  $\bar{X} = 82$ ,  $S_x = 8$  et  $n = 16$ ,  $\bar{Y} = 78$ ,  $S_y = 7$ . Comme les valeurs de  $S_x$  et  $S_y$  sont proches, nous admettons la condition que les deux populations (de notes) ont la même variance. On veut tester  $H_0 : \mu_1 = \mu_2$  contre  $H_1 : \mu_1 \geq \mu_2$  au niveau  $\alpha = 0.05$ . La statistique de test est

$$T = \frac{\bar{X} - \bar{Y}}{S_D},$$

avec

$$S_D^2 = (1/m + 1/n) \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}.$$

On rejette  $H_0$  si  $T \geq t_{m+n-2, 1-\alpha}$ . Pour  $\alpha = 0.05$ ,  $m = 25$ ,  $n = 16$  on a  $t_{39, 0.95} = 1.68$  et avec les données on obtient  $T = 1.63$ . Comme  $T \leq 1.68$ , on accepte  $H_0$ .

12.3 On veut tester  $H_0 : \mu = 39.6$  contre  $H_1 : \mu \neq 39.6$ . Nous utilisons la statistique

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

On rejette si  $T \leq t_{n-1, \alpha/2}$  ou si  $T \geq t_{n-1, 1-\alpha/2}$ . L'écart type et la moyenne des données sont  $S = 0.7631$  et  $\bar{X} = 39.485$ , d'où

$$T = \frac{39.485 - 39.6}{0.7631/\sqrt{20}} = -0.674.$$

On a  $t_{19, 0.025} = -2.093$  et  $t_{19, 0.975} = 2.093$  et  $-2.093 \leq T \leq 2.093$ . On accepte donc  $H_0$ .

12.4 On teste  $H_0 : \mu = 19.7$  contre  $H_1 : \mu \neq 19.7$ . Avec le test de Student on obtient  $T = 1.636$ ,  $t_{4, 0.025} = -2.776$  et  $t_{4, 0.975} = 2.776$ . On accepte donc  $H_0$ .

12.5 Les données sont appariées. On calcule donc les différences entre les poids avant et après le programme.

patients	:	1	2	3	4	5	6	7	8	9
différence	:	8	-2	8	-2	8	0	-3	-4	5

On note  $X_i, i = 1, \dots, n$  la série des différences ( $n = 9$ ). Les hypothèses à tester sont  $H_0 : \mu = 0$  contre  $H_1 : \mu > 0$ , où  $\mu$  représente la moyenne de la population des différences. On utilise la statistique

$$T = \frac{\bar{X} - 0}{S/\sqrt{n}} = \frac{2 - 0}{1.724} = 1.16.$$

La valeur critique est  $t_{n-1, 1-\alpha} = t_{8, 0.99} = 2.896$  et, comme  $T = 1.16 < 2.896$ ,  $H_0$  est accepté. Cela signifie que le programme diététique doit être considéré comme inefficace.