Université Joseph Fourier, Grenoble I
Licence Sciences et Technologies 2e année
STA230 : Méthodes Statistiques pour la Biologie

Exercices de Statistique

http ://ljk.imag.fr/membres/Bernard.Ycart/STA230/

Chaque thème commence par un rappel de cours et un exercice corrigé. Les calculs ont été effectués en utilisant un logiciel ; à cause des erreurs d'arrondis, il peut y avoir des différences mineures avec les calculs effectués à partir des tables de valeurs statistiques.

Table des matières

1 Données et Modèles

1.1 Distributions empiriques

Soit $(x_1, \ldots, x_n)$ un échantillon, c'est-à-dire les valeurs numériques prises par un même caractère sur un ensemble de n individus.

- Les *modalités* sont les valeurs prises.
- La *moyenne empirique* est $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.
- La *variance empirique* est $s_x^2 = \left(\dfrac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$.
- L'*écart-type empirique* est la racine carrée de la variance empirique.
- Un échantillon *centré et réduit* a pour moyenne 0 et pour variance 1. Pour *centrer et réduire* un échantillon, on retranche la moyenne à toutes les modalités, puis on les divise par l'écart-type.
- La *fréquence empirique* d'un intervalle est le rapport du nombre de valeurs prises dans cet intervalle, au nombre total d'individus.
- La *médiane* est la plus petite modalité telle qu'au moins 50% des valeurs prises soient inférieures.
- Le *premier quartile* est la plus petite modalité telle qu'au moins 25% des valeurs prises soient inférieures.
- Le *dernier quartile* est la plus petite modalité telle qu'au moins 75% des valeurs prises soient inférieures.
- On considère qu'un caractère est *continu* quand toutes les valeurs prises sont distinctes ou presque. Quand pour la plupart des modalités plusieurs individus ont la même valeur, le caractère est *discret*.

Exercice 1.1.1. On donne les effectifs par âge, de mères non fumeuses à l'accouchement.

âge	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
effectif	7	8	9	10	12	3	2	5	4	5	2	4	2	0	1

1. Quelles sont les modalités ?

   Les modalités sont les entiers de 21 *à* 35.

2. S'agit-il d'un caractère discret ou continu ?

   Compte tenu de la précision des données, plusieurs individus prennent la même modalité (sont considérés comme ayant le même âge). Il s'agit donc d'un caractère discret.

3. Calculer les fréquences empiriques des modalités.

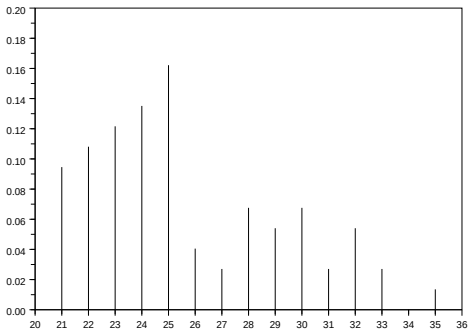
   Pour obtenir les fréquences empiriques, on divise les effectifs par le nombre total d'individus, soit ici 74.

âge	21	22	23	24	25	26	27
fréquence	$\frac{7}{74}$	$\frac{8}{74}$	$\frac{9}{74}$	$\frac{10}{74}$	$\frac{12}{74}$	$\frac{3}{74}$	$\frac{2}{74}$
val. arrondie	0.095	0.108	0.122	0.135	0.162	0.041	0.027

28	29	30	31	32	33	34	35
$\frac{5}{74}$	$\frac{4}{74}$	$\frac{5}{74}$	$\frac{2}{74}$	$\frac{4}{74}$	$\frac{2}{74}$	$\frac{0}{74}$	$\frac{1}{74}$
0.068	0.054	0.068	0.027	0.054	0.027	0	0.014

4. Représenter les fréquences empiriques sur un diagramme en bâtons.

   Le diagramme en bâtons consiste à tracer un segment vertical au-dessus de chaque modalité, de longueur proportionnelle à l'effectif ou à la fréquence empirique.



5. Calculer la moyenne, la variance et l'écart-type empiriques de l'échantillon.

   Pour calculer la moyenne empirique on effectue l'opération :

$$\overline{x} = \frac{1}{74}\Big(7\times 21 + 8\times 22 + \cdots + 0\times 34 + 1\times 35\Big) = 25.662\ .$$

   L'âge moyen dans cet échantillon est de 25 *ans et* 8 *mois environ.*
   Pour calculer la variance empirique on effectue l'opération :

$$s_x^2 = \frac{1}{74}\Big(7\times 21^2 + 8\times 22^2 + \cdots + 0\times 34^2 + 1\times 35^2\Big) - (25.662)^2 = 12.683\ .$$

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

$$s_x = \sqrt{12.683} = 3.561\ ,$$

   soit environ 3 *ans et* 7 *mois.*

6. Calculer les valeurs de la fonction de répartition empirique.

   Les valeurs de la fonction de répartition empirique sont les fréquences cumulées.

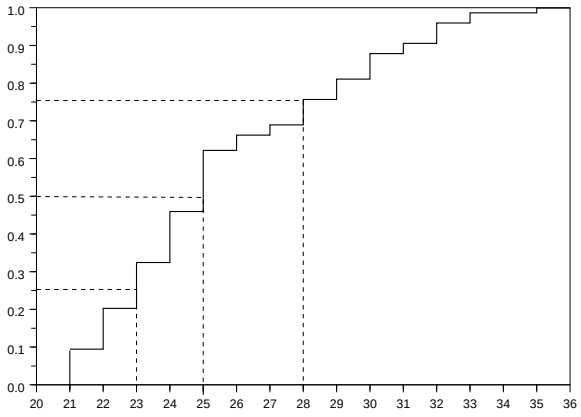
âge	21	22	23	24	25	26	27
fréq. cum.	$\frac{7}{74}$	$\frac{15}{74}$	$\frac{24}{74}$	$\frac{34}{74}$	$\frac{46}{74}$	$\frac{49}{74}$	$\frac{51}{74}$
val. arrondie	0.095	0.203	0.324	0.459	0.622	0.662	0.689

28	29	30	31	32	33	34	35
$\frac{56}{74}$	$\frac{60}{74}$	$\frac{65}{74}$	$\frac{67}{74}$	$\frac{71}{74}$	$\frac{73}{74}$	$\frac{73}{74}$	$\frac{74}{74}$
0.757	0.811	0.878	0.905	0.959	0.986	0.986	1

7. Quelle est la fréquence empirique de l'intervalle $[22\ ;\ 25]$?

   C'est la somme des fréquences empiriques des modalités 22, 23, 24, 25, *ou bien la différence de valeurs de la fonction de répartition empirique* $F(25) - F(21)$, *soit* $39/74 \simeq 0.527$. *Plus de la moitié des femmes de l'échantillon sont âgées de* 22 *à* 25 *ans.*

8. Représenter graphiquement la fonction de répartition empirique. Déterminer graphiquement la médiane et les quartiles de l'échantillon.



   La médiane est 25 *ans ; le premier quartile est* 23 *ans, le dernier quartile est* 28 *ans.*

9. Comparer d'une part la moyenne avec la médiane, d'autre part l'écart-type avec les distances entre la médiane et les quartiles.

   La moyenne est supérieure à la médiane, ce qui est normal pour une distribution qui est étirée vers la droite. Pour la même raison, l'écart entre le dernier quartile et la médiane est supérieur à l'écart entre la médiane et le premier quartile. Les deux sont inférieurs à l'écart-type : c'est le cas pour la plupart des distributions, qu'elles soient symétriques ou non.

Exercice 1.1.2. On donne les effectifs par âge, de mères fumeuses à l'accouchement.

âge	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
effectif	5	5	4	3	3	5	1	4	3	2	3	2	1	1	1

1. Quelles sont les modalités ?
2. S'agit-il d'un caractère discret ou continu ?

3. Calculer les fréquences empiriques des modalités.
4. Représenter les fréquences empiriques sur un diagramme en bâtons.
5. Calculer la moyenne, la variance et l'écart-type empiriques de l'échantillon.
6. Calculer les valeurs de la fonction de répartition empirique.
7. Quelle est la fréquence empirique de l'intervalle $[22\ ;\ 25]$?
8. Représenter graphiquement la fonction de répartition empirique. Déterminer la médiane et les quartiles de l'échantillon.
9. Comparer d'une part la moyenne avec la médiane, d'autre part l'écart-type avec les distances entre la médiane et les quartiles.

Exercice 1.1.3. On considère l'échantillon statistique $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. Quelle est sa moyenne empirique ?
2. Quelle est sa variance empirique ?
3. Quel échantillon centré et réduit peut-on lui associer ?
4. Si vous deviez proposer un modèle pour ces données : choisiriez-vous un modèle discret ou un modèle continu ?

Exercice 1.1.4. On considère l'échantillon statistique

$$(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4)\ .$$

1. Quelle est sa moyenne empirique ?
2. Quelle est sa variance empirique ?
3. Quel échantillon centré et réduit peut-on lui associer ?
4. Si vous deviez proposer un modèle pour ces données : choisiriez-vous un modèle discret ou un modèle continu ?

1.2 Probabilités et probabilités conditionnelles

- La *probabilité d'un événement* dans une population est la proportion des individus pour lesquels l'événement est réalisé.
- La *probabilité conditionnelle de A sachant B* est la proportion d'individus pour lesquels A est réalisé *parmi ceux pour lesquels B l'est aussi*. C'est le rapport de la probabilité de "A et B" à la probabilité de B :

$$\mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \text{ et } B]}{\mathbb{P}[B]}\ .$$

- La *formule des probabilités totales* donne la probabilité d'un événement A en fonction des probabilités conditionnelles sachant un autre événement B et son contraire $\overline{B}$:

$$\mathbb{P}[A] = \mathbb{P}[A \,|\, B]\,\mathbb{P}[B] + \mathbb{P}[A \,|\, \overline{B}]\,\mathbb{P}[\overline{B}]\ .$$

- La *formule de Bayes* permet d'échanger l'ordre des probabilités conditionnelles :

$$\mathbb{P}[B \,|\, A] = \frac{\mathbb{P}[A \,|\, B]\,\mathbb{P}[B]}{\mathbb{P}[A \,|\, B]\,\mathbb{P}[B] + \mathbb{P}[A \,|\, \overline{B}]\,\mathbb{P}[\overline{B}]} \; .$$

Exercice 1.2.1. Dans un élevage de moutons, on estime que 30% sont atteints par une certaine maladie. On dispose d'un test pour cette maladie. Si un mouton n'est pas atteint, il a 9 chances sur 10 d'avoir une réaction négative au test ; s'il est atteint, il a 8 chances sur 10 d'avoir une réaction positive. On soumet tous les moutons de l'élevage au test.

Pour tout l'exercice, on note M l'événement "le mouton est malade" et T l'événement "le mouton a une réaction positive au test. L'énoncé donne :

$$\mathbb{P}[M] = 0.3 \; , \quad \mathbb{P}[\overline{T} \,|\, \overline{M}] = 0.9 \; , \quad \mathbb{P}[T \,|\, M] = 0.8 \; .$$

1. Quelle est la probabilité qu'un mouton de cet élevage ne soit pas malade ?

$$\mathbb{P}[\overline{M}] = 1 - \mathbb{P}[M] = 1 - 0.3 = 0.7 \; .$$

2. Quelle est la probabilité conditionnelle qu'un mouton ait une réaction positive au test sachant qu'il n'est pas malade ?

$$\mathbb{P}[T \,|\, \overline{M}] = 1 - \mathbb{P}[\overline{T} \,|\, \overline{M}] = 1 - 0.9 = 0.1 \; .$$

3. Quelle est la probabilité qu'un mouton ne soit pas malade et ait une réaction positive au test ?

$$\mathbb{P}[T \text{ et } \overline{M}] = \mathbb{P}[T \,|\, \overline{M}]\,\mathbb{P}[\overline{M}] = 0.1 \times 0.7 = 0.07 \; .$$

4. Quelle proportion des moutons de l'élevage réagit positivement au test ?

   On peut utiliser la formule des probabilités totales ou raisonner directement, en distinguant, parmi les moutons ayant réagi positivement, ceux qui sont malades de ceux qui ne le sont pas.

$$\begin{aligned} \mathbb{P}[T] &= \mathbb{P}[T \text{ et } M] + \mathbb{P}[T \text{ et } \overline{M}] \\ &= \mathbb{P}[T \,|\, M]\,\mathbb{P}[M] + \mathbb{P}[T \,|\, \overline{M}]\,\mathbb{P}[\overline{M}] \\ &= 0.8 \times 0.3 + 0.1 \times 0.7 = 0.24 + 0.07 = 0.31 \; . \end{aligned}$$

5. Quelle est la probabilité qu'un mouton soit malade, sachant qu'il a réagi positivement ?

On peut utiliser directement la formule de Bayes ou bien la retrouver comme suit.

$$\begin{aligned}\mathbb{P}[M \,|\, T] &= \frac{\mathbb{P}[T \text{ et } M]}{\mathbb{P}[T]} \\ &= \frac{\mathbb{P}[T \,|\, M]\,\mathbb{P}[M]}{\mathbb{P}[T \,|\, M]\,\mathbb{P}[M] + \mathbb{P}[T \,|\, \overline{M}]\,\mathbb{P}[\overline{M}]} \\ &= \frac{0.8 \times 0.3}{0.8 \times 0.3 + 0.1 \times 0.7} \simeq 0.774\ .\end{aligned}$$

6. Quelle est la probabilité qu'un mouton ne soit pas malade, sachant qu'il a réagi négativement ?

   On peut utiliser directement la formule de Bayes ou bien la retrouver comme suit.

$$\begin{aligned}\mathbb{P}[\overline{M} \,|\, \overline{T}] &= \frac{\mathbb{P}[\overline{T} \text{ et } \overline{M}]}{\mathbb{P}[\overline{T}]} \\ &= \frac{\mathbb{P}[\overline{T} \,|\, \overline{M}]\,\mathbb{P}[\overline{M}]}{\mathbb{P}[\overline{T} \,|\, \overline{M}]\,\mathbb{P}[\overline{M}] + \mathbb{P}[\overline{T} \,|\, M]\,\mathbb{P}[M]} \\ &= \frac{0.9 \times 0.7}{0.9 \times 0.7 + 0.2 \times 0.3} \simeq 0.913\ .\end{aligned}$$

Exercice 1.2.2. Une plante comporte 3 espèces, hâtive, normale ou tardive. On sait que la plante peut être soit naine, soit grande. Dans un lot de plantes issues de 1000 graines, on a dénombré 600 naines, 200 tardives, 300 hâtives naines, 250 normales grandes, 100 tardives grandes. On considère la plante issue d'une graine choisie au hasard.

1. Quelle est la probabilité qu'elle soit hâtive ? normale ? tardive ? naine ? grande ?
2. On observe une plante naine. Quelle est la probabilité qu'elle soit hâtive ? normale ? tardive ?
3. On observe une plante grande. Quelle est la probabilité qu'elle soit hâtive ? normale ? tardive ?
4. On observe une plante hâtive. Quelle est la probabilité qu'elle soit naine ? grande ?

Exercice 1.2.3. Dans un lot de pièces fabriquées, il y a 5% de pièces défectueuses. On contrôle les pièces, mais le mécanisme de contrôle est aléatoire. Si la pièce est bonne, elle est acceptée avec une probabilité égale à 0.96 ; si la pièce est mauvaise, elle est refusée avec probabilité 0.98. On choisit au hasard une pièce que l'on contrôle.

1. Quelle est la probabilité que cette pièce soit refusée ?
2. Quelle est la probabilité que cette pièce soit bonne, sachant qu'elle est refusée ?

3. Quelle est la probabilité que cette pièce soit mauvaise sachant qu'elle est acceptée ?
4. Quelle est la probabilité qu'il y ait une erreur dans le contrôle (une bonne pièce est refusée ou une mauvaise est acceptée) ?

Exercice 1.2.4. Voici la répartition en pourcentages des différents groupes sanguins en France.

Groupe / Facteur	O	A	B	AB
Rhésus +	37.0	38.1	6.2	2.8
Rhésus –	7.0	7.2	1.2	0.5

1. Déterminer la distribution de probabilité des quatre groupes O, A, B, AB dans l'ensemble de la population.
2. Déterminer la distribution de probabilité des quatre groupes parmi les individus de rhésus positif.
3. Déterminer la distribution de probabilité des quatre groupes parmi les individus de rhésus négatif.
4. Si on choisit au hasard un individu de groupe O, quelle est la probabilité qu'il soit de rhésus négatif ? Même question pour un individu de groupe B.

1.3 Loi binomiale

- Au cours de n *expériences répétées indépendamment*, la variable aléatoire X égale au nombre de réalisations d'un même *événement de probabilité* p, suit la loi *binomiale* de paramètres n et p.
- La variable X peut prendre toutes les valeurs entières entre 0 et n.
- Pour tout entier k entre 0 et n, la variable X prend la valeur k avec la probabilité :

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

où

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \times (n-1) \cdots \times (n-k+1)}{k \times (k-1) \cdots \times 3 \times 2 \times 1}$$

est le nombre de manières de choisir k objets parmi n.

- L'espérance de X est np, sa variance est $np(1-p)$.

Exercice 1.3.1. On sait par expérience qu'une certaine opération chirurgicale a 90% de chances de réussir. On s'apprête à réaliser l'opération sur 5 patients. Soit X la variable aléatoire égale au nombre de réussites de l'opération sur les 5 tentatives.

1. Quel modèle proposez-vous pour X ?

   En supposant que les résultats (succès ou échec) des 5 opérations soient indépendants entre eux, le nombre de succès suit la loi binomiale de paramètres 5 *et* 0.9*. La variable aléatoire X prend ses valeurs dans l'ensemble* $\{0,1,2,3,4,5\}$*, et pour tout entier k dans cet ensemble :*

   $$\mathbb{P}[X=k] = \binom{5}{k} 0.9^k \, 0.1^{5-k} \; .$$

2. Quelle est la probabilité que l'opération rate les 5 fois ?

   $$\mathbb{P}[X=0] = 0.1^5 = 0.00001 \; .$$

3. Quelle est la probabilité que l'opération rate exactement 3 fois ?

   $$\mathbb{P}[X=2] = \binom{5}{2} 0.9^2 \, 0.1^3 = 0.0081 \; .$$

4. Quelle est la probabilité que l'opération réussisse au moins 3 fois ?

   $$\begin{aligned} \mathbb{P}[X \geqslant 3] &= \mathbb{P}[X=3] + \mathbb{P}[X=4] + \mathbb{P}[X=5] \\ &= \binom{5}{3} 0.9^3 \, 0.1^2 + \binom{5}{4} 0.9^4 \, 0.1^1 + \binom{5}{5} 0.9^5 \, 0.1^0 \\ &= 0.0729 + 0.32805 + 0.59049 = 0.99144 \; . \end{aligned}$$

Exercice 1.3.2. Quand un chasseur tire sur un lapin sans défense, il a une chance sur 10 de le toucher.

1. Deux chasseurs tirent indépendamment sur le même lapin. Calculer la probabilité que :
   (a) aucun ne le touche ;
   (b) un seul chasseur le touche ;
   (c) les deux chasseurs le touchent.
2. Quatre chasseurs tirent indépendamment sur le même lapin.
   (a) Quelle est la loi de probabilité du nombre de coups de fusils reçus par la pauvre bête ? Donner l'espérance et la variance de cette loi.
   (b) Quelle est la probabilité que le lapin reçoive au plus 2 coups de fusil ?
   (c) Quelle est la probabilité que le lapin reçoive au moins 2 coups de fusil ?
3. Dix chasseurs tirent indépendamment sur le même lapin.
   (a) Quelle est la probabilité que le lapin conserve l'étanchéité de sa fourrure ?

(b) Quelle est la probabilité que le lapin soit immangeable (s'il a reçu au moins 5 coups de fusil).

Exercice 1.3.3. Lors d'une séance d'identification, on propose à 6 témoins de désigner un coupable parmi 4 suspects, dont vous faites partie.

1. Si chacun des 6 témoins choisissait au hasard, quelles seraient vos chances :
   (a) de n'être jamais désigné ?
   (b) d'être désigné exactement une fois ?
   (c) d'être désigné deux fois ou plus ?
2. Il se trouve que 2 des 6 témoins vous ont désigné comme le coupable. Par référence au résultat de la question 1 (c), pensez-vous que le juge pourra attribuer cela au hasard ?
3. Et si 4 des 6 témoins vous avaient désigné ?

1.4 Loi hypergéométrique

- Dans un ensemble de N éléments, parmi lesquels m sont marqués, on en choisit au hasard n distincts. La variable aléatoire X égale au nombre d'éléments marqués parmi l'échantillon de n suit la la loi *hypergéométrique de paramètres* N, m, n.
- Dans le cas où $n \leq m$ et $n \leq N - m$, X peut prendre toutes les valeurs entières entre 0 et n.
- Pour tout entier k entre 0 et n, X prend la valeur k avec probabilité :

$$\mathbb{P}[X = k] = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} .$$

- L'espérance de X est nm/N.

Exercice 1.4.1. Un groupe d'étudiants est composé de 18 filles et de 11 garçons. On choisit au hasard dans ce groupe un échantillon de 5 personnes. Soit X la variable aléatoire égale au nombre de filles dans cet échantillon.

1. Quel modèle proposez-vous pour X ?

   La loi de X est la loi hypergéométrique de paramètres $N = 29$ (nombre total d'individus), $m = 18$ (les individus "marqués" sont les filles) et $n = 5$ (la taille de l'échantillon extrait). Les valeurs prises sont les entiers de 0 à 5. Pour tout entier $k = 0, 1 \ldots, 5$, on a :

$$\mathbb{P}[X = k] = \frac{\binom{18}{k}\binom{11}{5-k}}{\binom{29}{5}} .$$

2. Donner l'espérance de X.

   L'espérance de X est $5 \times 18/29 \simeq 3.1$. C'est la taille de l'échantillon, multipliée par la proportion de filles dans le groupe.

3. Calculer la probabilité que l'échantillon ne contienne que des filles.

$$\mathbb{P}[X = 5] = \frac{\binom{18}{5}}{\binom{29}{5}} \simeq 0.072 \ .$$

4. Calculer la probabilité que l'échantillon contienne au moins une fille.

   On doit calculer $\mathbb{P}[X \geqslant 1]$. On pourrait calculer $\mathbb{P}[X = 1] + \mathbb{P}[X = 2] + \mathbb{P}[X = 3] + \mathbb{P}[X = 4] + \mathbb{P}[X = 5]$, mais il est plus rapide de calculer $1 - \mathbb{P}[X = 0]$, ce qui revient au même :

$$\mathbb{P}[X \geqslant 1] = 1 - \mathbb{P}[X = 0] = 1 - \frac{\binom{11}{5}}{\binom{29}{5}} \simeq 0.996 \ .$$

5. Calculer la probabilité que l'échantillon contienne exactement 3 filles.

$$\mathbb{P}[X = 3] = \frac{\binom{18}{3}\binom{11}{2}}{\binom{29}{5}} \simeq 0.378 \ .$$

Exercice 1.4.2. Dans chacune des situations suivantes, on donnera la loi de probabilité de la variable aléatoire X et son espérance. On calculera la probabilité que X soit égal à 0, puis la probabilité que X soit supérieur ou égal à 2.

1. À la belote, huit cartes sont distribuées à chacun des quatre joueurs. Soit X le nombre d'as que reçoit un joueur donné.
2. À la belote, les quatre joueurs jouent par équipes de deux. Soit X le nombre de piques d'une équipe donnée.
3. Au bridge, treize cartes sont distribuées à chacun des quatre joueurs. Soit X le nombre de figures (valet, dame ou roi) d'un joueur donné.
4. Au loto, vous avez coché 6 numéros sur une grille qui en comporte 49. Soit X le nombre de bons numéros sur votre grille.

1.5 Loi normale

- Si on n'a pas de logiciel à disposition, on lit dans les tables pour la loi $\mathcal{N}(0, 1)$:
  - ⋆ les valeurs de la fonction de répartition F : pour une valeur de x la table retourne la probabilité $p = P[X \leqslant x] = F(x)$.
  - ⋆ les valeurs de la fonction quantile F^{-1} : pour une probabilité p la table retourne la valeur de $x = F^{-1}(p)$ telle que $p = \mathbb{P}[X \leqslant x]$.

- La densité de la loi $\mathcal{N}(0,1)$ est symétrique :
$$\mathbb{P}[X \leqslant -x] = \mathbb{P}[X \geqslant x] \, .$$
- Si une variable aléatoire X suit la loi $\mathcal{N}(\mu, \sigma^2)$, alors $(X-\mu)/\sqrt{\sigma^2}$ suit la loi $\mathcal{N}(0,1)$. Ainsi :
$$\begin{aligned} \mathbb{P}[a \leqslant X \leqslant b] &= P\left[\frac{a-\mu}{\sqrt{\sigma^2}} \leqslant \frac{X-\mu}{\sqrt{\sigma^2}} \leqslant \frac{b-\mu}{\sqrt{\sigma^2}}\right] \\ &= F\left(\frac{b-\mu}{\sqrt{\sigma^2}}\right) - F\left(\frac{a-\mu}{\sqrt{\sigma^2}}\right) \, , \end{aligned}$$
où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.
- Si X et Y sont deux variables aléatoires indépendantes, de lois respectives $\mathcal{N}(\mu_x, \sigma_x^2)$ et $\mathcal{N}(\mu_y, \sigma_y^2)$, alors $X+Y$ suit la loi $\mathcal{N}(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$ et $X-Y$ suit la loi $\mathcal{N}(\mu_x-\mu_y, \sigma_x^2+\sigma_y^2)$.

Exercice 1.5.1. La taille X des hommes en France est modélisée par une loi normale $\mathcal{N}(172, 196)$ (unité : le cm).

1. Quelle proportion de français a une taille inférieure à 160 cm ?
$$\mathbb{P}[X < 160] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} < \frac{160-172}{\sqrt{196}}\right] = F(-0.857) = 1 - F(0.857) = 0.1957 \, ,$$
où F désigne (comme dans tout l'exercice) la fonction de répartition de la loi $\mathcal{N}(0,1)$.
2. Quelle proportion de français mesure plus de deux mètres ?
$$\mathbb{P}[X > 200] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} > \frac{200-172}{\sqrt{196}}\right] = 1 - F(2) = 0.02275 \, .$$
3. Quelle proportion des français mesure entre 165 et 185 centimètres ?
$$\begin{aligned} \mathbb{P}[165 < X < 185] &= \mathbb{P}\left[\frac{165-172}{\sqrt{196}} < \frac{X-172}{\sqrt{196}} < \frac{185-172}{\sqrt{196}}\right] \\ &= F(0.928) - F(-0.5) = 0.8234 - 0.3085 = 0.5149 \, . \end{aligned}$$
4. Si on classait dix mille français choisis au hasard par ordre de taille croissante, quelle serait la taille du 9000-ième ?

   La question revient à trouver la taille telle que 90% *des français aient une taille inférieure, à savoir le quantile d'ordre* 0.9*, ou encore le neuvième décile. Soit x cette taille.*
$$\mathbb{P}[X < x] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} < \frac{x-172}{\sqrt{196}}\right] = 0.9$$

Donc $\frac{x-172}{\sqrt{196}}$ *est la valeur de la fonction quantile de la loi* $\mathcal{N}(0,1)$ *au point* 0.9*, à savoir* 1.2816*. On en déduit :*

$$x = 172 + 1.2816 \times \sqrt{196} \simeq 190 \ cm.$$

5. La taille Y des françaises est modélisée par une loi normale $\mathcal{N}(162, 144)$ (en centimètres). Quelle est la probabilité pour qu'un homme choisi au hasard soit plus grand qu'une femme choisie au hasard ?

   Si X *désigne la taille de l'homme et* Y *la taille de la femme, supposées indépendantes, alors* $X - Y$ *suit la loi normale* $\mathcal{N}(10, 340)$*. La probabilité que* X *soit supérieure à* Y *est la probabilité que la différence soit positive :*

$$\mathbb{P}\left[X - Y > 0\right] = \mathbb{P}\left[\frac{(X-Y)-10}{\sqrt{340}} > \frac{0-10}{\sqrt{340}}\right] = 1 - F(-0.5423) = 0.7062 \ .$$

Exercice 1.5.2. Soit X une variable aléatoire de loi $\mathcal{N}(0,1)$.

1. Exprimer à l'aide de la fonction de répartition de X, puis calculer à l'aide de la table les probabilités suivantes.
   (a) $\mathbb{P}[X > 1.45]$
   (b) $\mathbb{P}[-1.65 \leqslant X \leqslant 1.34]$
   (c) $\mathbb{P}[|X| < 2.05]$
2. Déterminer la valeur de u dans les cas suivants.
   (a) $\mathbb{P}[X < u] = 0.63$
   (b) $\mathbb{P}[X \geqslant u] = 0.63$
   (c) $\mathbb{P}[|X| < u] = 0.63$

Exercice 1.5.3. Soit X une variable aléatoire suivant la loi $\mathcal{N}(0,1)$. On pose $Y = 2X - 3$.

1. Quelle est la loi de Y ?
2. Calculer $\mathbb{P}[Y < -4]$.
3. Calculer $\mathbb{P}[-2 < Y < 3]$.

Exercice 1.5.4. Soit X une variable aléatoire de loi $\mathcal{N}(3, 25)$.

1. Exprimer à l'aide de la fonction de répartition de la loi $\mathcal{N}(0,1)$, puis calculer à l'aide de la table les probabilités suivantes.
   (a) $\mathbb{P}[X < 6]$
   (b) $\mathbb{P}[X > -2]$
   (c) $\mathbb{P}[-1 \leqslant X \leqslant 1.5]$
2. Déterminer la valeur de u dans les cas suivants.

(a) $\mathbb{P}[X < u] = 0.63$
(b) $\mathbb{P}[X > u] = 0.63$
(c) $\mathbb{P}[|X - 3| \leqslant u] = 0.63$

Exercice 1.5.5. Dans un pays donné, le taux de cholestérol sérique d'un individu pris au hasard est modélisé par une loi normale avec une moyenne de 200 mg/100 mL et un écart-type de 20 mg/100 mL.

1. Quelle est la probabilité qu'un individu pris au hasard dans ce pays ait un taux de cholestérol inférieur à 160 mg/100 mL ?
2. Quelle proportion de la population a un taux de cholestérol compris entre 170 et 230 mg/100 mL ?
3. Dans un autre pays, le taux moyen de cholestérol sérique est de 190 mg/100 mL, pour le même écart-type. Reprendre les questions précédentes.
4. On choisit un individu au hasard dans le premier pays, puis dans le second. Quelle est la probabilité que le premier individu ait un taux supérieur au second ?

Exercice 1.5.6. La taille d'un épi de blé dans un champ est modélisée par une variable aléatoire X de loi normale $\mathcal{N}(15, 36)$ (unité : le cm).

1. Quelle est la probabilité pour qu'un épi ait une taille inférieure à 16 cm ?
2. On admet qu'il y a environ 15 millions d'épis dans le champ, donner une estimation du nombre d'épis de plus de 20 cm.
3. Quelle est la probabilité pour que 10 épis prélevés dans le champ aient tous leur taille dans l'intervalle $[16\ ;\ 20]$?
4. On suppose que la taille d'un épi de blé d'un autre champ est modélisée par une variable aléatoire Y de loi normale $\mathcal{N}(10, 16)$ et que X et Y sont des variables indépendantes. Quelle est la probabilité pour qu'un épi pris dans le premier champ soit plus grand qu'un épi pris dans le second ?

1.6 Approximation d'une loi binomiale par une loi normale

- Pour n assez grand, on peut approcher la loi binomiale $\mathcal{B}(n, p)$ par la loi normale $\mathcal{N}(np, np(1-p))$, qui a la même espérance et la même variance.
- Dans ces conditions, si X suit la loi $\mathcal{B}(n, p)$, on calcule la probabilité que X se trouve dans l'intervalle $[a, b]$ par :

$$\begin{aligned}\mathbb{P}[a \leqslant X \leqslant b] &= P\left[\frac{a-np}{\sqrt{np(1-p)}} \leqslant \frac{X-np}{\sqrt{np(1-p)}} \leqslant \frac{b-np}{\sqrt{np(1-p)}}\right] \\ &\simeq F\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - F\left(\frac{a-np}{\sqrt{np(1-p)}}\right),\end{aligned}$$

où F est la fonction de répartition de la loi $\mathcal{N}(0, 1)$.

Exercice 1.6.1. On sait par expérience qu'une certaine opération chirurgicale a 90% de chances de réussir. Cette opération est réalisée dans une clinique 400 fois chaque année. Soit N le nombre de réussites dans une année. On utilisera l'approximation normale pour N.

1. Calculer l'espérance et la variance de N.

   L'espérance vaut $400 \times 0.9 = 360$*, la variance vaut* $400 \times 0.9 \times 0.1 = 36$*.*

2. Calculer la probabilité que la clinique réussisse au moins 345 opérations dans l'année.

$$\begin{aligned}\mathbb{P}[N \geqslant 345] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \geqslant \frac{345-360}{\sqrt{36}}\right] \\ &= 1 - F(-2.5) = F(2.5) = 0.9938\ .\end{aligned}$$

3. Calculer la probabilité que la clinique rate plus de 28 opérations dans l'année.

$$\begin{aligned}\mathbb{P}[N \leqslant 372] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \leqslant \frac{372-360}{\sqrt{36}}\right] \\ &= F(2) = 0.9772\ .\end{aligned}$$

4. L'assurance accepte de couvrir un certain nombre d'opérations ratées : ce nombre n'a que 1% de chances d'être dépassé. Quel est-il ?

   Soit n *le nombre d'opérations ratées cherché. Le nombre d'opérations réussies est* $400-n$*. Il vérifie* $\mathbb{P}[N \leqslant 400-n] = 0.01$*. Or :*

$$\begin{aligned}\mathbb{P}[N \leqslant 400-n] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \leqslant \frac{400-n-360}{\sqrt{36}}\right] \\ &= F\left(\frac{40-n}{\sqrt{36}}\right) = 0.01\ .\end{aligned}$$

   Le nombre $\frac{40-n}{\sqrt{36}}$ *est le quantile d'ordre* 0.01 *de la loi normale* $\mathcal{N}(0,1)$*, à savoir* -2.3236*. Donc :*

$$\frac{40-n}{\sqrt{36}} = -2.3263 \implies n = 40 + 2.3263\sqrt{36} \simeq 54\ .$$

   On peut aussi raisonner sur le nombre d'opérations ratées $R = 400-N$*. Il suit la loi binomiale* $\mathcal{B}(400, 0.1)$*, que l'on peut approcher par la loi normale* $\mathcal{N}(40, 36)$*.*

Le nombre cherché est tel que $\mathbb{P}[R > n] = 0.01$.

$$\begin{aligned}\mathbb{P}[R > n] &= \mathbb{P}\left[\frac{R-40}{\sqrt{36}} > \frac{n-40}{\sqrt{36}}\right] \\ &= 1 - F\left(\frac{n-40}{\sqrt{36}}\right) \\ &= F\left(\frac{40-n}{\sqrt{36}}\right) = 0.01\ .\end{aligned}$$

Bien sûr, le résultat est le même.

Exercice 1.6.2. On évalue à 0.4 la probabilité qu'une personne en âge d'être vaccinée contre la grippe demande effectivement à l'être. Sur une population de 150000 personnes en âge d'être vaccinées, soit N le nombre de personnes qui demanderont à l'être.

1. Quel modèle proposez-vous pour N ?
2. Si on prépare 60500 vaccins, quelle est la probabilité qu'il n'y en ait pas suffisamment ?
3. Calculer le nombre m de vaccins qu'il faudrait prévoir pour que la probabilité d'en manquer soit égale à 0.1.

Exercice 1.6.3. Un restaurant servant des repas uniquement sur réservation, dispose de 50 places. La probabilité qu'une personne ayant réservé ne vienne pas est 1/5. On note N le nombre de repas servis un jour donné. On utilisera l'approximation normale pour N.

1. Si le patron accepte 50 réservations, quelle est la probabilité qu'il serve plus de 45 repas ?
2. S'il accepte 55 réservations, quelle est la probabilité qu'il se retrouve dans une situation embarassante ?

Exercice 1.6.4. On suppose qu'il y a une probabilité égale à 0.1 d'être contrôlé lorsqu'on prend le tramway. Mr A. fait 700 voyages par an. On utilisera l'approximation normale pour le nombre de contrôles.

1. Quelle est la probabilité que Mr A. soit contrôlé entre 60 et 80 fois dans l'année ?
2. Mr A. est en fait un fraudeur et voyage toujours sans ticket. Sachant que le prix d'un ticket est de 1 euro, quelle amende minimale la régie de transports devrait-elle fixer pour que le fraudeur ait, sur une période d'une année, une probabilité supérieure à 0.75 d'être perdant ?

Exercice 1.6.5. Entre Grenoble et Valence TGV, deux bus de 50 places font le trajet le vendredi à 16h10. On estime que le nombre de personnes se présentant pour effectuer le trajet est en moyenne de 80 avec un écart-type de 10. On utilise l'approximation normale pour ce nombre.

1. Calculer la probabilité que les deux autobus soient pleins.
2. L'un des deux bus part de la gare, l'autre part de la place Victor Hugo : les voyageurs choisissent au hasard l'un ou l'autre, mais ne peuvent pas changer si le bus qu'ils ont choisi est plein. Supposons que 90 voyageurs veuillent aller de Grenoble à Valence. Quelle est la probabilité que l'un d'entre eux ne trouve pas de place ?
3. Avec les hypothèses de la question précédente, quelle devrait être la taille minimale des bus pour que la probabilité de refuser un voyageur soit inférieure à 0.05 ?

Exercice 1.6.6. On admet qu'en moyenne, un passager qui a acheté un billet d'avion, se présente à l'enregistrement avec probabilité 0.9. Un avion comporte 200 places.

1. Si la compagnie accepte 220 réservations, quelle est la probabilité qu'elle doive refuser des passagers ?
2. Combien de réservations peut-elle accepter au maximum pour que la probabilité de refuser un passager soit inférieure ou égale à 0.01 ?

2 Estimation paramétrique

2.1 Estimation ponctuelle

- Pour un paramètre inconnu, un estimateur est une fonction des données, qui prend des valeurs proches de ce paramètre. Il est *sans biais* si son espérance est égale au paramètre. Il est *convergent* si la probabilité qu'il prenne des valeurs à distance au plus ε du paramètre, tend vers 1 quand la taille de l'échantillon tend vers l'infini.
- La *fréquence empirique* d'un événement est un estimateur sans biais et convergent de la probabilité de cet événement.
- La *moyenne empirique* d'un échantillon est un estimateur sans biais et convergent de l'espérance théorique des variables.
- La *variance empirique* d'un échantillon est un estimateur convergent de la variance théorique des variables. On obtient un estimateur sans biais en multipliant la variance empirique par $n/(n-1)$, où n est la taille de l'échantillon.

Exercice 2.1.1. On considère l'échantillon statistique $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. Calculer sa moyenne et sa variance empiriques.

   On trouve :
   $$\overline{x} = \frac{6}{9} = \frac{2}{3} \quad \textit{et} \quad s_x^2 = \frac{4}{9} \, .$$

2. En supposant que les données de cet échantillon sont des réalisations d'une variable de loi inconnue, donner une estimation non biaisée de l'espérance et de la variance de cette loi.

   La moyenne empirique ($2/3$*) est une estimation non biaisée de l'espérance. On obtient une estimation non biaisée de la variance en multipliant* s_x^2 *par* $9/8$ *: on trouve* $1/2$*.*

3. On choisit de modéliser les valeurs de cet échantillon par une loi binomiale $\mathcal{B}(2, p)$. Utiliser la moyenne empirique pour proposer une estimation ponctuelle pour p.

   L'espérance de la loi $\mathcal{B}(2, p)$ *est* $2p$*. Elle est estimée par la moyenne empirique (ici :* $2/3$*). Donc la probabilité* p *peut être estimée par :*
   $$\frac{2/3}{2} = \frac{1}{3} \, .$$

4. Avec le même modèle, utiliser la variance empirique pour proposer une autre estimation de p.

   La variance de la loi $\mathcal{B}(2, p)$ *est* $2p(1-p)$*. Elle est estimée par* $1/2$*. On obtient une estimation de* p *en résolvant l'équation* $2p(1-p) = 1/2$*, dont la solution est* $p = 1/2$*.*

5. On choisit de modéliser les valeurs de cet échantillon par une loi de Poisson $\mathcal{P}(\lambda)$, qui a pour espérance λ. Quelle estimation ponctuelle proposez-vous pour λ ?

   On estime λ par la moyenne empirique, $2/3$.

Exercice 2.1.2. On considère l'échantillon statistique

$$(1, 3, 2, 3, 2, 2, 0, 2, 3, 1) \ .$$

1. En supposant que les variables de cet échantillon sont des réalisations d'une variable de loi inconnue, donner une estimation non biaisée de l'espérance et de la variance de cette loi.
2. On choisit de modéliser les valeurs de cet échantillon par une loi binomiale $\mathcal{B}(3, p)$. Utiliser la moyenne empirique pour proposer une estimation ponctuelle pour p.

Exercice 2.1.3. On considère l'échantillon statistique

$$(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4) \ .$$

1. On choisit de modéliser les valeurs de cet échantillon par une loi uniforme sur l'intervalle $[0, \theta]$. Quelle estimation ponctuelle proposez-vous pour θ ?
2. On choisit de modéliser les valeurs de cet échantillon par une loi normale $\mathcal{N}(\mu, \sigma^2)$. Quelle estimation ponctuelle proposez-vous pour μ et σ^2 ?

2.2 Intervalles de confiance pour un échantillon gaussien

Un échantillon gaussien est un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$. On note :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2 \ ,$$

la moyenne et la variance empiriques de l'échantillon.

- Si la variance théorique σ^2 est *connue*, on obtient un intervalle de confiance de niveau $1-\alpha$ pour μ par :

  $$\left[\overline{X} - u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}} \ ; \ \overline{X} + u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}}\right] \ ,$$

  où u_α est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0, 1)$.
- Si la variance théorique σ^2 est *inconnue*, on obtient un intervalle de confiance de niveau $1-\alpha$ pour μ par :

  $$\left[\overline{X} - t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}} \ ; \ \overline{X} + t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}}\right] \ ,$$

  où t_α est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre $n-1$.

- Si la variance théorique σ^2 est *inconnue*, on obtient un intervalle de confiance de niveau $1-\alpha$ pour σ^2 par :
$$\left[\frac{nS^2}{v_\alpha}\,;\,\frac{nS^2}{u_\alpha}\right],$$
où u_α est le quantile d'ordre $\alpha/2$ de la loi de khi-deux de paramètre $n-1$, et v_α est son quantile d'ordre $1-\alpha/2$.

Exercice 2.2.1. La force de compression d'un type de béton est modélisée par une variable gaussienne d'espérance μ et de variance σ^2. L'unité de mesure est le *psi* (pound per square inch). Dans les questions de 1. à 4., on supposera la variance σ^2 connue et égale à 1000. Sur un échantillon de 12 mesures, on a observé une moyenne empirique de 3250 psi.

1. Donner un intervalle de confiance de niveau 0.95 pour μ.

   Ici, $\alpha = 0.05$ *et* $1-\alpha/2 = 0.975$. *Le quantile d'ordre* 0.975 *de la loi* $\mathcal{N}(0,1)$ *est* 1.96. *L'intervalle de confiance est :*
$$\left[3250 - 1.96\frac{\sqrt{1000}}{\sqrt{12}}\,;\,3250 + 1.96\frac{\sqrt{1000}}{\sqrt{12}}\right] = [3232\,;\,3268]\,.$$
   Il est inutile de donner plus de chiffres que n'en a la moyenne empirique. On arrondit la borne inférieure par défaut, la borne supérieure par excès ; ainsi l'arrondi ne peut qu'agrandir l'intervalle, et on est assuré que le niveau de confiance de l'intervalle donné est au moins égal à 0.95.

2. Donner un intervalle de confiance de niveau 0.99 pour μ. Comparer sa largeur avec celle de l'intervalle précédent.

   Ici, $\alpha = 0.01$ *et* $1-\alpha/2 = 0.995$.*Le quantile d'ordre* 0.995 *de la loi* $\mathcal{N}(0,1)$ *est* 2.5758. *L'intervalle de confiance est :*
$$\left[3250 - 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\,;\,3250 + 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\right] = [3226\,;\,3274]\,.$$
   *L'intervalle est plus large que le précédent. Plus la probabilité que la moyenne appartienne à l'intervalle est grande (*0.99 *au lieu de* 0.95*), plus cet intervalle doit être large. Si on veut avoir plus confiance dans l'intervalle, il faut accepter qu'il soit moins précis.*

3. Si avec le même échantillon on donnait un intervalle de confiance de largeur 30 psi, quel serait son niveau de confiance ?

   La largeur de l'intervalle de confiance de niveau $1-\alpha$ *est :*
$$2u_\alpha\frac{\sqrt{1000}}{\sqrt{12}}\,.$$

Si cette largeur est égale à 30*, on obtient :*

$$u_\alpha = \frac{30\sqrt{12}}{2\sqrt{1000}} = 1.6432 \ .$$

Cette valeur est le quantile d'ordre $0.9498 = 1-\alpha/2$ *de la loi* $\mathcal{N}(0,1)$*. Donc* $\alpha = 0.1003$ *et* $1-\alpha = 0.8997$*.*

4. On souhaite maintenant estimer μ avec une précision de ± 15 psi, avec un niveau de confiance de 0.95. Quelle taille minimum doit avoir l'échantillon ?

   Pour un échantillon de taille n*, La précision de l'intervalle de confiance de niveau* 0.95 *est :*

$$\pm 1.96 \frac{\sqrt{1000}}{\sqrt{n}} \ .$$

   Si elle est égale à 15*, on obtient :*

$$n = \left(\frac{1.96\sqrt{1000}}{15} \right)^2 = 17.07 \ .$$

   L'échantillon doit donc être de taille 18 *au moins.*

5. La variance théorique est désormais supposée inconnue. On dispose de la donnée suivante (sur le même échantillon de taille 12) :

$$\sum_{i=1}^{12} x_i^2 = 126761700 \ .$$

   Donnez pour μ un intervalle de confiance de niveau 0.95 et comparez-le avec celui de la question 1, puis un intervalle de confiance de niveau 0.99 et comparez-le avec celui de la question 2.

   La variance estimée est :

$$s^2 = \frac{1}{12} \times 126761700 - (3250)^2 = 975 \ .$$

   Le quantile d'ordre 0.975 *de la loi de Student* $\mathcal{T}(n–1)$ *est* 2.201*, le quantile d'ordre* 0.995 *est* 3.106*. L'intervalle de confiance de niveau* 0.95 *est :*

$$\left[3250 - 2.201 \frac{\sqrt{975}}{\sqrt{11}} \ ; \ 3250 + 2.201 \frac{\sqrt{975}}{\sqrt{11}} \right] \ = \ [3229 \, ; \, 3271] \ .$$

   L'intervalle de confiance de niveau 0.99 *est :*

$$\left[3250 - 3.106 \frac{\sqrt{975}}{\sqrt{11}} \ ; \ 3250 + 3.106 \frac{\sqrt{975}}{\sqrt{11}} \right] \ = \ [3220 \, ; \, 3280] \ .$$

À niveau de confiance égal, et bien que la variance estimée soit inférieure à la variance théorique, l'intervalle de confiance calculé avec la loi de Student (variance supposée inconnue) est plus large, donc moins précis, que celui calculé avec la loi normale (variance connue). Cela tient au fait que les lois de Student sont plus dispersées que la loi normale $\mathcal{N}(0,1)$ *: l'intervalle contenant* 95% *des valeurs pour la loi* $\mathcal{T}(11)$ *est* $[-2.201\,;\,+2.201]$*, au lieu de* $[-1.96\,;\,+1.96]$ *pour la loi* $\mathcal{N}(0,1)$*. Il est raisonnable de s'attendre à une moins grande précision quand on dispose de moins d'information sur le modèle.*

6. Donner un intervalle de confiance de niveau 0.95 pour la variance, et pour l'écart-type.

   Le quantile d'ordre 0.025 *pour la loi de khi-deux* $\mathcal{X}^2(11)$ *est* $u_\alpha = 3.816$*. Le quantile d'ordre* 0.975 *est* $v_\alpha = 21.92$*. L'intervalle de confiance de niveau* 0.95 *pour la variance est :*

$$\left[\frac{12\times 975}{21.92}\,;\,\frac{12\times 975}{3.816}\right] \;=\; [533\,;\,3067]\,.$$

   En prenant la racine carrée des deux bornes, on obtient un intervalle de confiance pour l'écart-type :

$$\left[\sqrt{\frac{12\times 975}{21.92}}\,;\,\sqrt{\frac{12\times 975}{3.816}}\right] \;=\; [23.1\,;\,55.4]\,.$$

   Les intervalles de confiance pour la variance ou l'écart-type pour de petits échantillons sont en général très imprécis.

Exercice 2.2.2. On a mesuré le poids de raisin produit par pied sur 10 pieds pris au hasard dans une vigne. On a obtenu les résultats suivants exprimés en kilogrammes :

$$2.4 \quad 3.4 \quad 3.6 \quad 4.1 \quad 4.3 \quad 4.7 \quad 5.4 \quad 5.9 \quad 6.5 \quad 6.9\,.$$

On modélise le poids de raisin produit par une souche de cette vigne par une variable aléatoire de loi $\mathcal{N}(\mu,\sigma^2)$.

1. Calculer la moyenne et la variance empiriques de l'échantillon.
2. Donner un intervalle de confiance de niveau 0.95 pour μ.
3. Donner un intervalle de confiance de niveau 0.95 pour σ^2.
4. On suppose désormais que l'écart-type des productions par pied est connu et égal à 1.4. Donner un intervalle de confiance de niveau 0.95 pour μ.
5. Quel nombre de pieds au minimum devrait-on observer pour estimer μ au niveau de confiance 0.99 avec une précision de plus ou moins 500 grammes ?

Exercice 2.2.3. Une étude faite sur la vitesse coronarienne a donné les résultats suivants sur 18 individus :

75, 77, 78, 77, 77, 72, 72, 72, 70, 71, 69, 69, 68, 66, 64, 66, 62, 61.

On modélise les valeurs de cet échantillon par une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma^2)$, où μ et σ^2 sont deux paramètres a priori inconnus.

1. Calculer la moyenne et la variance de l'échantillon.
2. Calculer les intervalles de confiance de μ aux niveaux 0.95, 0.98 et 0.99.
3. Calculer les intervalles de confiance de σ^2 aux niveaux 0.95, 0.98 et 0.99.
4. Que seraient les intervalles de confiance de μ, si on supposait que la variance σ^2 était connue et égale à 26 ?

Exercice 2.2.4. Un laboratoire utilise un appareil de mesure optique destiné à mesurer la concentration des solutions de fluoresceïne. Les résultats des mesures sont modélisés par une variable aléatoire normale dont l'espérance est égale à la concentration réelle de la solution, et l'écart-type, garanti par le constructeur, est connu : $\sigma = 0.05$.

1. On effectue 9 mesures à partir d'une solution donnée. La moyenne empirique des 9 mesures est 4.38 mg/l. Donner un intervalle de confiance pour la concentration réelle de la solution, au niveau de confiance 0.99.
2. Pour le même échantillon, quel est le niveau de confiance de l'intervalle $[4.36\ ;\ 4.40]$?
3. Quelle devrait être la taille de l'échantillon pour connaître la concentration réelle de la solution, au niveau de confiance 0.99, avec une précision de ± 0.01 mg/l ?
4. Sur le même échantillon de 9 mesures, on a observé un écart-type empirique de 0.08 mg/l. Donner un intervalle de confiance pour l'écart-type réel, de niveau de confiance 0.99. Que pensez-vous de la garantie du constructeur ?
5. Reprendre la première question, en supposant cette fois que l'écart-type de la loi des mesures est inconnu, et estimé par l'écart-type empirique.

Exercice 2.2.5. Pour étudier la pourriture des pommes de terre, un chercheur injecte à 13 pommes de terre des bactéries qui causent cette pourriture. Il mesure ensuite la surface pourrie (en mm^2) sur ces 13 pommes de terre. Il obtient une moyenne empirique de 7.84 mm^2 pour une variance empirique de 14.13. On modélise la surface pourrie d'une pomme de terre par une loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Calculer un intervalle de confiance pour μ au niveau 0.95 puis 0.99.
2. Calculer un intervalle de confiance pour σ^2 au niveau 0.95 puis 0.99.

Exercice 2.2.6. On désire estimer la production d'une nouvelle espèce de pommier. On modélise la production d'un pommier de cette espèce par une loi normale d'espérance μ et d'écart-type σ inconnus.

1. Sur un échantillon de 15 pommiers, on a observé une récolte moyenne de 52 kg avec un écart-type de 5 kg. Donner un intervalle de confiance pour la production moyenne des pommiers de cette espèce, de niveau 0.95, puis 0.99.
2. Donner un intervalle de confiance pour l'écart-type σ, de niveau 0.95.

2.3 Int. de conf. d'une espérance pour un grand échantillon

Pour un grand échantillon, on obtient un intervalle de confiance de niveau approché $1-\alpha$ pour l'espérance par :

$$\left[\overline{X} - u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}} \; ; \; \overline{X} + u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}}\right] ,$$

où u_α est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Exercice 2.3.1. On a effectué 90 mesures de concentration d'une solution de fluorescеïne. On a observé une moyenne empirique de 4.38 mg/l et un écart-type empirique de 0.08 mg/l. Donner un intervalle de confiance pour la concentration réelle de la solution, au niveaux de confiance 0.95 et 0.99.

Le quantile d'ordre 0.975 *de la loi* $\mathcal{N}(0,1)$ *est* 1.96. *L'intervalle de confiance de niveau* 0.95 *est :*

$$\left[4.38 - 1.96\frac{0.08}{\sqrt{90}} \; ; \; 4.38 + 1.96\frac{0.08}{\sqrt{90}}\right] \; = \; [4.363\,;\, 4.397] \; .$$

Le quantile d'ordre 0.995 *de la loi* $\mathcal{N}(0,1)$ *est* 2.5758. *L'intervalle de confiance de niveau* 0.99 *est :*

$$\left[4.38 - 2.5758\frac{0.08}{\sqrt{90}} \; ; \; 4.38 + 2.5758\frac{0.08}{\sqrt{90}}\right] \; = \; [4.358\,;\, 4.402] \; .$$

Exercice 2.3.2. On désire estimer la production d'une nouvelle espèce de pommier. Sur un échantillon de 80 pommiers, on observe une récolte moyenne de 51.5 kg, avec un écart-type de 4.5 kg. Donner un intervalle de confiance pour la production moyenne des pommiers de cette espèce, de niveau 0.95, puis 0.99.

Exercice 2.3.3. On a mesuré la longueur en millimètres de 152 œufs de coucou, et obtenu une moyenne empirique de 40.8 mm, pour une variance empirique de 14.7 mm^2. Donner un intervalle de confiance pour la longueur moyenne d'un œuf de coucou, au niveau de confiance 0.95, puis 0.98, puis 0.99.

Exercice 2.3.4. On a mesuré la longueur de 150 coquilles de noix et obtenu une moyenne empirique de 27.6 mm, pour un écart-type empirique de 3.7 mm. Donner un intervalle de confiance pour la longueur moyenne d'une coquille de noix, au niveau de confiance 0.99, puis 0.998.

Exercice 2.3.5. On administre des somnifères à deux groupes de malades A et B comprenant 50 et 100 individus. Le groupe A reçoit un nouveau somnifère, le groupe B reçoit l'ancien. Les patients du groupe A ont dormi 7.82 heures en moyenne avec un écart-type de 0.24 h ; ceux du groupe B ont dormi 6.75 heures en moyenne avec un écart-type de 0.30 h.

1. Calculer l'intervalle de confiance pour le nombre moyen d'heures de sommeil d'un patient recevant le nouveau somnifère, aux niveaux 0.90, puis 0.95 et 0.99.
2. Même question pour un patient recevant l'ancien somnifère.
3. Pensez-vous que le nouveau somnifère soit plus efficace que l'ancien ?

2.4 Int. de conf. d'une probabilité pour un grand échantillon

Pour un grand échantillon binaire, on obtient un intervalle de confiance de niveau approché $1-\alpha$ pour la probabilité de l'événement par :

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}} \; ; \; \overline{X} + u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}}\right] ,$$

où n est la taille de l'échantillon, $\overline{X}$ est la fréquence empirique de l'événement et u_α est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Exercice 2.4.1. Afin d'étudier l'influence des rayons X sur la spermatogénèse de Bombyx Mori, on a irradié des mâles au deuxième jour et au quatrième jour du stade larvaire ; ces mâles ont été accouplés avec des femelles non irradiées. On a compté le nombre d'œufs fertiles dans la ponte des femelles, et on a obtenu 4998 œufs fertiles pour 5646 œufs récoltés en tout. On a aussi accouplé des mâles et des femelles non irradiés, avec un résultat de 5834 œufs fertiles sur 6221 œufs récoltés.

1. Donner un intervalle de confiance de niveau 0.95 pour la proportion d'œufs fertiles après irradiation des mâles.

   La fréquence empirique des œufs fertiles après irradiation des mâles est :

   $$F = \frac{4998}{5646} = 0.885 .$$

   L'intervalle de confiance de niveau 0.95 *est :*

   $$\left[0.885 - 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}} \; ; \; 0.885 + 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}}\right]$$

   $$= \; [0.876\,;\, 0.894] .$$

2. Donner un intervalle de confiance de niveau 0.95 pour la proportion d'œufs fertiles de couples non irradiés.

   La fréquence empirique des œufs fertiles parmi les couples non irradiés est :

   $$F = \frac{5834}{6221} = 0.938 .$$

L'intervalle de confiance de niveau 0.95 *est :*

$$\left[0.938 - 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\ ;\ 0.938 + 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\right]$$

$$= \quad [0.931\ ;\ 0.944]\ .$$

3. Que pensez-vous de l'influence de l'irradiation sur la fertilité des œufs ?

   Les deux intervalles de confiance ont une intersection vide ; la proportion d'œufs fertiles est donc significativement plus basse pour les mâles irradiés.

Exercice 2.4.2. On a observé un échantillon de taille $n = 500$ d'adolescents de 15 ans, dans lequel 210 présentent un surpoids. Soit p la proportion d'adolescents de 15 ans qui présentent un surpoids. Donner un intervalle de confiance pour p, aux niveaux de confiance 0.95 et 0.99.

Exercice 2.4.3. Une clinique a proposé une nouvelle opération chirurgicale, et a connu 40 échecs, sur 200 tentatives. On note p le pourcentage de réussite de cette nouvelle opération.

1. Quelle estimation de p proposez-vous ?
2. En utilisant l'approximation normale, donner un intervalle de confiance pour p de niveau de confiance 0.95.
3. Combien d'opérations la clinique devrait-elle réaliser pour connaître le pourcentage de réussite avec une précision de plus ou moins 1%, au niveau de confiance 0.95 ?

3 Tests statistiques

3.1 Règle de décision, seuil et p-valeur

- Dans un test, l'*hypothèse nulle* $\mathcal{H}_0$ est celle dont on choisit de maîtriser la probabilité de rejet à tort. C'est celle à laquelle on tient le plus, celle qu'il serait le plus dangereux ou le plus coûteux de rejeter à tort.
- Le *seuil* du test, encore appelé *risque de première espèce* est la probabilité de rejeter $\mathcal{H}_0$ à tort :
$$\mathbb{P}_{\mathcal{H}_0}[\text{ Rejet de } \mathcal{H}_0\,] = \alpha\ .$$
- La *statistique de test* est une fonction des données, dont on connaît la distribution de probabilité sous l'hypothèse nulle $\mathcal{H}_0$.
- La *règle de décision* spécifie, en fonction des valeurs de la statistique de test, dans quel cas on rejette l'hypothèse $\mathcal{H}_0$.
- Un test peut être :
  - ⋆ *bilatéral* si la règle de décision est :
$$\text{Rejet de } \mathcal{H}_0 \iff T \notin [l, l']$$
(rejet des valeurs trop grandes ou trop petites). On convient habituellement de choisir l et l' de sorte que $\mathbb{P}_{\mathcal{H}_0}[T < l] = \mathbb{P}_{\mathcal{H}_0}[T > l'] = \alpha/2$.
  - ⋆ *unilatéral* si la règle de décision est :
$$\text{Rejet de } \mathcal{H}_0 \iff T < l$$
(rejet des valeurs trop petites),
ou bien :
$$\text{Rejet de } \mathcal{H}_0 \iff T > l$$
(rejet des valeurs trop grandes).
- La *p-valeur* est le seuil pour lequel la valeur observée de la statistique de test serait la limite de la région de rejet. C'est la probabilité sous $\mathcal{H}_0$ que la statistique de test soit au-delà de la valeur déjà observée.
- Le *risque de deuxième espèce* est la probabilité d'accepter $\mathcal{H}_0$ à tort, où encore la probabilité d'accepter $\mathcal{H}_0$ quand l'*hypotèse alternative* $\mathcal{H}_1$ est vraie :
$$\mathbb{P}_{\mathcal{H}_1}[\text{ accepter } \mathcal{H}_0\,] = \beta\ .$$
La *puissance* du test est $1-\beta$. C'est la probabilité de rejeter $\mathcal{H}_0$ en ayant raison.

Exercice 3.1.1. Chez un individu adulte, le logarithme du dosage en d-dimères, variable que nous noterons X, est modélisé par une loi normale d'espérance μ et de variance σ^2. La variable X est un indicateur de risque cardio-vasculaire : on considère que chez les individus sains, μ vaut -1, alors que chez les individus à risque, μ vaut 0. Dans les deux cas, la valeur de σ^2 est la même : 0.09.

1. Le Dr. House ne souhaite pas alarmer inutilement ses patients. Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ choisira-t-il de tester ? Donner la règle de décision pour son test, au seuil de 1%, et au seuil de 5%.

   Si Dr. House ne veut pas alarmer inutilement un patient, l'hypothèse qu'il considère comme dangereux de rejeter à tort est que celui-ci n'est pas à risque, donc que sa variable X (la statistique de test) a pour espérance -1. Son hypothèse $\mathcal{H}_0$ est donc $\mu = -1$ (le patient ne présente pas de risque), qu'il teste contre $\mathcal{H}_1$: $\mu = 0$ (le patient présente un risque). Il choisira de rejeter des valeurs trop élevées de X. La règle de décision sera donc :

$$\textit{Rejet de } \mathcal{H}_0 \iff X > l \ ,$$

   où :

$$\mathbb{P}_{\mathcal{H}_0}[X > l] = \alpha \ .$$

   Sous l'hypothèse $\mathcal{H}_0$, la statistique de test X suit la loi $\mathcal{N}(-1, 0.09)$, donc $\frac{X-(-1)}{\sqrt{0.09}}$ suit la loi $\mathcal{N}(0,1)$. Une règle de décision équivalente est :

$$\textit{Rejet de } \mathcal{H}_0 \iff \frac{X-(-1)}{\sqrt{0.09}} > \frac{l-(-1)}{\sqrt{0.09}} \ .$$

   Donc $\frac{l-(-1)}{\sqrt{0.09}}$ est la valeur qui a probabilité α d'être dépassée pour une variable de loi $\mathcal{N}(0,1)$: 1.6449 pour $\alpha = 0.05$, 2.3263 pour $\alpha = 0.01$. Au seuil 0.05 la règle de décision du test est :

$$\begin{aligned} \textit{Rejet de } \mathcal{H}_0 &\iff \frac{X-(-1)}{\sqrt{0.09}} > 1.6449 \\ &\iff X > 1.6449\sqrt{0.09} + (-1) = -0.5065 \ . \end{aligned}$$

   On déclare que le patient présente un risque cardio-vasculaire quand son dosage en d-dimères est supérieur à -0.5065.
   Au seuil 0.01 la règle de décision du test est :

$$\begin{aligned} \textit{Rejet de } \mathcal{H}_0 &\iff \frac{X-(-1)}{\sqrt{0.09}} > 2.3263 \\ &\iff X > 2.3263\sqrt{0.09} + (-1) = -0.3021 \ . \end{aligned}$$

   Plus le seuil est faible, moins la règle de décision rejette d'individus à risque : ce qui doit se produire pour rejeter $\mu = -1$ au seuil 0.01 est plus extraordinaire qu'au seuil 0.05.

2. Calculer le risque de deuxième espèce et la puissance des tests de la question précédente.

Le risque de deuxième espèce est la probabilité de rejeter $\mathcal{H}_1$ à tort. Sous l'hypothèse $\mathcal{H}_1$, $\mu = 0$, la variable X suit la loi $\mathcal{N}(0, 0.09)$.
Pour le test de seuil 0.05*, la probabilité d'accepter $\mathcal{H}_0$ à tort (déclarer à tort qu'un patient ne présente pas de risque) est :*

$$\beta = \mathbb{P}_{\mathcal{H}_1}[X \leqslant -0.5065] = \mathbb{P}_{\mathcal{H}_1}\left[\frac{X-0}{\sqrt{0.09}} \leqslant \frac{-0.5065-0}{\sqrt{0.09}}\right]$$

Or sous l'hypothèse $\mathcal{H}_1$, $\frac{X-0}{\sqrt{0.09}}$ suit la loi $\mathcal{N}(0,1)$. Nous devons donc calculer la probabilité, pour une variable de loi $\mathcal{N}(0,1)$ de tomber en-dessous de $\frac{-0.5065-0}{\sqrt{0.09}} = -1.6885$: c'est la valeur de la fonction de répartition de la loi $\mathcal{N}(0,1)$ au point -1.6885*, à savoir* 0.0457*. La puissance est :*

$$1 - \beta = 1 - 0.0457 = 0.9543\ .$$

Pour le test de seuil 0.01*, le raisonnement est le même, en remplaçant la valeur limite* -0.5065 *par* -0.3021*. On trouve un risque de deuxième espèce égal à* 0.1570 *et une puissance égale à* 0.8430*.*
Quand on abaisse le seuil, on diminue le risque de rejeter $\mathcal{H}_0$ à tort, mais on augmente aussi le risque de l'accepter à tort, et on diminue la puissance. Pour le test de seuil 0.01*, la probabilité que le médecin se trompe en déclarant qu'un patient n'est pas à risque est de l'ordre de 16%.*

3. Un patient présente une valeur de X égale à -0.46. Calculer la p-valeur du test du Dr. House.

   La p-valeur est le seuil pour lequel -0.46 *serait la valeur limite. Au vu des résultats de la première question, comme* -0.46 *est entre* -0.5065 *et* -0.3021*, la p-valeur est comprise entre* 0.05 *et* 0.01*. Elle est égale à la probabilité sous $\mathcal{H}_0$, que la variable X soit supérieure à* -0.46*.*

$$\mathbb{P}_{\mathcal{H}_0}[X > -0.46] = \mathbb{P}_{\mathcal{H}_0}\left[\frac{X-(-1)}{\sqrt{0.09}} > \frac{-0.46-(-1)}{\sqrt{0.09}}\right] = \mathbb{P}_{\mathcal{H}_0}\left[\frac{X-(-1)}{\sqrt{0.09}} > 1.8\right].$$

   Or sous $\mathcal{H}_0$, $\frac{X-(-1)}{\sqrt{0.09}}$ suit la loi $\mathcal{N}(0,1)$. La probabilité cherchée est $1 - F(1.8)$, où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$, à savoir 0.0359*.*

4. Le Dr. Cuddy a pour point de vue qu'il vaut mieux alarmer à tort un patient plutôt que de ne pas l'avertir d'un risque réel. Quelles hypothèses $\mathcal{H}'_0$ et $\mathcal{H}'_1$ choisira-t-elle de tester ? Donner la règle de décision pour son test, au seuil de 1%, et au seuil de 5%.

   Si Dr. Cuddy ne veut pas manquer un patient à risque, l'hypothèse qu'elle considère comme dangereux de rejeter à tort est que celui-ci est à risque, donc que sa variable X (la statistique de test) a pour espérance 0*. Son hypothèse $\mathcal{H}'_0$ est donc $\mu = 0$ (le patient présente un risque), qu'elle teste contre $\mathcal{H}'_1$: $\mu = -1$ (le patient*

ne présente pas de risque). Elle choisira de rejeter des valeurs trop basses de X. La règle de décision sera donc :

$$\textit{Rejet de } \mathcal{H}'_0 \iff X < l' \ ,$$

où :

$$\mathbb{P}_{\mathcal{H}'_0}[X < l'] = \alpha \ .$$

Sous l'hypothèse $\mathcal{H}'_0$, la statistique de test X suit la loi $\mathcal{N}(0, 0.09)$, donc $\frac{X-0}{\sqrt{0.09}}$ suit la loi $\mathcal{N}(0,1)$. Une règle de décision équivalente est :

$$\textit{Rejet de } \mathcal{H}'_0 \iff \frac{X-0}{\sqrt{0.09}} < \frac{l'-0}{\sqrt{0.09}} \ .$$

Donc $\frac{l'-0}{\sqrt{0.09}}$ est la valeur telle qu'une variable de loi $\mathcal{N}(0,1)$ tombe en-dessous avec probabilité α : -1.6449 pour $\alpha = 0.05$, -2.3263 pour $\alpha = 0.01$. Au seuil 0.05 la règle de décision du test est :

$$\begin{aligned} \textit{Rejet de } \mathcal{H}_0 &\iff \frac{X-0}{\sqrt{0.09}} < -1.6449 \\ &\iff X < -1.6449 \times \sqrt{0.09} + 0 = -0.4935 \ . \end{aligned}$$

Le Dr. Cuddy déclare que le patient ne présente pas de risque cardio-vasculaire quand son dosage en d-dimères est inférieur à -0.4935.
Au seuil 0.01 la règle de décision du test est :

$$\begin{aligned} \textit{Rejet de } \mathcal{H}'_0 &\iff \frac{X-0}{\sqrt{0.09}} < -2.3263 \\ &\iff X < -2.3263 \times \sqrt{0.09} + 0 = -0.6980 \ . \end{aligned}$$

5. Selon le seuil, pour quelles valeurs de X les deux médecins seront-ils d'accord ?

   Si $X < \min\{l, l'\}$, Le Dr. House accepte $\mathcal{H}_0$, le Dr. Cuddy rejette $\mathcal{H}'_0$. Dans les deux cas, la conclusion pour le patient est la même : il n'est pas à risque. À l'inverse, si $X > \max\{l, l'\}$ le Dr. House rejette $\mathcal{H}_0$, le Dr. Cuddy accepte $\mathcal{H}'_0$ et la conclusion est identique : le patient est à risque.

   Les conclusions des deux médecins diffèrent pour les patients dont la valeur de X se situe entre l et l'.

   Au seuil 0.05 les valeurs limites des deux tests sont $l = -0.5065$ et $l' - 0.4935$. Pour un patient dont la variable X est entre -0.5065 et -0.4935 (par exemple -0.5), le Dr. House déclare qu'il est à risque (il rejette $\mathcal{H}_0$), le Dr.Cuddy déclare qu'il n'est pas à risque (elle rejette $\mathcal{H}'_0$).

Au seuil 0.01*, les valeurs limites sont* $l = -0.3021$ *et* $l' = -0.6980$*. Pour un patient dont la variable* X *est entre* -0.6980 *et* -0.3021 *(par exemple* -0.5*), le Dr. House déclare qu'il n'est pas à risque (il accepte* $\mathcal{H}_0$*), le Dr. Cuddy déclare qu'il est à risque (elle accepte* $\mathcal{H}'_0$*).*

6. Donner la règle de décision du test de seuil 0.05, pour l'hypothèse nulle $\mathcal{H}''_0 : \mu = -1$ contre l'hypothèse alternative $\mathcal{H}''_1 : \mu \neq -1$.

   Il s'agit ici d'un test bilatéral. La règle de décision sera donc :

   $$\textit{Rejet de } \mathcal{H}''_0 \iff X \notin [l_1, l_2] \, ,$$

   où :

   $$\mathbb{P}_{\mathcal{H}''_0}[\, X \notin [l_1, l_2]\,] = 0.05 \, .$$

   Sous l'hypothèse $\mathcal{H}''_0$*, la statistique de test* X *suit la loi* $\mathcal{N}(-1, 0.09)$*, donc* $\frac{X-(-1)}{\sqrt{0.09}}$ *suit la loi* $\mathcal{N}(0,1)$*. Une règle de décision équivalente est :*

   $$\textit{Rejet de } \mathcal{H}''_0 \iff \frac{X-(-1)}{\sqrt{0.09}} \notin \left[\frac{l_1-(-1)}{\sqrt{0.09}} \, ; \, \frac{l_2-(-1)}{\sqrt{0.09}}\right] \, .$$

   L'intervalle $\left[\frac{l_1-(-1)}{\sqrt{0.09}} \, ; \, \frac{l_2-(-1)}{\sqrt{0.09}}\right]$ *doit contenir* 95% *des valeurs d'une variable suivant la loi* $\mathcal{N}(0,1)$*. On choisit l'intervalle centré en* 0 *:* $[-1.96\,;\,+1.96]$*. Donc :*

   $$\frac{l_1-(-1)}{\sqrt{0.09}} = -1.96 \implies l_1 = (-1) - 1.96\sqrt{0.09} = -1.588 \, ,$$

   et :

   $$\frac{l_2-(-1)}{\sqrt{0.09}} = +1.96 \implies l_2 = (-1) + 1.96\sqrt{0.09} = -0.412 \, ,$$

   Au seuil 0.05 *la règle de décision du test est :*

   $$\textit{Rejet de } \mathcal{H}_0 \iff X \notin [-1.588\,;\,-0.412] \, .$$

   On déclare que le patient présente un dosage significativement différent de -1 *quand sa variable* X *est soit inférieure à* -1.588*, soit supérieure à* -0.488*.*

7. Un patient présente une valeur de X égale à -0.46. Calculer la p-valeur du test de la question précédente.

   La p-valeur est le seuil pour lequel la valeur observée serait limite de la région de rejet. Cette région de rejet est centrée en -1*. L'autre valeur limite devrait donc être* $-1-(-0.46-(-1)) = -1.54$*.*

La p-valeur est la probabilité suivante.

$$\begin{aligned}
& \mathbb{P}_{\mathcal{H}_0''}[X \notin [-1.54\,;\,-0.46]\,] \\
= \;& \mathbb{P}_{\mathcal{H}_0''}\left[\frac{X-(-1)}{\sqrt{0.09}} \notin \left[\frac{-1.54-(-1)}{\sqrt{0.09}}\,;\,\frac{-0.46-(-1)}{\sqrt{0.09}}\right]\right] \\
= \;& \mathbb{P}_{\mathcal{H}_0''}\left[\frac{X-(-1)}{\sqrt{0.09}} \notin [-1.8\,;\,+1.8]\right] .
\end{aligned}$$

Sous l'hypothèse $\mathcal{H}_0''$, la variable $\frac{X-(-1)}{\sqrt{0.09}}$ suit la loi $\mathcal{N}(0,1)$: la probabilité cherchée est 0.0719. *La p-valeur que l'on trouve est le double de celle du test unilatéral de la question 3.*

Exercice 3.1.2. Une machine à emballer est censée produire des paquets de 1 kg. Le poids réel des paquets est modélisé par une variable aléatoire suivant une loi normale dont l'écart-type vaut 20 g. Par contre, il est possible de régler le poids moyen des paquets.

1. Le responsable de la production décide de ne pas mettre à la vente les paquets dont le poids s'écarterait trop de la valeur nominale de 1 kg. Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ doit-il tester ? Établir la règle de décision de ce test aux seuils de 5% et 1%.
2. Le patron de l'usine prétend que les paquets mis à la vente sont souvent trop lourds, ce qui fait perdre de l'argent à l'usine. Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ le responsable de production doit-il tester ? Établir la règle de décision de ce test aux seuils de 5% et 1%.
3. On pèse un paquet de 1018 grammes. Quelle est la p-valeur du test de la question précédente ? Quelle est la p-valeur du test de la question 1 ?
4. Une association de consommateurs accuse l'usine de mettre à la vente des paquets de poids trop faible. Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ le responsable de production doit-il tester ? Établir la règle de décision de ce test aux seuils de 5% et 1%.
5. On pèse un paquet de 982 grammes. Quelle est la p-valeur du test de la question précédente ? Quelle est la p-valeur du test de la question 1. ?

Exercice 3.1.3. Une concentration en paracétamol de plus de 150 mg par kilogramme de poids corporel est considérée comme dangereuse. Les mesures de paracétamol dans les tests sanguins sont modélisées par une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma^2)$. L'écart-type, lié à la procédure de test est supposé connu et égal à 5 mg.

1. Donner les hypothèses et la règle de décision du test décidant, au seuil de 5%, si un patient court un risque, au vu du résultat d'un test sanguin (vous êtes un docteur prudent).

2. Un patient montrant des signes d'empoisonnement au paracétamol arive à l'hôpital. On effectue un test sanguin et on trouve une concentration de 140 mg. Donner la p-valeur du test de la question précédente. Doit-on considérer que ce patient court un risque ?

Exercice 3.1.4. Soit X l'indice de pollution mesuré près d'une usine. On modélise X par une loi $\mathcal{N}(\mu, \sigma^2)$. On admet que l'écart-type σ est connu, et vaut 4. Les normes fixent à 30 l'indice moyen de pollution maximal.

1. Le directeur de l'usine souhaite montrer que celle-ci est aux normes. Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ doit-il tester ? Établir la règle de décision de ce test aux seuils de 5% et 1%.
2. Une association écologiste veut démontrer que l'usine est hors-normes. Quelles hypothèses $\mathcal{H}'_0$ et $\mathcal{H}'_1$ doit-elle tester ? Établir la règle de décision de ce test aux seuils de 5% et 1%.

3.2 Tests sur un échantillon

On note :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$$

la moyenne et la variance empiriques de l'échantillon. L'espérance de la loi inconnue est μ, sa variance est σ^2. Les statistiques de test à utiliser et leur loi de probabilité sous l'hypothèse nulle $\mathcal{H}_0$ sont les suivantes.

- Test de valeurs de l'espérance, échantillon gaussien, σ^2 connu.

$$\mathcal{H}_0 \ : \mu = \mu_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2}}\right) \text{ suit la loi normale } \mathcal{N}(0,1) \ .$$

- Test de valeurs de l'espérance, échantillon gaussien, σ^2 inconnu.

$$\mathcal{H}_0 \ : \mu = \mu_0 \quad ; \quad T = \sqrt{n-1}\left(\frac{\overline{X} - \mu_0}{\sqrt{S^2}}\right) \text{ suit la loi de Student } \mathcal{T}(n-1) \ .$$

- Test de valeurs de la variance, échantillon gaussien, σ^2 inconnu.

$$\mathcal{H}_0 \ : \sigma^2 = \sigma_0^2 \quad ; \quad T = n\left(\frac{S^2}{\sigma_0^2}\right) \text{ suit la loi de khi-deux } \mathcal{X}^2(n-1) \ .$$

- Test de valeurs de l'espérance, grand échantillon, σ^2 connu ou non.

$$\mathcal{H}_0 \ : \mu = \mu_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sqrt{S^2}}\right) \text{ suit la loi normale } \mathcal{N}(0,1) \ .$$

- Test de valeurs d'une probabilité, échantillon binaire de grande taille.

$$\mathcal{H}_0 \ : p = p_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}}\right) \text{ suit la loi normale } \mathcal{N}(0,1) \ .$$

Exercice 3.2.1. Chez un individu adulte, le logarithme du dosage en d-dimères, variable que nous noterons X, est modélisé par une loi normale d'espérance μ et de variance σ^2. La variable X est un indicateur de risque cardio-vasculaire : on considère que chez les individus sains, μ vaut -1, alors que chez les individus à risque, μ vaut 0. On souhaite étudier l'influence de la consommation d'huile d'olive sur le risque cardio-vasculaire.

1. On a fait suivre un régime à base d'huile d'olive à un groupe de 13 patients, précédemment considérés comme à risque. Après le régime, on a mesuré la valeur de X pour chaque patient, et obtenu une moyenne empirique de -0.15. On suppose σ^2 connu et égal à 0.09. Donner la règle de décision du test de $\mathcal{H}_0 \ : \mu = 0$ contre $\mathcal{H}_1 \ : \mu = -1$, au seuil de 5%. Quelle est la p-valeur correspondant à -0.15 ? Quelle est votre conclusion ? Calculer le risque de deuxième espèce et la puissance du test.

   On se trouve dans le cas d'un échantillon gaussien avec variance connue, et on construit un test sur la valeur de l'espérance. La statistique de test est :

$$T = \sqrt{13}\frac{\overline{X} - 0}{\sqrt{0.09}} \ .$$

   Sous l'hypothèse $\mathcal{H}_0$, T suit la loi normale $\mathcal{N}(0,1)$. On rejette l'hypothèse $\mathcal{H}_0$ quand T prend des valeurs trop basses. Au seuil de 5% la valeur limite est -1.6449. La règle de décision est :

$$\textit{Rejet de } \mathcal{H}_0 \iff T < -1.6449 \iff \overline{X} < -0.1369 \ .$$

   Pour $\overline{X} = -0.15$, la statistique du test prend la valeur -1.8028, la p-valeur correspondante est 0.0357. Au seuil de 5% on rejette $\mathcal{H}_0$, c'est-à-dire qu'on déclare qu'il y a eu une amélioration significative. Mais à un seuil inférieur à 3.57%, on ne peut pas rejeter $\mathcal{H}_0$.
   Sous l'hypothèse $\mathcal{H}_1$, $\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}}$ suit la loi $\mathcal{N}(0,1)$. Le risque de deuxième espèce est la probabilité d'accepter $\mathcal{H}_0$ à tort, à savoir :

$$\begin{aligned}
\beta &= \mathbb{P}_{\mathcal{H}_1}[\overline{X} > -0.1369] \\
&= \mathbb{P}_{\mathcal{H}_1}\left[\sqrt{13}\frac{\overline{X} - (-1)}{\sqrt{0.09}} > \sqrt{13}\frac{-0.1369 - (-1)}{\sqrt{0.09}}\right] \\
&= \mathbb{P}_{\mathcal{H}_1}\left[\sqrt{13}\frac{\overline{X} - (-1)}{\sqrt{0.09}} > 10.3732\right] \\
&\simeq 0 \ .
\end{aligned}$$

Le risque de deuxième espèce est très proche de 0 *(inférieur à* 10^{-20}*), et la puissance très proche de* 1*.*

2. Pour le même groupe de 13 patients, on a observé un écart-type empirique égal à 0.37. Donner la règle de décision du test de $\mathcal{H}_0 : \sigma^2 = 0.09$, contre $\mathcal{H}_1 : \sigma^2 \neq 0.09$, au seuil de 5%. Quelle est votre conclusion ?

   Il s'agit de tester une valeur de la variance pour un échantillon gaussien. La statistique de test est :

$$T = 13\frac{S^2}{0.09} .$$

   Sous l'hypothèse $\mathcal{H}_0$*, elle suit la loi de khi-deux de paramètre* 12*. On souhaite un test bilatéral, donc une règle de décision qui écarte les valeurs trop basses ou trop hautes.*

$$\textit{Rejet de } \mathcal{H}_0 \iff T \notin [l, l'] ,$$

   où l *et* l' *sont les quantiles d'ordre* 0.025 *et* 0.975 *de la loi de khi-deux de paramètre* 12 *:* $l = 4.4038$ *et* $l' = 23.3367$*. Ici, la statistique de test prend la valeur* 19.7744*. C'est une valeur élevée, mais pas suffisamment pour rejeter l'hypothèse que la variance théorique est de* 0.09*.*

3. En supposant la variance inconnue, et en utilisant l'estimation de la question précédente, donner la règle de décision du test de $\mathcal{H}_0 : \mu = 0$, contre $\mathcal{H}_1 : \mu < 0$, au seuil de 5%. Quelle est votre conclusion ?

   On se trouve dans le cas d'un échantillon gaussien avec variance inconnue, et on construit un test sur la valeur de l'espérance. La statistique de test est :

$$T = \sqrt{12}\frac{\overline{X} - 0}{\sqrt{S^2}} .$$

   Sous l'hypothèse $\mathcal{H}_0$*,* T *suit la loi de Student* $\mathcal{T}(12)$*. On rejette l'hypothèse* $\mathcal{H}_0$ *quand* T *prend des valeurs trop basses. Au seuil de* 5% *la valeur limite est* -1.7823*. La règle de décision est :*

$$\textit{Rejet de } \mathcal{H}_0 \iff T < -1.7823 .$$

   Pour $\overline{X} = -0.15$ *et* $\sqrt{S^2} = 0.37$*, la statistique du test* T *prend la valeur* -1.4044*, donc on ne peut pas rejeter* $\mathcal{H}_0$ *(la p-valeur correspondante est* 0.0928*), c'est-à-dire qu'on déclare qu'il n'y a pas eu d'amélioration significative.*

4. On reprend l'expérience sur un groupe de 130 patients, pour lesquels on observe une moyenne empirique de -0.12 et un écart-type de 0.32. Donner la règle de décision du test de $\mathcal{H}_0 : \mu = 0$ contre $\mathcal{H}_1 : \mu < 0$, au seuil de 5%. Quelle est la p-valeur correspondant à -0.12 ? Quelle est votre conclusion ?

   On doit maintenant tester une valeur de l'espérance pour un grand échantillon.

La statistique de test est :

$$T = \sqrt{130}\frac{\overline{X} - 0}{\sqrt{S^2}} .$$

Sous l'hypothèse $\mathcal{H}_0$, T suit la loi normale $\mathcal{N}(0,1)$. On rejette l'hypothèse $\mathcal{H}_0$ quand T prend des valeurs trop basses. Au seuil de 5% la valeur limite est -1.6449. La règle de décision est :

$$\textit{Rejet de } \mathcal{H}_0 \iff T < -1.6449 .$$

Pour $\overline{X} = -0.12$ et $\sqrt{S^2} = 0.32$, la statistique du test prend la valeur -4.2757, la p-valeur correspondante est proche de 10^{-5}. On peut donc conclure sans hésiter que pour ce groupe de patients, le dosage moyen est significativement inférieur à 0.

5. On avait mesuré le dosage en d-dimères des 130 patients avant le régime. À l'issue du régime, le dosage a baissé pour 78 patients, monté pour 52 patients. Construire un test permettant de décider si le régime à base d'huile d'olive a amélioré l'état d'une proportion significative des patients. Avec les observations dont vous disposez, quelle est la p-valeur de ce test, quelle est votre conclusion ?

   Notons p la probabilité que le régime à base d'huile d'olive améliore l'état du patient, c'est-à-dire fasse baisser son dosage en d-dimères. Si le régime n'avait pas d'effet, les fluctuations du dosage seraient purement aléatoires et il y aurait autant d'augmentations que de diminutions : la proportion d'améliorations serait de 50%. Nous devons donc tester, pour un grand échantillon binaire, l'hypothèse $\mathcal{H}_0 : p = 0.5$, contre $\mathcal{H}_1 : p > 0.5$. La statistique de test est :

$$T = \sqrt{n}\frac{\overline{X} - 0.5}{\sqrt{0.5(1-0.5)}} .$$

   Sous l'hypothèse $\mathcal{H}_0$, la statistique de test suit la loi normale $\mathcal{N}(0,1)$. Ici $\overline{X}$ est la proportion observée d'améliorations, à savoir $78/130$. La statistique de test prend la valeur 2.2804, la p-valeur correspondante (probabilité qu'une variable de loi $\mathcal{N}(0,1)$ dépasse 2.2804) est 0.0113. Au seuil de 5% on peut conclure que l'amélioration est significative, mais pas tout à fait au seuil de 1%.

Exercice 3.2.2. Une machine à emballer est censée produire des paquets de 1 kg. Le poids réel des paquets est modélisé par une variable aléatoire suivant une loi normale dont l'écart-type vaut 20 g. Il est possible de régler le poids moyen des paquets. Pour contrôler que la machine est bien réglée, on prélève un échantillon de 10 paquets que l'on pèse pour calculer la moyenne empirique de leurs poids.

1. Soit $\mathcal{H}_0$ l'hypothèse : "le poids moyen est de 1 kg". Construire un test au seuil 1%, de $\mathcal{H}_0$ contre l'hypothèse $\mathcal{H}_1$: "le poids moyen est différent de 1 kg". Calculer la p-valeur de ce test, pour un échantillon sur lequel on a observé une moyenne empirique de 1011 grammes.

2. Reprendre la question précédente pour l'hypothèse $\mathcal{H}_1$: "le poids moyen est supérieur à 1 kg".
3. Reprendre les deux questions précédentes pour un échantillon de 100 paquets, de poids moyen 1005 g.
4. Sur un échantillon de 10 paquets, on a observé un poids moyen de 1011 g, avec un écart-type empirique de 32 grammes. Au seuil de 1%, cette observation est-elle compatible avec la valeur de 20 g donnée pour l'écart-type théorique ?
5. Pour l'échantillon de la question précédente, en supposant la variance inconnue, peut-on dire que les paquets sont significativement trop lourds en moyenne au seuil de 1% ?

Exercice 3.2.3. Une concentration en paracétamol de plus de 150 mg par kilogramme de poids corporel est considérée comme dangereuse. Les mesures de paracétamol dans les tests sanguins sont modélisées par une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma^2)$. L'écart-type, lié à la procédure de test est supposé connu et égal à 5 mg. Par sécurité, on effectue 4 tests, dont les résultats sont supposés être des réalisations indépendantes de la même loi normale. $\mathcal{N}(\mu, \sigma^2)$.

1. Donner les hypothèses et la règle de décision du test décidant, au seuil de 5%, si un patient court un risque, au vu de ses 4 résultats (vous êtes un docteur prudent).
2. Sur un certain patient, les 4 tests ont donné des concentrations en paracétamol de 140, 133, 148, 144. Calculer la p-valeur du test de la question précédente. Ce patient court-il un risque ?
3. À partir de cette question, l'écart-type est supposé *inconnu*. Donner la statistique de test et la règle de décision du test décidant, au seuil de 5%, si un patient court un risque, au vu de ses 4 résultats.
4. Pour le patient de la question 2, donner un intervalle contenant la p-valeur du test de la question précédente. Quelle est votre conclusion ?

Exercice 3.2.4. Dans une population donnée, le poids des nouveaux-nés est modélisé par une loi normale. Dans l'ensemble de la population, l'écart-type des poids à la naissance est de 380 g. Le poids moyen d'un nouveau-né dont la mère ne fume pas est de 3400 g. Afin d'étudier l'effet du tabac sur le poids d'un nouveau-né, on relève le poids de 30 nouveau-nés dont les mères fument et on obtient une moyenne empirique de 3240 g, avec un écart-type de 426 g.

1. En supposant que l'écart-type de l'échantillon est connu et égal à celui de l'ensemble de la population, donner la p-valeur du test permettant de décider, si les nouveaux-nés de l'échantillon sont significativement plus légers en moyenne. Quelle est votre conclusion, au seuil de 5% ?
2. En supposant l'écart-type inconnu, donner une statistique de test et une région de rejet, pour tester les mêmes hypothèses qu'à la question précédente. Quelle est votre conclusion ?

3. L'écart-type observé est-il significativement supérieur à celui de l'ensemble de la population ?
4. Reprendre la question 1. avec un échantillon de 300 nouveaux-nés, pour lesquels on a observé un poids moyen de 3340 g.

Exercice 3.2.5. On dispose de l'échantillon suivant, de 15 longueurs d'œufs de coucou (exprimées en millimètres) :

19.8, 22.1, 21.5, 20.9, 22.0, 21.0, 22.3, 21.0, 20.3, 20.9, 22.0, 22.0, 20.8, 21.2, 21.0 .

On donne :

$$\sum x_i = 318.8 \quad \text{et} \quad \sum x_i^2 = 6782.78 \, .$$

On modélise la longueur d'un œuf de coucou par une loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Calculer la moyenne empirique et la variance empirique de cet échantillon.
2. Tester l'hypothèse $\mathcal{H}_0 \ : \sigma^2 = 0.4$ contre $\mathcal{H}_1 \ : \sigma^2 > 0.4$, au seuil de 5%.
3. Tester l'hypothèse $\mathcal{H}_0 \ : \mu = 21$" contre $\mathcal{H}_1 \ : \mu > 21$, au seuil de 5%.
4. Donner un encadrement de la p-valeur pour le test de la question précédente.

Exercice 3.2.6. À la suite d'un traitement sur une variété de rongeurs, on prélève un échantillon de 10 animaux et on les pèse. On obtient les poids en grammes suivants :

$$83 \, , \quad 81 \, , \quad 84 \, , \quad 80 \, , \quad 85 \, , \quad 87 \, , \quad 89 \, , \quad 84 \, , \quad 82 \, , \quad 80 \, .$$

On donne :

$$\sum x_i = 835 \quad \text{et} \quad \sum x_i^2 = 69801 \, .$$

On sait que les rongeurs non traités ont un poids moyen de 87.6 g. On modélise le poids d'un rongeur traité par loi normale.

1. Au seuil de 5%, tester l'hypothèse "le traitement n'a pas d'effet sur le poids moyen" contre "le traitement diminue le poids moyen".
2. Donner un encadrement de la p-valeur pour le test de la question précédente.

Exercice 3.2.7. Une société de location de voiture met en place une expérience afin de trancher entre deux types de pneus. Onze voitures sont conduites sur un parcours précis avec des pneus de type A. Les pneus sont alors remplacés par ceux de type B et les voitures sont de nouveau conduites sur le même parcours. Les consommations en litres pour 100 km des voitures en question sont modélisées par une loi normale. Voici les observations :

Voiture	1	2	3	4	5	6	7	8	9	10	11
Pneus A	4.2	4.7	6.6	7	6.7	4.5	5.7	6	7.4	4.9	6.1
Pneus B	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6

1. En admettant que les différences de consommation observées suivent une loi normale, quelle statistique de test proposez-vous ?
2. Quelles hypothèses allez-vous tester pour décider si les pneus ont un effet sur la consommation ?
3. Quelles hypothèses allez-vous tester pour décider si les pneus de type B sont significativement meilleurs en moyenne ?
4. Au seuil de 5% quelles sont vos conclusions ?

Exercice 3.2.8. Neuf malades présentant des symptômes d'anxiété reçoivent un tranquillisant. On évalue l'état du malade avant et après traitement par un indice que le médecin traitant calcule d'après les réponses à une série de questions. Si le traitement est efficace, l'indice doit diminuer. Les valeurs observées de cet indice sur les 9 patients sont les suivantes :

Avant	1.83	0.5	1.62	2.48	1.68	1.88	1.55	3.06	1.3
Après	0.88	0.65	0.59	2.05	1.06	1.29	1.06	3.14	1.29

1. En modélisant les valeurs des indices par une loi normale, quelle statistique de test proposez-vous ?
2. Donner un encadrement de la p-valeur pour le test permettant de décider si le tranquillisant apporte une amélioration significative en moyenne. Quelle est votre conclusion ?

Exercice 3.2.9. Une usine doit livrer des baguettes dont la longueur est modélisée par une loi normale d'espérance 40 mm. Les baguettes sont inutilisables si elles sont plus petites que 39 mm ou plus grandes que 41 mm, et l'usine garantit que moins de 1% des baguettes livrées sont inutilisables.

1. En supposant que la machine produit des baguettes à la bonne longueur en moyenne, quel doit être l'écart-type des longueurs pour que 1% des baguettes seulement soient inutilisables ?
2. Sur un échantillon de 15 baguettes, on a observé une moyenne empirique de 40.3 mm, avec un écart-type de 0.6 mm. L'écart-type observé est-il significativement supérieur à l'écart-type théorique de la question précédente ?
3. Les baguettes sont elles significativement trop longues en moyenne ?
4. Un client se plaint d'avoir reçu 112 baguettes inutilisables sur un lot de 10000. A-t-il raison de se plaindre ?

Exercice 3.2.10. Le pourcentage des femmes de 35 ans présentant des rides est de 25%. Sur 200 femmes de 35 ans ayant suivi un traitement antirides, on a observé que 40 avaient des rides. Au risque de 5%, peut-on dire que le traitement est efficace ?

Exercice 3.2.11. Pour une certaine maladie, on dispose d'un traitement satisfaisant dans 70% des cas. Un laboratoire propose un nouveau traitement et affirme qu'il donne satisfaction plus souvent que l'ancien traitement. Sur 200 malades ayant suivi ce nouveau traitement, on a observé une guérison pour 148 d'entre eux. En tant qu'expert chargé d'autoriser la mise sur le marché de ce nouveau traitement, que concluez-vous ?

Exercice 3.2.12. Voici le tableau des fréquences en France des principaux groupes sanguins :

Groupe / Facteur	O	A	B	AB
Rhésus +	0.370	0.381	0.062	0.028
Rhésus –	0.070	0.072	0.012	0.005

Le centre de transfusion sanguine de Pau a observé la répartition suivante sur 5000 donneurs.

Groupe / Facteur	O	A	B	AB
Rhésus +	2291	1631	282	79
Rhésus –	325	332	48	12

On souhaite répondre statistiquement aux questions ci-dessous. Dans chaque cas, on calculera la valeur prise par la statistique du test, on donnera la p-valeur, et on conclura.

1. Le type O+ est-il significativement plus fréquent à Pau ?
2. Parmi les individus de rhésus positif, la fréquence du groupe O est-elle significativement différente à Pau ?
3. Parmi les individus de groupe O, la fréquence du rhésus positif est elle significativement plus élevée à Pau ?

3.3 Comparaison de deux échantillons indépendants

Pour le premier échantillon :

$$\overline{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \quad \text{et} \quad S_x^2 = \left(\frac{1}{n_x}\sum_{i=1}^{n_x} X_i^2\right) - \overline{X}^2 \ ,$$

l'espérance de la loi inconnue est μ_x, sa variance est σ_x^2.

Pour le second échantillon :

$$\overline{Y} = \frac{1}{n_y}\sum_{j=1}^{n_y} Y_j \quad \text{et} \quad S_Y^2 = \left(\frac{1}{n_y}\sum_{j=1}^{n_y} Y_j^2\right) - \overline{Y}^2 \ ,$$

l'espérance de la loi inconnue est μ_y, sa variance est σ_y^2.

Les statistiques de test à utiliser et leur loi de probabilité sous l'hypothèse nulle $\mathcal{H}_0$ sont les suivantes.

- Test de Fisher : comparaison des variances, échantillon gaussien.

$$\mathcal{H}_0 \ : \sigma_x^2 = \sigma_y^2 \quad ; \quad T = \frac{\frac{n_x}{n_x-1}S_x^2}{\frac{n_y}{n_y-1}S_y^2} \text{ suit la loi de Fisher } \mathcal{F}(n_x-1, n_y-1) \ .$$

  Si $T < 1$, échanger le rôle de X et Y (ce qui revient à remplacer T par $1/T$) et comparer au quantile d'ordre $1-\alpha/2$ de la loi de Fisher $\mathcal{F}(n_x-1, n_y-1)$).
- Test de Student : comparaison des espérances, échantillon gaussien.

$$\mathcal{H}_0 \ : \mu_x = \mu_y \quad ; \quad T = \frac{\sqrt{n_x+n_y-2}}{\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \frac{\overline{X}-\overline{Y}}{\sqrt{n_x S_x^2 + n_y S_y^2}} \ ,$$

  suit la loi de Student $\mathcal{T}(n_x+n_y-2)$, si $\sigma_x = \sigma_y$.
- Test de comparaison des espérances, grands échantillons.

$$\mathcal{H}_0 \ : \mu_x = \mu_y \quad ; \quad T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}} \text{ suit la loi normale } \mathcal{N}(0,1) \ .$$

Exercice 3.3.1. On désire savoir si, chez les individus qui consomment régulièrement de l'huile d'olive, le risque cardio-vasculaire est diminué. On utilise pour cela le logarithme du dosage en d-dimères, modélisé par une loi normale. Sur un échantillon de 9 individus consommant de l'huile d'arachide, on a observé une moyenne de -0.78, avec un écart-type de 0.27. Sur un échantillon de 13 individus consommant de l'huile d'olive, on a observé une moyenne de -0.97, avec un écart-type de 0.32.

1. Tester l'hypothèse d'égalité des variances au seuil 0.05.

   Il s'agit d'appliquer le test de Fisher pour évaluer si la différence entre les variances observées des deux échantillons est significative ou non. On calcule la statistique du test de Fisher. Si on met au numérateur la variance la plus faible, on obtient :

$$T = \frac{\frac{9}{8}0.27^2}{\frac{13}{12}0.32^2} = 0.7393 \ .$$

   On doit tester $\mathcal{H}_0 \ : \sigma_x = \sigma_y$ *contre* $\mathcal{H}_1 \ : \sigma_x \neq \sigma_y$. *C'est donc un test bilatéral : il rejette les valeurs à l'extérieur de l'intervalle* $[l, l']$, *où* l *et* l' *sont les quantiles d'ordre* 0.025 *et* 0.975 *de la loi de* T *sous* $\mathcal{H}_0$, *qui est la loi de Fisher* $\mathcal{F}(8,12)$. *Or le quantile d'ordre* 0.025 *de la loi* $\mathcal{F}(8,12)$ *est l'inverse du quantile d'ordre* 0.975 *de la loi* $\mathcal{F}(12,8)$. *Il est donc plus simple d'échanger le rôle de* X *et* Y, *ce qui revient à calculer* $1/T = 1.3526$. *Cette valeur doit être comparée au quantile d'ordre* 0.975 *de la loi de Fisher de paramètres* 12 *et* 8 *(et non pas* 8 *et* 12 *puisqu'on a dû échanger* X *et* Y*). Cette valeur limite est* 4.1997. *La valeur observée* 1.3526 *est inférieure, donc on accepte l'hypothèse d'égalité des variances au seuil de* 5%.

2. Au seuil de 0.05, quel test proposez-vous pour décider si l'huile d'olive abaisse significativement le risque cardio-vasculaire ? Quelle est votre conclusion ? Donner un encadrement de la p-valeur.

   Le fait d'avoir accepté l'hypothèse d'égalité des variances justifie l'application du test de Student d'égalité des espérances. En notant X la variable "logarithme du dosage en d-dimères chez un individu consommant de l'huile d'arachide", et Y la même variable chez les individus consommant de l'huile d'olive, on souhaite tester :

   $$\mathcal{H}_0 \ : \mu_x = \mu_y \quad \textit{contre} \quad \mathcal{H}_1 \ : \mu_x > \mu_y \ .$$

   On utilise pour cela la statistique de test :

   $$T = \frac{\sqrt{n_x + n_y - 2}}{\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \frac{\overline{X} - \overline{Y}}{\sqrt{n_x S_x^2 + n_y S_y^2}} \ ,$$

   dont on rejettera les valeurs trop hautes.

   $$\textit{Rejet de } \mathcal{H}_0 \iff T > l \ .$$

   La valeur limite l est telle qu'une variable de loi de Student de paramètre $9+13-2 = 20$ soit supérieure avec probabilité 0.05, soit $l = 1.7247$. La statistique du test de Student prend la valeur 1.3055, donc on ne rejette pas l'hypothèse $\mathcal{H}_0$ d'égalité des espérances : la diminution observée en moyenne n'est pas significative au seuil de 5%. La p-valeur est la probabilité qu'une variable suivant la loi de Student $\mathcal{T}(20)$ soit supérieure à 1.3055. Sur la table, 1.3055 est entre les quantiles d'ordre 0.8 et 0.9, proche du quantile d'ordre 0.9. La p-valeur cherchée est donc comprise entre 0.1 et 0.2. La valeur numérique est 0.1033.

3. On effectue des dosages sur 110 individus consommant de l'huile d'arachide, pour lesquels on observe une moyenne de -0.82, avec un écart-type de 0.29, et sur 130 individus consommant de l'huile d'olive, pour lesquels on observe une moyenne de -0.93, avec un écart-type de 0.31. Calculer la p-valeur du test permettant de décider si l'amélioration est significative. Au seuil de 0.05, que concluez-vous ?

   Il s'agit d'un test de comparaison des espérances pour de grands échantillons. La statistique de test est :

   $$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} \ ,$$

   qui suit la loi $\mathcal{N}(0,1)$ sous l'hypothèse $\mathcal{H}_0$. Or elle prend la valeur 2.8366. La p-valeur est la probabilité pour une variable de loi $\mathcal{N}(0,1)$ de dépasser 2.8366, soit 0.0023. À tout seuil inférieur à 0.23% (et bien sûr en particulier aux seuils de 5% et 1%), on rejette $\mathcal{H}_0$, donc on décide que l'huile d'olive améliore de manière significative le risque cardio-vasculaire.

Exercice 3.3.2. On étudie l'activité de l'enzyme sérique PDE, en fonction de différents facteurs dans l'espèce humaine. Les résultats sont exprimés en unité internationale par litre de sérum. Chez deux groupes de femmes, enceintes ou non, on obtient les résultats suivants :

non enceinte	1.5	1.6	1.4	2.9	2.2	1.8	2.7	1.9
enceinte	4.2	5.5	4.6	5.4	3.9	5.4	2.7	3.9

non enceinte	2.2	2.8	2.1	1.8	3.7	1.8	2.1
enceinte	4.1	4.1	4.6	3.9	3.5		

(Indications numériques : $\sum x_i = 32.5$, $\sum x_i^2 = 75.83$, $\sum y_i = 55.8$, $\sum y_i^2 = 247.32$).

1. Préciser les hypothèses de modélisation.
2. Tester l'hypothèse d'égalité des variances au seuil de 5%.
3. Peut-on affirmer que l'activité de l'enzyme sérique PDE est significativement différente chez les femmes enceintes et chez les femmes non enceintes ?
4. Peut-on affirmer que l'activité de l'enzyme sérique PDE est significativement supérieure chez les femmes enceintes ?

Exercice 3.3.3. Les QI de 9 enfants d'un quartier d'une grande ville ont pour moyenne empirique 107 et écart-type empirique 10. Les QI de 12 enfants d'un autre quartier ont pour moyenne empirique 112 et écart-type empirique 9.

1. Préciser les hypothèses de modélisation.
2. Tester l'égalité des variances au seuil de 5%.
3. Les QI des enfants du deuxième quartier sont-il significativement supérieurs en moyenne à ceux des enfants du premier quartier ? Donner un encadrement de la p-valeur du test correspondant.

Exercice 3.3.4. Les tensions maximales des muscles gastrocnémiens (exprimées en g) de la grenouille varient selon que ces muscles sont normaux ou dénervés. Lors d'une expérience faite sur 9 grenouilles, on a relevé les mesures suivantes :

Muscles normaux	75	96	32	41	50	39	59	45	30
Muscles dénervés	53	67	32	29	35	27	37	30	21

1. Préciser les hypothèses de modélisation.
2. Tester l'hypothèse d'égalité des variances au seuil de 5%.
3. Au seuil de 5%, peut-on affirmer que la tension maximale moyenne est différente pour les muscles normaux et pour les muscles dénervés ? Donner un encadrement de la p-valeur de ce test.

Exercice 3.3.5. Au cours d'une étude destinée à comparer diverses méthodes d'échantillonnage de sols forestiers, on a mesuré les teneurs en K_20, d'une part pour 20 échantillons de terre prélevés individuellement, et d'autre part pour 10 échantillons mélangés obtenus chacun à partir de 25 terres différentes. On a obtenu pour les échantillons individuels :

$$\sum x_i = 259.2 \quad \text{et} \quad \sum x_i^2 = 3662.08 \ ,$$

et pour les échantillons mélangés :

$$\sum y_i = 109.2 \quad \text{et} \quad \sum y_i^2 = 1200.8 \ .$$

On s'attend à ce que les deux méthodes d'échantillonnage donnent des variances très différentes. Justifier cela intuitivement et vérifiez le par le test de Fisher.

Exercice 3.3.6. Pour déterminer le poids moyen d'un épi de blé appartenant à deux variétés, on procède à 9 pesées pour chaque variété. On donne les moyennes et variances empiriques des deux échantillons :

$$\overline{x} = 170.7 \ ; \quad \overline{y} = 168.5 \ ; \quad s_x^2 = 432.90 \ ; \quad s_y^2 = 182.70 \ .$$

1. Préciser les hypothèses de modélisation.
2. Tester au seuil de 5% l'hypothèse d'égalité des variances.
3. Donner un encadrement de la p-valeur pour le test permettant de décider si les deux variétés sont significativement différentes. Quelle est votre conclusion ?

Exercice 3.3.7. Dans une coopérative agricole, on désire tester l'effet d'un engrais sur la production de blé. Pour cela, on choisit 200 lots de terrain de même superficie. La moitié de ces lots est traitée avec l'engrais, et l'autre ne l'est pas. Les récoltes en tonnes obtenues pour les 100 lots non traités donnent $\sum x_i = 61.6$, $\sum x_i^2 = 292.18$ et pour les lots traités $\sum y_i = 66.8$, $\sum y_i^2 = 343.48$.

Tester l'hypothèse "l'engrais n'est pas efficace" contre "l'engrais est efficace" aux seuils 0.01 et 0.05.

Exercice 3.3.8. Dans un échantillon de 300 personnes, prélevé dans la population d'une ville A, il y en a 36 qui fument au moins deux paquets de cigarettes par jour. Dans une autre ville B et pour un échantillon de 100 personnes, on trouve 8 personnes qui fument au moins deux paquets de cigarettes par jour. On veut tester $\mathcal{H}_0$: "il n'y a aucune différence entre les deux villes" contre $\mathcal{H}_1$: "il y a plus de personnes qui fument au moins deux paquets de cigarettes par jour dans la ville A que dans la ville B".

1. On note p_A (resp. p_B) la proportion d'individus qui fument au moins deux paquets de cigarettes dans la ville A (resp. B). Quelles variables proposez-vous pour modéliser le problème ? Donner leurs espérances et leurs variances en fonction de p_A et p_B.

2. Quel test proposez-vous pour décider s'il y a significativement plus de gros fumeurs dans la ville A que dans la ville B ?
3. Donnez la p-valeur de ce test pour les données de l'énoncé. Quelle est votre conclusion ?

Exercice 3.3.9. Soit p_A la probabilité de guérison d'une maladie donnée grâce à un traitement A. Un groupe de 50 malades est soumis à ce traitement et 28 guérissent. Un autre traitement B permet de soigner cette maladie, avec probabilité p_B. Sur 60 malades soumis à ce nouveau traitement, 38 guérissent.

1. Quel test proposez-vous pour décider si le nouveau traitement est meilleur que l'ancien ?
2. Donnez la p-valeur de ce test pour les données de l'énoncé. Quelle est votre conclusion ?

3.4 Test du khi-deux d'ajustement

On note r le nombre de classes. Pour $i = 1, \ldots, r$, on note n_i l'effectif *observé* de la classe i, et np_i son effectif *théorique*.

- La statistique du test du khi-deux est :

$$T = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \, .$$

- Sous l'hypothèse nulle où le modèle théorique est le bon, T suit la loi du khi-deux de paramètre $d = r-1-k$:
  - ⋆ r est le nombre de classes,
  - ⋆ k est le nombre de paramètres qui ont été estimés à partir des données pour établir la distribution théorique.
- Le test s'applique à un grand échantillon ($n \geqslant 50$). Les effectifs théoriques de chaque classe doivent être assez grands ($np_i \geqslant 8$). On peut être amené à regrouper des classes pour satisfaire la seconde condition.

Exercice 3.4.1. On effectue le croisement entre des pois à fleurs blanches et des pois à fleurs rouges. On obtient en deuxième génération sur 600 plantes les effectifs suivants :

Phénotype	Rouge	Rose	Blanc
Effectif	141	325	134

On a formé ensuite 150 bouquets de 4 plantes, parmi lesquels on a observé le nombre de plantes à fleurs blanches. Les effectifs ont été les suivants.

Nbre. fleurs blanches	0	1	2	3	4
Effectif	53	68	23	4	2

1. Donner les proportions théoriques de la répartition mendélienne pour les trois couleurs. Calculer la statistique de test pour le test du khi-deux. Donner un encadrement de la p-valeur. Quelle est votre conclusion ?

   Notons R l'allèle induisant la couleur rouge et B l'allèle induisant la couleur blanche. On suppose que les phénotypes "fleurs rouges", "fleurs roses" et "fleurs blanches" correspondent respectivement aux génotypes RR, RB et BB. Si on croise deux individus de génotypes respectifs RR et BB, on obtient forcément des individus de génotype RB à la première génération. À la seconde génération, on obtiendra théoriquement un quart de génotypes RR, la moitié de génotypes RB et un quart de génotypes BB ; on devrait donc observer théoriquement un quart de plantes à fleurs rouges, la moitié à fleurs roses, et un quart à fleurs blanches. Les effectifs théoriques correspondants sont 150, 300, 150.
   La statistique de test du khi-deux prend la valeur :

$$T = \frac{(141-150)^2}{150} + \frac{(325-300)^2}{300} + \frac{(134-150)^2}{150} = 4.33 \; .$$

   Cette valeur doit être comparée aux quantiles de la loi du khi-deux de paramètre $3-1=2$. *La p-valeur est la probabilité qu'une variable suivant la loi* $\mathcal{X}^2(2)$ *dépasse* 4.33. *D'après la table, elle est comprise entre* 0.1 *et* 0.2. *La valeur exacte est* 0.1147. *On accepte l'hypothèse d'adéquation de la loi observée avec la loi théorique.*

2. Quel modèle théorique proposez-vous pour le nombre de plantes à fleurs blanches sur un bouquet de 4 ? Effectuez un regroupement en classes approprié. Calculer la statistique de test pour le test du khi-deux. Donner un encadrement de la p-valeur. Quelle est votre conclusion ?

   Si les bouquets sont formés au hasard, la loi du nombre de plantes à fleurs blanches sur un bouquet de 4 *est la loi binomiale de paramètres* 4 *(le nombre total de plantes) et* 1/4 *(la proportion théorique de plantes à fleurs blanches). Pour* $i = 0, \ldots, 4$, *l'effectif théorique du nombre de bouquets avec k plantes à fleurs blanches est :*

$$np_i = 150 \binom{4}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{4-k} \; .$$

Nbre. fleurs blanches	0	1	2	3	4
Effectif observé	53	68	23	4	2
Effectif théorique	47.46	63.28	31.64	7.03	0.59

   Pour atteindre un effectif théorique au moins égal à 8 *dans chaque classe, on peut regrouper les 3 dernières classes.*

Nbre. fleurs blanches	0	1	2, 3, 4
Effectif observé	53	68	29
Effectif théorique	47.46	63.28	39.26

La statistique de test du khi-deux prend la valeur 3.6786. *La p-valeur est la probabilité qu'une variable suivant la loi* $\mathcal{X}^2(2)$ *dépasse* 3.6786. *Sur la table, la p-valeur est entre* 0.1 *et* 0.2. *La valeur exacte est* 0.1589. *On accepte l'hypothèse d'adéquation de la loi observée avec la loi théorique.*

3. Soit $\widehat{p}$ la proportion observée de plantes à fleurs blanches. Pour les bouquets de 4 plantes, tester l'adéquation de la distribution observée avec la loi binomiale $\mathcal{B}(4, \widehat{p})$: calculer la statistique de test et donner un encadrement de la p-valeur.

   Le nombre total de plantes à fleurs blanches est de 134, *leur proportion est donc de* $\widehat{p} = 134/600 \simeq 0.2233$. *On calcule maintenant les effectifs théoriques par rapport à la loi binomiale* $\mathcal{B}(4, \widehat{p})$.

Nbre. fleurs blanches	0	1	2, 3, 4
Effectif observé	53	68	29
Effectif théorique	54.59	62.78	32.64

   La statistique de test du khi-deux prend la valeur 0.8855. *Puisqu'on a estimé un paramètre pour établir la distribution théorique, le paramètre de la loi du khi-deux est* $3 - 1 - 1 = 1$. *Sur la table, la p-valeur est entre* 0.3 *et* 0.4, *la valeur exacte est* 0.3467. *On accepte l'hypothèse d'adéquation de la loi observée avec la loi théorique.*

Exercice 3.4.2. Voici le tableau des fréquences en France des principaux groupes sanguins :

Groupe / Facteur	O	A	B	AB
Rhésus +	0.370	0.381	0.062	0.028
Rhésus –	0.070	0.072	0.012	0.005

Le centre de transfusion sanguine de Pau a observé la répartition suivante sur 5000 donneurs.

Groupe / Facteur	O	A	B	AB
Rhésus +	2291	1631	282	79
Rhésus –	325	332	48	12

On souhaite répondre statistiquement aux questions ci-dessous. Dans chaque cas, on écrira le tableau des distributions observée et théorique, on calculera la valeur prise par la statistique du test, on donnera un encadrement de la p-valeur, et on concluera.

1. La répartition paloise des 8 types groupe-rhésus est-elle différente de la répartition nationale ?
2. La répartition paloise des rhésus est-elle différente de la répartition nationale ?

3. Parmi les individus de groupe O, la répartition paloise des rhésus est-elle différente de la répartition nationale ?
4. Parmi les individus de rhésus positif, la répartition paloise des groupes est-elle différente de la répartition nationale ?
5. Parmi les individus de rhésus négatif, la répartition paloise des groupes est-elle différente de la répartition nationale ?

Exercice 3.4.3. On a demandé à 162 étudiant(e)s d'estimer le temps mensuel en heures qu'ils passent à préparer la cuisine :

Heures	[0 ; 5[	[5 ; 10[	[10 ; 15[	$\geqslant 15$
Étudiants	63	49	19	31

Des études antérieures dans l'ensemble de la population ont permis d'établir la répartition suivante :

Heures	[0 ; 5[	[5 ; 10[	[10 ; 15[	$\geqslant 15$
Proportion	40%	35%	15%	10%

Tester l'adéquation de la distribution observée avec la distribution connue. Donner un encadrement de la p-valeur. Quelle est votre conclusion ?

Exercice 3.4.4. On s'intéresse au temps de sommeil d'un enfant de douze ans et sur un échantillon de taille $n = 50$ on a observé les temps de sommeil (exprimés en heures). On donne $\sum x_i = 424$ et $\sum x_i^2 = 3828$, ainsi que la répartition en classes suivante :

Class	$\leqslant 8$	]8 ; 9]	]9 ; 10]	> 10
Number	19	12	9	10

1. Il est généralement admis que le temps de sommeil d'un enfant de cet âge suit la loi normale $\mathcal{N}(9, 3)$. Réaliser le test d'adéquation de la distribution observée avec cette hypothèse théorique. Donner la valeur prise par la statistique de test, un encadrement de la p-valeur et votre conclusion.
2. Calculer la moyenne empirique $\overline{x}$ et la variance empirique s^2. Reprendre la question précédente en remplaçant la loi $\mathcal{N}(9, 3)$ par la loi $\mathcal{N}(\overline{x}, s^2)$.

Exercice 3.4.5. Une étude biométrique faite sur la longueur d'œufs de coucou a donné les résultats suivants. On donne : $n = 152$, $\sum x_i = 6200$, $\sum x_i^2 = 255200$, ainsi que la répartition en classes suivante :

classe	< 32	[32; 34[	[34; 36[	[36; 38[	[38; 40[	[40; 42[	[42; 44[	[44; 46[	[46; 48[	$\geqslant 48$
effectif	2	7	6	18	25	40	23	20	6	5

1. Des études antérieures avaient montré que les longueurs d'œufs de coucou suivent une loi normale d'espérance 40 et d'écart-type 4. Réaliser le test d'adéquation de la distribution observée avec cette hypothèse théorique. Donner la valeur prise par la statistique de test, un encadrement de la p-valeur et votre conclusion.
2. Calculer la moyenne empirique $\overline{x}$ et la variance empirique s^2. Reprendre la question précédente en remplaçant la loi $\mathcal{N}(40, 4^2)$ par la loi $\mathcal{N}(\overline{x}, s^2)$.

3.5 Test du khi-deux de contingence

C'est un cas particulier du test du khi-deux d'ajustement, qui permet de tester l'indépendance de deux caractères discrets.

- La *table de contingence* présente les *effectifs conjoints*. À la ligne i, colonne j, on trouve n_{ij}, qui est le nombre d'individus dans la classe i pour le premier caractère et dans la classe j pour le second. Si le nombre de modalités des deux caractères sont r et s, la table a r lignes et s colonnes.
- Les *effectifs marginaux* sont les sommes par ligne ou par colonne de la table de contingence ; $n_{i\bullet} = \sum_j n_{ij}$ est le nombre total d'individus dans la classe i pour le premier caractère ; $n_{\bullet j} = \sum_i n_{ij}$ est le nombre total d'individus dans la classe j pour le second caractère. Le nombre total d'individus est $n = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$.
- La statistique du test est :

$$T = n\left(-1 + \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}}\right).$$

- Sous l'hypothèse nulle où les deux caractères sont indépendants, T suit la loi du khi-deux de paramètre $d = (r-1)(s-1)$.

Exercice 3.5.1. Le centre de transfusion sanguine de Pau a observé la répartition suivante sur 5000 donneurs.

Groupe / Facteur	O	A	B	AB
Rhésus +	2291	1631	282	79
Rhésus –	325	332	48	12

1. Compléter la table de contingence par les effectifs marginaux.

   L'énoncé donne les effectifs conjoints. Il suffit de les sommer pour avoir les effectifs marginaux.

Groupe / Facteur	*O*	*A*	*B*	*AB*	*Total*
Rhésus +	2291	1631	282	79	4283
Rhésus –	325	332	48	12	717
Total	2616	1963	330	91	5000

2. Calculer la valeur prise par la statistique du test du khi-deux de contingence.

   On calcule :

$$T = 5000\left(-1 + \frac{2291^2}{2616\times 4283} + \cdots + \frac{12^2}{717\times 91}\right) = 18.5104\,.$$

3. Au seuil de 1% que concluez-vous ?

   Sous l'hypothèse d'indépendance, la statistique de test suit la loi de khi-deux de paramètre $(4-1)(2-1) = 3$. *Le quantile d'ordre* 0.99 *de cette loi est* 11.3449. *Comme* 18.5104 *est supérieur, on conclut qu'il y a dépendance entre le groupe sanguin et le rhésus, au vu de ces données. La p-valeur exacte est de* 0.000345.

Exercice 3.5.2. Les résultats observés de l'évolution d'une certaine maladie à la suite de l'emploi de l'un ou l'autre des traitements A et B pour 1000 patients figurent dans le tableau ci-dessous :

Effet / Traitement	Guérison	Amélioration	Etat stationnaire
A	280	210	110
B	220	90	90

1. Compléter cette table de contingence.
2. Calculer la valeur prise par la statistique du khi-deux de contingence pour cette table.
3. Donner un encadrement de la p-valeur pour le test du khi-deux de contingence. Diriez-vous que les traitements A et B sont significativement différents quant à leur efficacité ?

Exercice 3.5.3. On a observé pendant dix ans 240 individus. Parmi-ceux-ci :
- 110 ont consommé de l'huile d'arachide
- 25 ont consommé de l'huile d'olive et ont eu des problèmes cardio-vasculaires
- 78 ont consommé de l'huile d'arachide et n'ont eu aucun problème.

1. Écrire la table de contingence correspondant à ces observations.
2. Calculer la valeur prise par la statistique du khi-deux de contingence pour ce tableau.
3. Donner un encadrement de la p-valeur pour le test du khi-deux de contingence. Diriez-vous que le risque cardio-vasculaire est indépendant du type d'huile consommée ?

Exercice 3.5.4. L'observation d'un couple (X, Y) de variables physiologiques pour les 100 individus d'une population a conduit, après choix de deux classes pour X et de trois classes pour Y, à la table de contingence suivante :

X \ Y	1	2	3	Total
1	4	11	7	22
2	16	39	23	78
Total	20	50	30	100

1. Calculer la valeur prise par la statistique du khi-deux de contingence.
2. Donner un encadrement de la p-valeur pour le test du khi-deux de contingence. Quelle est votre conclusion ?

Exercice 3.5.5. À la suite du même traitement d'une certaine maladie, pour 70 patients jeunes, on a observé 40 cas d'amélioration et pour 100 patients âgés, on en a observé 50.

1. Écrire la table de contingence correspondant à ces observations.
2. Calculer la valeur prise par la statistique du khi-deux de contingence.
3. Donner un encadrement de la p-valeur pour le test du khi-deux de contingence. Diriez-vous que l'effet du traitement dépend de l'âge du patient ?

Exercice 3.5.6. On considère la table de contingence suivante concernant 592 femmes réparties selon la couleur de leurs yeux et celle de leurs cheveux :

Cheveux / Yeux	Bruns	Châtains	Roux	Blonds
Marrons	68	119	26	7
Noisette	15	54	14	10
Verts	5	29	14	16
Bleus	20	84	17	94

1. Compléter cette table de contingence.
2. Calculer la valeur prise par la statistique du khi-deux de contingence.
3. Donner un encadrement de la p-valeur pour le test du khi-deux de contingence. Diriez-vous qu'il y a indépendance entre la couleur des yeux et celle des cheveux ?

4 Régression linéaire

4.1 Droite de régression et prédiction ponctuelle

Les données sont n couples de réels. La première coordonnée est un caractère considéré comme *déterministe* et *explicatif*. Le second est considéré comme *aléatoire* et *à expliquer*. On calcule :

- la moyenne du caractère explicatif : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
- la moyenne du caractère à expliquer : $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$
- la variance du caractère explicatif : $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$
- la variance du caractère à expliquer : $s_y^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$
- la *covariance* des deux caractères : $c_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$
- le *coefficient de corrélation* : $r_{xy} = \dfrac{c_{xy}}{\sqrt{s_x^2\, s_y^2}}$.
- la *pente* de la droite de régression linéaire : $\widehat{a} = \dfrac{c_{xy}}{s_x^2}$
- l'*ordonnée à l'origine* : $\widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$
- la *variance estimée* : $\widehat{\sigma}^2 = \dfrac{n}{n-2} s_y^2(1 - r_{xy}^2)$
- la *prédiction* d'une ordonnée pour une abscisse x_* donnée : $y_* = \widehat{a}\, x_* + \widehat{b}$.

Exercice 4.1.1. Pour mesurer la dépendance entre l'âge et le risque cardio-vasculaire, on a observé 12 patients, pour lesquels on dispose de l'âge en années (variable X), et du logarithme du dosage en d-dimères (variable Y). On donne les quantités suivantes :

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Calculer le coefficient de corrélation linéaire de X et Y.

   On trouve :

   $$\overline{x} = 49.667 \; ; \quad \overline{y} = -0.43333 \; ; \quad s_x^2 = 236.139 \; ; \quad s_y^2 = 0.17056 \; ;$$

   $$c_{xy} = 5.8072 \; ; \quad r_{xy} = 0.91506 \; .$$

   Le fait que r_{xy} soit proche de 1 indique une forte corrélation.

2. Calculer l'équation de la droite de régression linéaire de Y sur X.

   On trouve :

   $$\widehat{a} = 0.02459 \; ; \quad \widehat{b} = -1.6548 \; .$$

L'équation de la droite de régression linéaire est $y = 0.02459\,x - 1.6548$. *Elle est croissante* ($a > 0$) *car la corrélation est positive : le logarithme du dosage en d-dimères tend à augmenter avec l'âge.*

3. Calculer la variance estimée de la régression.

   On trouve $\widehat{\sigma}^2 = 0.0333$.

4. Quelle valeur de Y prévoyez-vous pour un individu de 60 ans ?

   La valeur prédite pour $x_* = 60$ *est* $y_* = 0.02459 \times 60 - 1.6548 = -0.1792$.

Exercice 4.1.2. On étudie la pollution de l'air dans 41 villes américaines par la variable Y, mesurant le volume de SO_2 dans l'air en micro-grammes par m^3, en fonction de la température moyenne annuelle X, exprimée en degrés Fahrenheit. On donne les résultats numériques suivants :

$$\sum x_i = 2286\ ;\ \sum y_i = 1232\ ;\ \sum x_i^2 = 129549\ ;\ \sum y_i^2 = 59050\ ;\ \sum x_i y_i = 74598\ .$$

1. Calculer le coefficient de corrélation linéaire de X et Y.
2. Donner l'équation de la droite de régression de Y par rapport à X.
3. Quelle valeur de Y prédisez-vous pour une ville où la température moyenne est de 60°F ?

Exercice 4.1.3. Dans le cadre de travaux de recherche sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne (variable Y en degrés Celsius) ainsi que l'altitude (variable X en mètres) de chaque station données dans le tableau ci-dessous :

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
température	7.4	6	4.5	3.8	2.9	1.9	1	−1.2	−1.5	−4.5

On donne :

$$\sum x_i = 19690;\ \sum y_i = 20.3;\ \sum x_i^2 = 42925500;\ \sum y_i^2 = 162.41;\ \sum x_i y_i = 17671\ .$$

1. Calculer le coefficient de corrélation linéaire.
2. Calculer les estimations des paramètres a, b et σ^2 pour la régression linéaire de Y sur X.
3. Quelle température moyenne prévoyez-vous à 1100 m ?

Exercice 4.1.4. On souhaite vérifier si la prise de poids d'un jeune mouton en un an (variable Y en kilogrammes) dépend de son poids initial (variable X également en kilogrammes). Sur 10 moutons, on donne les résultats suivants :

$$\sum x_i = 406\ ;\ \sum y_i = 423\ ;\ \sum x_i^2 = 16570\ ;\ \sum y_i^2 = 18057\ ;\ \sum x_i y_i = 17280\ .$$

1. Calculer le coefficient de corrélation linéaire.
2. Estimer les paramètres a, b et σ^2 pour la régression linéaire de Y sur X.
3. Selon ce modèle combien un mouton de poids initial 50 kg devrait-il prendre de poids ? Même question pour un mouton de 30 kg.

Exercice 4.1.5. Le volume d'air expiré Y est une mesure standard du fonctionnement pulmonaire. Pour identifier une population possédant un fonctionnement pulmonaire anormal, il faut établir un modèle pour le volume d'air expiré dans une population normale. Pour cela, on mesure le volume Y en litres et la taille X en centimètres sur 12 garçons âgés de 10 à 15 ans.

On obtient les résumés numériques suivants :

$$\sum x_i = 1872 \; ; \; \sum y_i = 32.3 \; ; \; \sum x_i^2 = 294320 \; ; \; \sum y_i^2 = 93.11 \; ; \; \sum x_i y_i = 5156.20 \; .$$

1. Calculer le coefficient de corrélation.
2. Calculer les estimations des coefficients de la droite de régression linéaire de Y sur X et de la variance.
3. Quel volume d'air devrait expirer un garçon mesurant 1.60 m ?

Exercice 4.1.6. On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour faire une régression linéaire, on effectue un changement de variable en posant $Y = \ln(H)$ et $X = \ln(D)$. Voici les mesures faites sur 5 arbres :

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

On donne :

$$\sum x_i = -4.71 \; ; \; \sum y_i = 12.12 \; ; \; \sum x_i^2 = 5.4095 \; ;$$
$$\sum y_i^2 = 29.5282 \; ; \; \sum x_i y_i = -11.0458 \; .$$

1. Donner le coefficient de corrélation linéaire entre X et Y.
2. Donner l'équation de la droite de régression de Y par rapport à X.
3. Donner la hauteur prévue d'un arbre de diamètre 0.7.

4.2 Intervalles de confiance et de prédiction

Les intervalles donnés dans ce qui suit sont de niveau $1-\alpha$, et t_α désigne le quantile d'ordre $1-\alpha/2$ de la loi de Student $\mathcal{T}(n-2)$.

- Intervalle de confiance pour la pente a :

$$\left[\widehat{a} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2}{n s_x^2}} \right] \; .$$

- Intervalle de confiance pour $ax_* + b$:

$$\left[\widehat{a}x_* + \widehat{b} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2(s_x^2 + (x_* - \overline{x})^2)}{ns_x^2}} \right] .$$

- Intervalle de *prédiction* pour $Y_* = ax_* + b + E$:

$$\left[\widehat{a}x_* + \widehat{b} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}{ns_x^2}} \right] .$$

Exercice 4.2.1. Pour mesurer la dépendance entre l'âge et le risque cardio-vasculaire, on a observé 12 patients, pour lesquels on dispose de l'âge en années (variable X), et du logarithme du dosage en d-dimères (variable Y). On donne les quantités suivantes :

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Donner un intervalle de confiance de niveau 0.99 pour la pente de la droite de régression linéaire.

   Le quantile d'ordre 0.995 *de la loi de Student de paramètre* $12 - 2 = 10$ *est* 3.169. *L'intervalle de confiance est* $[0.0137\,;\,0.0355]$.

2. Donner un intervalle de confiance de niveau 0.99 pour l'ordonnée à l'origine de la droite de régression linéaire.

   On obtient un intervalle de confiance pour b *en posant* $x_* = 0$ *dans la formule donnant l'intervalle de confiance pour* $ax_* + b$. *L'intervalle cherché est* $[-2.2195\,;\,-1.0900]$.

3. Donner un intervalle de confiance de niveau 0.99 pour la valeur moyenne de Y parmi les individus de 60 ans.

   On cherche un intervalle de confiance pour $ax_* + b$, *avec* $x_* = 60$. *L'intervalle est* $[-0.380\,;0.022\,]$.

4. Donner un intervalle de prédiction de niveau 0.99 pour la valeur de Y chez un individu de 60 ans particulier.

   On cherche un intervalle de prédiction pour $Y_* = ax_* + b + E$, *avec* $x_* = 60$. *L'intervalle est* $[-0.791\,;0.433\,]$. *Attention à ne pas confondre :*
   - estimer *la valeur* moyenne *des dosages en d-dimères chez* les *individus de* 60 *ans*
   - prédire *la valeur du dosage en d-dimères chez* un *individu de* 60 *ans en particulier.*

   Dans le second cas, l'intervalle est forcément plus large que dans le premier.

Exercice 4.2.2. On étudie la pollution de l'air dans 41 villes américaines par la variable Y, mesurant le volume de SO_2 dans l'air en micro-grammes par m^3, en fonction de

la température moyenne annuelle X, exprimée en degrés Fahrenheit. On donne les résultats numériques suivants :

$$\sum x_i = 2286,\ \sum y_i = 1232,\ \sum x_i^2 = 129549,\ \sum y_i^2 = 59050,\ \sum x_i y_i = 74598\ .$$

1. Donner un intervalle de confiance de niveau 0.95 pour la pente et l'ordonnée à l'origine de la droite de régression.
2. Donner un intervalle de confiance de niveau 0.95 pour la valeur moyenne de Y dans les villes où la température est de 60°F.
3. Donner un intervalle de prédiction de niveau 0.95 pour la valeur de Y dans une ville où la température est de 60°F.

Exercice 4.2.3. Dans le cadre de travaux de recherche sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne (en degrés Celsius) ainsi que l'altitude (en mètres) de chaque station sont données dans le tableau ci-dessous :

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
température	7.4	6	4.5	3.8	2.9	1.9	1	−1.2	−1.5	−4.5

On donne :

$$\sum x_i = 19690;\ \sum y_i = 20.3;\ \sum x_i^2 = 42925500;\ \sum y_i^2 = 162.41;\ \sum x_i y_i = 17671\ .$$

1. Donner un intervalle de confiance de niveau 0.95 pour la pente et l'ordonnée à l'origine de la droite de régression.
2. Donner un intervalle de confiance de niveau 0.95 pour la température moyenne à 1100 m.
3. Donner un intervalle de prédiction de niveau 0.95 pour la température moyenne d'un endroit situé à 1100 m.

Exercice 4.2.4. On souhaite vérifier si la prise de poids d'un jeune mouton en un an (variable Y en kilogrammes) dépend de son poids initial (variable X également en kilogrammes). Sur 10 moutons, on donne les résultats suivants :

$$\sum x_i = 406\ ;\ \sum y_i = 423\ ;\ \sum x_i^2 = 16570\ ;\ \sum y_i^2 = 18057\ ;\ \sum x_i y_i = 17280\ .$$

1. Donner un intervalle de confiance de niveau 0.99 pour la pente et l'ordonnée à l'origine de la droite de régression.
2. Donner un intervalle de confiance de niveau 0.99 pour le gain de poids moyen des moutons de 30 kg.
3. Donner un intervalle de prédiction de niveau 0.99 pour le gain de poids d'un mouton de 30 kg.

Exercice 4.2.5. Le volume d'air expiré Y est une mesure standard du fonctionnement pulmonaire. Pour identifier une population possédant un fonctionnement pulmonaire anormal, il faut établir un modèle pour le volume d'air expiré dans une population normale. Pour cela, on mesure le volume Y en litres et la taille X en centimètres sur 12 garçons âgés de 10 à 15 ans.

On obtient les résumés numériques suivants :

$$\sum x_i = 1872 \; ; \; \sum y_i = 32.3 \; ; \; \sum x_i^2 = 294320 \; ; \; \sum y_i^2 = 93.11 \; ; \; \sum x_i y_i = 5156.20 \; .$$

1. Donner un intervalle de confiance de niveau 0.99 pour la pente et l'ordonnée à l'origine de la droite de régression.
2. Donner un intervalle de confiance de niveau 0.99 pour le volume d'air expiré en moyenne par les garçons de 1.60 m.
3. Donner un intervalle de prédiction de niveau 0.99 pour le volume d'air expiré par un garçon de 1.60 m.

Exercice 4.2.6. On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour faire une régression linéaire, on effectue un changement de variable en posant $Y = \ln(H)$ et $X = \ln(D)$. Voici les mesures faites sur 5 arbres.

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

On donne :

$$\sum x_i = -4.71, \; \sum y_i = 12.12, \; \sum x_i^2 = 5.4095,$$
$$\sum y_i^2 = 29.5282, \; \sum x_i y_i = -11.0458.$$

1. Donner un intervalle de confiance de niveau 0.95 pour la pente et l'ordonnée à l'origine de la droite de régression.
2. Donner un intervalle de confiance de niveau 0.95 pour la hauteur moyenne des arbres de diamètre 0.7.
3. Donner un intervalle de prédiction de niveau 0.95 pour la hauteur d'un arbre de diamètre 0.7.

4.3 Tests sur une régression

Sous l'hypothèse $\mathcal{H}_0$, le modèle est $Y = ax+b+E$, où E suit la loi normale $\mathcal{N}(0, \sigma^2)$. Les paramètres a, b et σ^2 sont inconnus. On les estime par $\widehat{a}$, $\widehat{b}$ et $\widehat{\sigma}^2$. Pour tester des valeurs particulières, on utilise les résultats suivants, donnant la loi des statistiques de test sous $\mathcal{H}_0$.

- $\sqrt{\dfrac{n s_x^2}{\widehat{\sigma}^2}} \left(\widehat{a} - a \right)$ suit la loi de Student $\mathcal{T}(n-2)$.

- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2(s_x^2+(x_*-\overline{x})^2)}}\left(\widehat{a}x_*+\widehat{b}-ax_*-b\right)$ suit la loi de Student $\mathcal{T}(n-2)$.
- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2((n+1)s_x^2+(x_*-\overline{x})^2)}}\left(Y_*-\widehat{a}x_*-\widehat{b}\right)$ suit la loi de Student $\mathcal{T}(n-2)$.
- $(n-2)\dfrac{\widehat{\sigma}^2}{\sigma^2}$ suit la loi du khi-deux $\mathcal{X}^2(n-2)$.

Le test de *pertinence* ou de validité de la régression consiste à tester $\mathcal{H}_0 : a = 0$ contre $\mathcal{H}_1 : a \neq 0$, en utilisant le premier des résultats précédents. On conclut que la régression est pertinente en rejetant $\mathcal{H}_0$.

Exercice 4.3.1. Pour mesurer la dépendance entre l'âge et le risque cardio-vasculaire, on a observé 12 patients, pour lesquels on dispose de l'âge en années (variable X), et du logarithme du dosage en d-dimères (variable Y). On donne les quantités suivantes :

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Tester la pertinence de la régression au seuil de 1%.

   Il s'agit d'un test bilatéral de $\mathcal{H}_0 : a = 0$ *contre* $\mathcal{H}_1 : a \neq 0$. *La statistique de test est :*

   $$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2}}\,\widehat{a}\,.$$

   Sous l'hypothèse $\mathcal{H}_0$, T *suit la loi de Student de paramètre* 10. *La règle de décision est :*

   $$\textit{Rejet de } \mathcal{H}_0 \implies T \notin [-t_\alpha\,;\,+t_\alpha]\,,$$

   où t_α *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student* $\mathcal{T}(10)$, *à savoir* 3.169. *Ici, la valeur prise par* T *est* 7.177. *On rejette* $\mathcal{H}_0$, *donc on déclare que la régression est pertinente.*

2. Des études précédentes avaient donné une dépendance linéaire entre l'âge et le dosage en d-dimères sous la forme $Y = 0.02x - 2$. Tester au seuil de 1% si les valeurs de a et b précédemment admises peuvent être conservées.

   Nous testons d'abord $\mathcal{H}_0 : a = 0.02$ *contre* $\mathcal{H}_1 : a \neq 0.02$. *La statistique de test est :*

   $$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2}}\,(\widehat{a}-0.02)\,.$$

   Elle prend la valeur 1.341 *qui est dans l'intervalle* $[-3.169\,;\,+3.169]$. *Donc on accepte* $\mathcal{H}_0$ *(on déclare que la valeur estimée de* a *n'est pas significativement éloignée de* 0.02*).*

   Nous testons maintenant $\mathcal{H}_0 : b = -2$ *contre* $\mathcal{H}_1 : b \neq -2$. *La statistique de test est :*

   $$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2(s_x^2+(0-\overline{x})^2)}}(\widehat{b}-(-2))\,.$$

Elle prend la valeur 1.935 *qui est dans l'intervalle* $[-3.169\,;\,+3.169]$. *Donc on accepte* $\mathcal{H}_0$ *(on déclare que la valeur estimée de* b *n'est pas significativement éloignée de* -2*).*

Au total les deux tests acceptent les valeurs antérieures de a *et* b.

3. Un patient de 60 ans présente une valeur de Y égale à 0.14 : est-ce inquiétant ?

   Nous testons ici une valeur de $Y_* = ax_* + b + E$, *avec* $x_* = 60$. *La statistique de test est :*

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}}(Y_* - \widehat{a}x_* - \widehat{b}) \ .$$

   Elle prend la valeur 5.028. *Cette valeur est supérieure au quantile d'ordre* 0.0005 *de la loi* $\mathcal{T}(10)$, *elle est donc anormalement élevée (par rapport aux données fournies).*

4. Tester au seuil de 1% l'hypothèse $\mathcal{H}_0 \ : \sigma^2 = 0.03$ contre $\mathcal{H}_1 \ : \sigma^2 > 0.03$.

   La statistique de test est :

$$10\frac{\widehat{\sigma}^2}{0.03} \ .$$

   Sous l'hypothèse $\mathcal{H}_0$, T *suit la loi du khi-deux de paramètre* 10. *Au seuil de* 1%, *on rejette les valeurs suprérieures au quantile d'ordre* 0.99 *de la loi* $\mathcal{X}^2(10)$, *à savoir* 23.21. *Ici,* T *prend la valeur* 11.09, *donc on accepte* $\mathcal{H}_0$.

Exercice 4.3.2. On étudie la pollution de l'air dans 41 villes américaines par la variable Y, mesurant le volume de SO_2 dans l'air en micro-grammes par m^3, en fonction de la température moyenne annuelle X, exprimée en degrés Fahrenheit. On donne les résultats numériques suivants :

$$\sum x_i = 2286,\ \sum y_i = 1232,\ \sum x_i^2 = 129549,\ \sum y_i^2 = 59050,\ \sum x_i y_i = 74598 \ .$$

1. Tester la pertinence de la régression au seuil de 5%.
2. Tester $\mathcal{H}_0 \ : a = 3$ contre $\mathcal{H}_1 \ : a < 3$, au seuil de 5%.
3. Si vous deviez fixer une limite maximale de pollution pour une ville dont la température moyenne est de 60 degrés, qui ne soit dépassée que dans 5% des cas, quelle limite choisiriez-vous ?

Exercice 4.3.3. Dans le cadre de travaux de recherche sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne (en degrés Celsius) ainsi que l'altitude (en mètres) de chaque station sont données dans le tableau ci-dessous :

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
température	7.4	6	4.5	3.8	2.9	1.9	1	-1.2	-1.5	-4.5

On donne :

$$\sum x_i = 19690;\ \sum y_i = 20.3;\ \sum x_i^2 = 42925500;\ \sum y_i^2 = 162.41;\ \sum x_i y_i = 17671\,.$$

1. Tester la pertinence de la régression au seuil de 1%.
2. Dans un endroit situé à 1100 mètres d'altitude, on a relevé une température moyenne de 3.2 degrés. Au seuil de 1%, diriez-vous que cette température est anormalement basse ?

Exercice 4.3.4. On souhaite vérifier si la prise de poids d'un jeune mouton en un an (variable Y en kilogrammes) dépend de son poids initial (variable X également en kilogrammes). Sur 10 moutons, on donne les résultats suivants :

$$\sum x_i = 406\ ;\ \sum y_i = 423\ ;\ \sum x_i^2 = 16570\ ;\ \sum y_i^2 = 18057\ ;\ \sum x_i y_i = 17280\ .$$

1. Tester la pertinence de la régression au seuil de 1%.
2. La sagesse populaire dit que le poids d'un mouton doit doubler en un an. Au seuil de 1%, pouvez-vous confirmer ?
3. Un mouton de poids initial 30 kg, n'a pris que 20 kg au bout d'un an. Au seuil de 1%, est-ce inquiétant ?

Exercice 4.3.5. Le volume d'air expiré Y est une mesure standard du fonctionnement pulmonaire. Pour identifier une population possédant un fonctionnement pulmonaire anormal, il faut établir un modèle pour le volume d'air expiré dans une population normale. Pour cela, on mesure le volume Y en litres et la taille X en centimètres sur 12 garçons âgés de 10 à 15 ans.

On obtient les résumés numériques suivants :

$$\sum x_i = 1872\ ;\ \sum y_i = 32.3\ ;\ \sum x_i^2 = 294320\ ;\ \sum y_i^2 = 93.11\ ;\ \sum x_i y_i = 5156.20\ .$$

1. Tester la pertinence de la régression au seuil de 1%.
2. Un garçon mesurant 1.60 m expire 2.1 litres : est-ce alarmant ?

Exercice 4.3.6. On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour faire une régression linéaire, on effectue un changement de variable en posant $Y = \ln(H)$ et $X = \ln(D)$. Voici les mesures faites sur 5 arbres.

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

On donne :

$$\sum x_i = -4.71,\ \sum y_i = 12.12,\ \sum x_i^2 = 5.4095,$$
$$\sum y_i^2 = 29.5282,\ \sum x_i y_i = -11.0458.$$

1. Tester la pertinence de la régression au seuil de 5%.
2. On a abattu un arbre de diamètre 0.7 qui mesurait 20 m. Était-il anormalement grand ?