Recueil d'exercices de travaux dirigés de statistique
SY02

Thierry Denœux, Ghislaine Gayraud, Jean-Benoist Leger, Philippe Xu

Version courante :
- Base : Automne 2021
- Date : 2022-02-04 10:16:33
- Version : A2021-8-g49f22b1

Table des matières

Consignes :

- Les exercices de TD doivent être préparés avant chaque séance.
- Les étudiants pourront être interrogés dès le début des séances.
- Tous les exercices ne seront pas systématiquement traités en séance. Il appartient aux élèves de traiter en autonomie les exercices non abordés en séance.

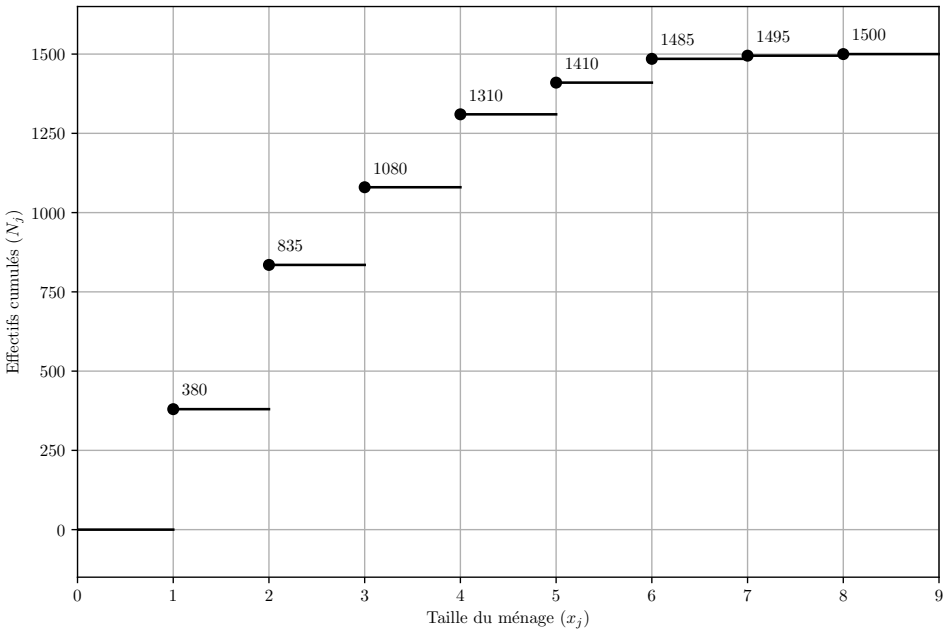
— Les étudiants sont invités à échanger sur le forum du Moodle afin de compléter les corrections données en TD.
— Des exercices supplémentaires pourront être proposés sur le Moodle afin d'approfondir des notions du cours.

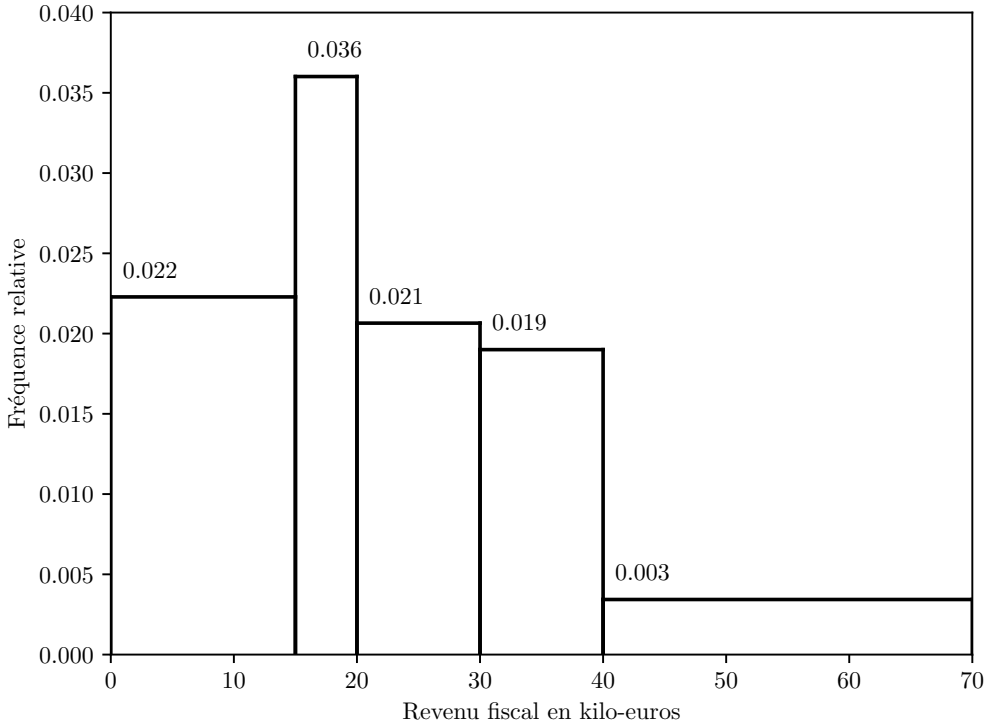
Chapitre 1

Statistiques descriptives

Exercice 1.1 *À préparer intégralement.*

Une enquête menée auprès de 1500 ménages d'une certaine région géographique rurale s'est intéressée à la variable correspondant à la taille du ménage, c'est-à-dire au nombre de personnes constituant le ménage, ainsi qu'à celle correspondant au revenu fiscal du ménage. Les données recueillies ont permis de dessiner la fonction de répartition et l'histogramme ci-dessous.





1. De quel type sont les variables « taille du ménage » et « revenu fiscal » ? Justifier.
2. Pour la variable « taille du ménage », calculer le premier décile, le premier quartile, la médiane, le troisième quartile, le neuvième décile, la moyenne empirique et l'écart-type empirique de l'échantillon.
3. À partir de l'histogramme, construire le tableau de fréquences associé et donner un encadrement du revenu fiscal médian.

Exercice 1.2

Les données ci-dessous sont le résultat d'une étude effectuée entre le 25/09/86 et le 18/10/87 dans la forêt équatoriale de Nouvelle-Guinée. Elles concernent les quantités de poisson (en kg) pêchées par chaque habitant de sexe masculin d'un village. Sont indiqués, pour chaque individu : l'âge, la situation familiale (M = marié, W = veuf, B = célibataire, BM = marié au cours de la période d'étude, Y = jeune homme, C = enfant), le niveau en chasse et pêche (D = débutant, I = intermédiaire, C = confirmé, E = expert), le nombre de nuits passées dans les zones de chasse et de pêche du village, et les quantités (en kg) de poissons pêchées par deux procédés différents.

Nom	Age	Sit.	Niveau	Nuits	Javelot	Hameçon
Bisaeo	45	M	I	327	8.1	2.5
Gugwi	45	M	E	363	54.1	13.6
Wodai	45	W	I	43	1.0	12.1
Mamo	40	M	I	310	11.7	3.4
Simo	35	M	C	295	3.3	21.8
Gwase	28	BM	C	346	10.3	12.0
Tufa	25	BM	D	155	2.2	2.7
Gwuho	25	B	I	136	2.0	7.1
Filifi	25	B	C	274	30.7	0.9
Sinio	22	M	E	267	52.2	8.0
Maubo	20	Y	E	362	16.2	27.9
Dogo	15	Y	C	314	11.8	14.9
Hegogwa	15	Y	I	122	3.8	15.9
Gawua	10	C	D	263	0.5	2.0
Okre	3	C	D	355	0.0	0.0

1. Indiquer la nature de chacune des variables du tableau. Une attention particulière sera portée à la justification de chaque réponse.
2. Résumer les distributions des variables « situation familiale » et « niveau » sous forme de tableaux de fréquences.
3. Tracer la fonction de répartition empirique de la variable « âge ».
4. Représenter la distribution de la variable « nombre de nuits » sous la forme d'un histogramme.
5. Représenter les distributions des variables « javelot » et « hameçon » à l'aide de diagrammes en tige et feuilles.
6. Calculer la moyenne empirique, la variance empirique (corrigée et non corrigée), l'écart-type empirique (corrigé et non corrigé), la médiane, les quartiles et l'étendue interquartiles des variables « javelot » et « hameçon ».
7. À l'aide d'un diagramme en boîtes, comparer les quantités de poissons obtenues par les deux modes de pêche.
8. Tracer la quantité de poisson pêchée par un javelot en fonction de celle par un hameçon pour tous les individus. Calculer le coefficient de corrélation de Pearson et commenter.

Exercice 1.3 *Non traité en TD.*

On a reporté dans le tableau suivant les prénoms d'un groupe d'étudiants avec une indication du nombre de livres lus dans l'année (A = peu, B = moyen, C = beaucoup, D = exceptionnel).

Pierre	Paul	Jacques	Gregory	Clara	Chloé	Henri
C	C	A	B	A	B	C
Paulette	Fanny	Laure	Kevin	Carole	Claire	Jeanine
B	B	C	D	B	A	C
Julie	Ernest	Cindy	Vanessa	José	Aurélien	
C	C	C	D	C	C	

1. Indiquer la nature de la variable ainsi mesurée.
2. Résumer la distribution de cette variable sous forme d'un tableau de fréquences.
3. Représenter cette distribution à l'aide d'un diagramme à bandes.

Exercice 1.4 *Non traité en TD.*

Un atelier réalise le séchage de boues d'origine industrielle. Il obtient à la fin du processus des déchets. On a observé les masses suivantes mesurées en kg de déchets après le traitement de 100 kg de boues :

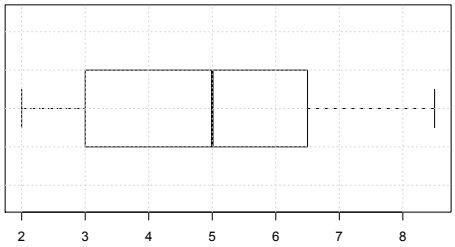
4.7 4.3 4.5 4.9 4.2 4.7 4.0 4.2 5.0 3.9 4.6 4.6.

1. Tracer le diagramme en tige et feuilles de cet échantillon.
2. Tracer la fonction de répartition empirique.
3. Calculer la moyenne empirique, la variance empirique corrigée, l'écart-type empirique corrigé, la médiane, les quartiles, l'étendue et l'étendue interquartiles.
4. Tracer le diagramme en boîte.
5. Supposons que la 9^{e} valeur soit 50 et non 5.0 (à cause d'une erreur d'unité dans la saisie des données). Que deviennent alors les résumés numériques et le diagramme en boîte de la nouvelle distribution ?

Exercice 1.5 *Non traité en TD.*

On considère le jeu de données ci-dessous, et le diagramme en boîte correspondant. On a calculé $\sum_{i=1}^{n} x_i = 96.5$, $\sum_{i=1}^{n} x_i^2 = 531.75$.

2.5	6.0	7.0	8.5	2.0
5.0	7.0	3.0	3.5	5.5
6.5	3.0	3.0	3.5	5.5
5.0	2.5	6.5	6.5	4.5



1. Sans aucune autre information que celle de l'énoncé, quel type de variable utilisez-vous dans votre modélisation ? Justifier.
2. Calculer la moyenne, la moyenne tronquée d'ordre 1, la variance empirique, et la variance empirique corrigée.
3. Déterminer la médiane, les quartiles, l'étendue et l'étendue inter-quartiles.

Exercice 1.6 *Non traité en TD.*

Une enquête menée auprès d'un échantillon de 40 habitants d'une certaine commune afin d'étudier leurs habitudes de lecture du journal trimestriel de la commune fournit le tableau suivant (la variable N correspond au nombre de personnes vivant dans le foyer, Fl les habitudes de lecture et S le sexe).

Age	N	Fl	S	Age	N	Fl	S
17	4	régulièrement	F	10	3	jamais	H
12	2	rarement	H	40	5	régulièrement	F
15	3	rarement	F	54	5	rarement	F
87	1	toujours	F	25	3	régulièrement	H
32	1	jamais	F	53	4	rarement	F
33	2	régulièrement	H	27	3	rarement	F
45	4	jamais	H	57	4	régulièrement	H
46	1	rarement	H	59	2	régulièrement	F
29	2	régulièrement	H	13	5	rarement	F
38	3	rarement	F	53	3	régulièrement	H
76	2	toujours	H	67	3	toujours	F
65	2	toujours	F	16	5	rarement	H
59	6	régulièrement	F	55	4	rarement	H
12	2	jamais	H	49	6	régulièrement	F
14	4	régulièrement	H	58	2	jamais	F
15	2	rarement	H	21	2	jamais	H
66	2	rarement	F	95	2	rarement	F
38	2	rarement	F	28	3	régulièrement	H
40	4	régulièrement	F	65	2	régulièrement	F
42	5	régulièrement	H	89	1	toujours	H

1. Indiquer la nature de chacune des variables du tableau.
2. Tracer les diagrammes en bâton des variables discrètes ou qualitatives et les histogrammes des variables continues.

Exercice 1.7 *Non traité en TD.*

Le tableau suivant indique les nombres de voitures particulières vendues au États-Unis en 1990, par classes de masses. La signification des notations est la suivante :

— c_k : centre de la classe k,
— $[a_k, a_{k+1}[$: intervalle définissant la classe k,
— n_k : nombre de véhicules vendus en 1990 (en milliers) pour la classe k.

c_k	$[a_k, a_{k+1}[$	n_k
1750	[1500,2000[	510
2250	[2000,2500[	1290
2750	[2500,3000[	4070
3500	[3000,4000[	4094
5000	[4000,6000[	720

1. Calculer les fréquences relatives et les fréquences relatives cumulées.
2. Tracer un histogramme de la distribution des ventes de véhicules en 1990.
3. Donner la moyenne, l'écart-type et la classe modale de cette distribution.

Chapitre 2

Probabilités

Exercice 2.1 *À préparer intégralement.*

Soit f et g les fonctions suivantes :

$$f(x) = \begin{cases} ax & \text{si } 0 \leq x \leq b \\ 0 & \text{sinon,} \end{cases} \tag{0.1}$$

$$g(x) = \begin{cases} \frac{c}{x^2} & \text{si } x \geq d \\ 0 & \text{sinon,} \end{cases} \tag{0.2}$$

où $b > 0$ et $d > 0$.

1. Trouver les constantes a en fonction de b, respectivement c en fonction de d, pour que les fonctions (0.1) et respectivement (0.1) soient des densités de probabilité.
2. Donner ensuite les expressions de l'espérance mathématique et de la variance (si elles existent) des v.a. correspondantes :
3. Tracer les fonctions de densité et les fonctions de répartition correspondantes.

Exercice 2.2

On a constaté que la répartition du taux de cholestérol dans une population de grande taille est la suivante :

— taux inférieur à 165 cg : 58 % ;
— taux compris entre 165 et 180 cg : 38 % ;
— taux supérieur à 180 cg : 4 %.

1. On considère que le taux de cholestérol est distribué selon une loi normale. Discuter du choix de modèle.
2. En supposant le modèle justifié, calculer la valeur moyenne et l'écart-type du taux de cholestérol dans la population.

3. Vérifier que le choix du modèle n'est pas aberrant.
4. On admet que les personnes dont le taux est supérieur à 183 cg doivent suivre un traitement. Quel est le nombre de personnes à soigner dans une population d'un million de personnes ?

Exercice 2.3

Soit X une variable aléatoire telle que :

$$\mathbb{P}(X = a) = \begin{cases} 1/4 & \text{si } a = -1 \\ 1/2 & \text{si } a = 0 \\ 1/4 & \text{si } a = 1 \end{cases}$$

Soit Y une variable aléatoire telle que :

$$\mathbb{P}(Y = b) = \begin{cases} 1/2 & \text{si } b = -1 \\ 1/2 & \text{si } b = 1 \end{cases}$$

On suppose X et Y indépendantes.
On introduit

$$U = XY$$

1. Calculer la loi de U.
2. Calculer $\mathbb{P}(X = -1 \text{ et } U = -1)$. En déduire si U et X sont indépendantes.
3. En oubliant la question précédente, calculer $\mathbb{P}(U = 1 \mid X = 1)$ en déduire si X et U sont indépendantes.

Exercice 2.4

L'objectif de cet exercice est de présenter une méthode générique de simulation/génération d'une v.a. à partir de la réciproque de sa fonction de répartition et de la simulation d'une v.a. $U \sim \mathcal{U}([0,1])$.

1. Soit X un v.a. de loi absolument continue de fonction de répartition F_X, et de fonction quantile F_X^{-1} (F_X^{-1} est la fonction réciproque de F_X).
   En posant $Y = F_X^{-1}(U)$, avec $U \sim \mathcal{U}([0,1])$, montrer que Y suit la même loi que X.
2. Application à la loi Exponentielle $\mathcal{E}(\lambda)$ avec λ un réel strictement positif. On note u une réalisation de la v.a. $U \sim \mathcal{U}([0,1])$ (cela signifie que u est une valeur simulée de la loi $\mathcal{U}([0,1])$). Proposer une méthode de simulation de la loi $\mathcal{E}(\lambda)$, i.e., proposer une méthode numérique qui conduise à une réalisation x de la v.a. $X \sim \mathcal{E}(\lambda)$.

Exercice 2.5

On considère n variables aléatoires $X_1, \dots, X_n$ indépendantes de même distribution caractérisées par une fonction de répartition F.

1. Déterminer la fonction de répartition G de la variable aléatoire

$$Y = \max_{1 \leq i \leq n} X_i$$

   en fonction de F.

2. On suppose que les variables aléatoires X_i ont une densité f. En déduire que Y a une densité g que l'on calculera.

3. On suppose maintenant que X suit une loi uniforme $\mathcal{U}_{[0,\theta]}$, avec $\theta > 0$. Calculer l'espérance et la variance de la variable aléatoire Y.

Exercice 2.6 *Non traité en TD.*

L'équipe de football A affronte régulièrement l'équipe B. Soit X une v.a. discrète à valeurs dans $V_X = \{-1, 0, +1\}$ et représentant le résultat d'un match entre ces deux équipes. On utilise la convention suivante :

— $X = -1$: défaite de l'équipe A
— $X = \ \ 0$: match nul
— $X = +1$: victoire de l'équipe A

On suppose que la loi de probabilité de X est définie en fonction d'un paramètre θ de la façon suivante :

$$\begin{aligned} \mathbb{P}(X = -1) &= \theta/2 \\ \mathbb{P}(X = \ \ 0) &= 1/2 \\ \mathbb{P}(X = +1) &= (1-\theta)/2. \end{aligned}$$

1. Vérifier que la loi ci-dessus est bien une loi de probabilité. Si besoin, appliquer des conditions sur θ.

2. Calculer en fonction de θ l'espérance et la variance de X.

3. On note $X_1, \dots, X_n$ les résultats de n matchs suivants la même distribution que X. Peut-on supposer que les variables aléatoires sont indépendantes ?

4. On note N_1 le nombre de victoires de l'équipe A. Donner la loi, l'espérance et la variance de N_1.

5. En prenant $\theta = 0.2$ et $n = 5$, calculer la probabilité que l'équipe A remporte au moins trois des cinq matchs contre l'équipe B.

Exercice 2.7 *Non traité en TD.*

On considère le nombre de passages de véhicules en un point d'une autoroute reliant deux villes A et B durant un intervalle d'une minute. On suppose que dans le sens de A à B, ce nombre X est une variable de Poisson de paramètre 17.8 ; dans le sens de B à A, le nombre Y de passages est une variable de Poisson de paramètre 7.1.

1. Justifiez du choix de la loi de X et de Y.
2. Quelle est la distribution de nombre total de passages durant une minute ?
3. On fait l'hypothèse d'indépendance entre X et Y. Justifier que cette hypothèse peut être fausse.
4. Donner une approximation de la probabilité pour que le nombre total de passages enregistré durant une heure soit strictement plus grand que 1500.

Exercice 2.8 *Non traité en TD.*

L'objectif de cet exercice est de démontrer le résultat suivant, très utile dans le cours, connu sous le nom de théorème de Fisher. Soit

$$X_i \overset{i.i.d.}{\sim} \mathcal{N}\left(\mu, \sigma^2\right),$$

avec

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

alors on a :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \sim \chi^2_{n-1}.$$

Plusieurs moyennes pourront apparaître au fil des calculs, il vous est conseillé de ne jamais noter $\overline{X}$, mais $\overline{X}_n$ avec :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Nous noterons également :

$$S_n = \sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Ainsi, l'objectif sera de montrer que :

$$\frac{1}{\sigma^2}S_n \sim \chi^2_{n-1}.$$

1. Exprimer $\overline{X}_{n+1}$ en fonction de $\overline{X}_n$ et de X_{n+1}.
2. En déduire $\overline{X}_n$ en fonction de $\overline{X}_{n+1}$ et de X_{n+1}.

3. Montrer que pour $n \geq 2$, on a $S_{n+1} - S_n = Z_n^2$ avec

$$Z_n = \frac{1}{\sqrt{n(n+1)}} \left(nX_{n+1} - \sum_{i=1}^{n} X_i \right). \tag{0.3}$$

   Pour mener à bien les calculs, on s'efforcera de ne jamais développer et de faire apparaître des différences de carrés pour factoriser.

4. Montrer que $S_2 = Z_1^2$.
5. Montrer que pour $n \geq 2$ on a $S_n = \sum_{k=1}^{n-1} Z_k^2$.
6. On s'intéresse à présent à la loi des Z_k.
   (a) Montrer que les Z_k sont des variables normales.
   (b) Montrer que $\mathbb{E}Z_k = 0$.
   (c) Montrer que $\mathrm{Var}\, Z_k = \sigma^2$.
   (d) Montrer que les Z_k sont indépendants deux à deux.
   (e) Conclure sur la loi des Z_k.
7. Conclure sur la loi de $\frac{1}{\sigma^2} S_n$ en utilisant les questions précédente et la définition d'une loi du χ^2.

Exercice 2.9 *Non traité en TD.*

Suite de l'exercice 2.8.

Cet exercice nécessite des outils hors programme pour SY02. Un fort prérequis d'algèbre linéaire (fournit dans le programme de TC ou des CPGE par exemple) est nécessaire. Sa réalisation n'est pas au programme de SY02. Toutefois, il permet de démontrer le même résultat que l'exercice 2.8 de manière générale, et fournit un lemme qui est utilisé pour démontrer un grand nombre de résultats du cours de SY02.

Dans cette méthode nous montrerons d'abord un lemme général, et nous appliquerons ce lemme dans le cas particulier qui nous intéresse.

Dans cette partie la notation anglo-saxonne de la transposition sera utilisée (M^T). Vous pouvez également utiliser la notation française (${}^tM$). De manière générale, tout mélange entre les deux notations est à proscrire (M^t et ${}^TM$).

Soit $Y_i \sim \mathcal{N}(0,1)$ indépendants. On notera Y le vecteur colonne :

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

L'objectif est de démontrer que si M est symétrique et que c'est une matrice de projecteur, alors :

$$Y^T M Y \sim \chi^2_{\mathrm{rg}\, M}, \quad \text{avec } \mathrm{rg}\, M = \mathrm{tr}\, M$$

avec $\mathrm{rg}\, M$ le rang de M, et $\mathrm{tr}\, M$ la trace de M.
Dans toute la suite, M désignera une matrice $\mathbf{R}^{n\times n}$ symétrique de projecteur quelconque.

Cas général

1. Montrer que M se diagonalise en base orthonormale, de la forme :

$$M = PDP^T \quad \text{avec} \begin{cases} P \text{ une matrice orthogonale} \\ D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \\ (\lambda_1, \cdots, \lambda_n) = (\underbrace{1, \cdots, 1}_{\mathrm{rg}(M)}, \underbrace{0, \cdots, 0}_{n-\mathrm{rg}(M)}) \end{cases}$$

2. Montrer que $\mathrm{rg}(M) = \mathrm{tr}(M)$.
3. Soit $Z = P^T Y$. Justifier que les Z_k sont indépendants de loi $\mathcal{N}(0,1)$.
4. Montrer que :

$$Y^T M Y = \sum_{i=1}^{\mathrm{rg}(M)} Z_i^2$$

5. En déduire le résultat du lemme.

Application dans le cas particulier
Soient (X_i) et S_n tels que définis en préliminaire.
Soit X le vecteur colonne des (X_i) :

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

1. Démontrer que $X - \overline{X}_n$ s'exprime en fonction de X sous forme de produit matriciel, avec une matrice M.

$$X - \overline{X}_n = \begin{bmatrix} X_1 - \overline{X}_n \\ \vdots \\ X_n - \overline{X}_n \end{bmatrix} = MX$$

   On exprimera M en fonction des matrices I et U (respectivement matrice identité et matrice constituée uniquement de 1).
2. Montrer que M est symétrique.
3. Démontrer que M est une matrice de projecteur. Déterminer son rang.
4. Montrer que $S_n = X^T M X$.

5. En posant $Y = (X - \mu)/\sigma$, montrer que

$$Y^T M Y = \frac{1}{\sigma^2} S_n$$

6. Vérifier que Y vérifie les propriétés du lemme, appliquer icelui, et en déduire le résultat du problème.

Dans le cours de SY02, de nombreux autres résultats font intervenir un χ^2 avec un nombre de degré de liberté différent suivant le problème. Ces résultats sont démontrables en utilisant ce lemme et en calculant le rang des projecteurs impliqués.

Chapitre 3

Échantillonnage. Théorème central de la limite

Exercice 3.1 *À préparer intégralement.*

Un coureur à pied couvre en moyenne 142 cm à chaque foulée, avec un écart-type de 5 cm.

1. Modéliser le phénomène et introduire les notations.
2. Soit S_n la distance parcourue par un coureur en n foulées. Donner une loi approchée de S_n lorsque n est grand. On justifira le choix du modèle.
3. En déduire la probabilité qu'un coureur parcoure 100 m en moins de 70 foulées.

Exercice 3.2

La DREES publie des informations sur la répartition des chirurgiens [1] suivant les spécialités en France, la partie utile est ici reproduite :

Spécialité	Effectif
Chirurgie générale	3611
Chirurgie infantile	300
Chirurgie maxillo-faciale et stomatologie	1170
Chirurgie orthopédique et traumatologie	3157
Chirurgie plastique reconstructrice et esthétique	869
Chirurgie thoracique et cardio-vasculaire	436
Chirurgie urologique	1171
Chirurgie vasculaire	523
Chirurgie viscérale et digestive	977

1. https://drees.solidarites-sante.gouv.fr/IMG/pdf/rpps_medecins_-_synthese_des_effectifs_au_1er_jan2015.pdf

On vous demande d'estimer le revenu moyen des chirurgiens en France. Pour des raisons budgétaires, il ne vous est permis que d'interroger au plus 100 médecins. Construire un échantillonnage, et exhibez la moyenne empirique sur l'échantillon considéré permettant d'atteindre ce but au mieux avec les informations dont vous disposez.

Exercice 3.3

Un avion de ligne peut accueillir jusqu'à 350 passagers. On suppose que la masse d'un passager est une v.a. de moyenne 70 kg et d'écart-type 10 kg, et la masse des bagages d'un passager est une v.a. de moyenne 45 kg et d'écart-type 5 kg. La capacité l d'un avion représente la limite que la masse total de la cargaison (passagers et bagages) ne doit pas excéder pour naviguer en toute sécurité. Pour qu'un service aérien soit considéré comme sûr, la probabilité que la masse de la cargaison d'un avion dépasse sa capacité l doit être inférieure au risque $\alpha = 10^{-8}$.

1. Modéliser le problème.
2. Considérer une v.a. représentant la masse d'un passager avec ses bagages. Transférer le modèle sur cette variable.
3. Que doit valoir l, au minimum, pour que cette consigne soit respectée ?

Exercice 3.4

Soit $X_1, \dots, X_n$ des variables aléatoires représentant des mesures répétées d'un même phénomène. On considère que pour tout i, $\mathbb{E}(X_i) = \mu$ et $\mathrm{Var}(X_i) = \sigma^2$. Toutefois, on ne considère pas l'indépendance entre les variables. On considère que les mesures successives sont corrélées, et que la corrélation entre deux mesures est d'autant plus faible que les mesures sont espacées dans le temps :

$$\forall i, k,\ \mathrm{Cor}(X_i, X_{i+k}) = \rho^k,$$

avec $|\rho| < 1$. On note $S_n = \sum_{i=1}^{n} X_i$. Le calcul suivant est fourni.

$$\begin{aligned}
\frac{\mathrm{Var}(S_n)}{\sigma^2} &= \frac{\mathrm{Var}(S_{n-1} + X_n)}{\sigma^2} \\
&= \frac{\mathrm{Var}(S_{n-1})}{\sigma^2} + 1 + \frac{2\,\mathrm{Cov}\left(\sum_{i=1}^{n-1} X_i, X_n\right)}{\sigma^2} \\
&= \frac{\mathrm{Var}(S_{n-1})}{\sigma^2} + 1 + 2\sum_{i=1}^{n-1} \mathrm{Cor}(X_i, X_n) \\
&= \frac{\mathrm{Var}(S_{n-1})}{\sigma^2} + 1 + 2\sum_{i=1}^{n-1} \rho^i \\
&= \frac{\mathrm{Var}(S_{n-1})}{\sigma^2} + 1 + 2\frac{\rho - \rho^n}{1 - \rho}
\end{aligned}$$

Or $\mathrm{Var}(S_0) = \mathrm{Var}(0) = 0$ donc :

$$\begin{aligned}
\frac{\mathrm{Var}(S_n)}{\sigma^2} &= \sum_{i=1}^{n} 1 + 2\frac{\rho - \rho^i}{1-\rho} \\
&= n + 2\rho\frac{n - \sum_{i=1}^{n}\rho^{i-1}}{1-\rho} \\
&= n + 2\rho\frac{n - \frac{1-\rho^n}{1-\rho}}{1-\rho} \\
&\approx n + 2\rho\frac{n}{1-\rho} && (\text{car } \tfrac{1-\rho^n}{1-\rho} \ll n \text{ pour } n \text{ grand}) \\
&\approx n\frac{1+\rho}{1-\rho}
\end{aligned}$$

On introduit $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. Calculer $\mathbb{E}\left(\overline{X}_n\right)$ et $\mathrm{Var}\left(\overline{X}_n\right)$.
2. Montrer que $\overline{X}_n$ converge en probabilité vers μ. Pouvait-on obtenir ce résultat avec la loi faible des grand nombres ?
3. Pour $\rho = -\frac{1}{2}$, $\rho = -\frac{1}{4}$, $\rho = 0$, $\rho = \frac{1}{4}$, $\rho = \frac{1}{2}$, calculer $\mathrm{Var}\left(\overline{X}_n\right)$. Comparer ces valeurs entre elles et à ce qui est attendu dans un cas d'indépendance. Interpréter.

Exercice 3.5 *Non traité en TD.*

Un local doit être éclairé en permanence au moyen d'une ampoule ; lorsque l'ampoule tombe en panne, elle est immédiatement remplacée. Il y a deux qualités d'ampoules : les ampoules de qualité A ont une durée de vie (en heures) qui est distribuée exponentiellement avec le paramètre $\theta_A = 0.01$, les ampoules de qualité B ont une durée de vie (en heures) qui est distribuée exponentiellement avec le paramètre $\theta_B = 0.02$. On a stocké 40 ampoules de qualité A et 60 ampoules de qualité B. Quelle est la probabilité pour que cette réserve d'ampoules soit suffisante pour un éclairage de 6500 heures du local ? Le problème sera modélisé avant de répondre aux questions.

Exercice 3.6 *Non traité en TD.*

On considère n variables aléatoires $X_1, \ldots, X_n$ indépendantes de même distribution de Poisson $\mathscr{P}(\lambda)$ avec $\lambda = 1$.

1. Montrer que la somme de deux v.a. indépendantes de Poisson suit encore une loi de Poisson, dont on précisera la paramètre. En déduire la distribution de la variable aléatoire $Y_n = \sum_{i=1}^{n} X_i$.
2. Déterminer la quantité $\mathbb{P}(Y_n \leq n)$ pour $n = 20$ en utilisant les tables statistiques.

3. Déterminer la même quantité mais de manière approximative en vous appuyant sur le théorème central limite. Qu'en est-il pour $n = 50,\ 100,\ 200$?
4. Dans la table statistique de la loi de Poisson, on propose d'approcher la fonction de répartition d'une v.a. de Poisson $X \sim \mathscr{P}(\lambda)$ lorsque $\lambda \geq 20$ par

$$\mathbb{P}(X \leq x) \approx \Phi\left(\frac{x + 0.5 - \lambda}{\sqrt{\lambda}}\right).$$

   Déterminer, toujours de manière approximative, les quantités calculées à la question précédente mais en utilisant cette nouvelle approximation qui introduit un terme de correction.
5. En utilisant un ordinateur, on peut retrouver, sans approximation, que pour $n = 50$, 100 et 200, on obtient respectivement les valeurs 0.5375, 0.5266 et 0.5188. Que peut-on en conclure ?

Exercice 3.7 *Non traité en TD.*

Jeu de pile ou face.

1. Un joueur lance 100 fois une pièce de monnaie parfaitement équilibrée. Trouver la probabilité qu'il obtienne 60 faces ou plus.
2. Le gérant du casino considère maintenant chacun des 500 joueurs présents dans son établissement. Chaque joueur lance 100 fois une pièce de monnaie parfaitement équilibrée. Quelle est la probabilité pour que 10 joueurs ou plus obtiennent 60 faces ou plus ?

Exercice 3.8 *Non traité en TD.*

On suppose que le nombre de clients moyen entrant dans un magasin un jour donné est de 12.

En faisant des hypothèses que l'on précisera, calculer la probabilité d'avoir au moins 250 entrées de clients durant un mois de 22 jours ouvrables.

Chapitre 4

Estimation, méthode des moments

Exercice 4.1 *À préparer intégralement.*

Soit $X_1, \dots, X_n$ un échantillon i.i.d. de v.a. parente X de densité de probabilité :

$$f(x) = \begin{cases} 1-\theta & \text{si } x \in]-1/2, 0], \\ 1+\theta & \text{si } x \in]0, +1/2], \\ 0 & \text{sinon,} \end{cases}$$

où $\theta \in]-1; 1[$ est un paramètre inconnu.

1. Calculer l'espérance de X et en déduire une expression de θ en fonction de cette dernière.
2. Trouver un estimateur $\hat{\theta}_m$ de θ par la méthode des moments.
3. Cet estimateur est-il sans biais ? Est-il convergent ?

Exercice 4.2

On considère un échantillon $X_1, \dots, X_n$ iid dont la variable aléatoire parente X suit la loi

$$\begin{aligned} \mathbb{P}(X=0) &= 1 - p_1 - p_2 \\ \mathbb{P}(X=1) &= p_1 \\ \mathbb{P}(X=2) &= p_2, \end{aligned}$$

où p_1 et p_2 sont des paramètres strictement positifs.

1. Calculer $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ et $\mathrm{Var}(X)$.
2. En utilisant les statistiques $\overline{X} = \frac{1}{n}\sum_i X_i$ et $\widehat{m}_2 = \frac{1}{n}\sum_i X_i^2$, déterminer des estimateurs $\hat{p}_1$ et $\hat{p}_2$ de p_1 et p_2 par la méthode des moments.
3. Montrer que ces estimateurs sont sans biais.

Exercice 4.3

Soit X une v.a. suivant une loi continue uniforme sur l'intervalle $[-\theta, \theta]$, $\theta \in \mathbb{R}_+^*$ étant un paramètre inconnu, et $X_1, \dots, X_n$ un échantillon i.i.d. de variable parente X.

1. Déterminer un estimateur $\hat{\theta}_1$ de θ par la méthode des moments.

   L'inégalité de Jensen implique que si g est une fonction strictement convexe et Y une v.a. non dégénérée (de variance non nulle), alors

   $$\mathbb{E}[g(Y)] > g(\mathbb{E}[Y]).$$

   L'estimateur $\hat{\theta}_1$ est-il biaisé ?
2. On considère maintenant la variable aléatoire $Y = |X|$ et l'échantillon associé $Y_1, \dots, Y_n$ où $Y_i = |X_i|$. Calculer l'espérance et la variance de Y.
3. Déterminer un second estimateur $\hat{\theta}_2$ de θ par la méthode des moments à partir de $\mathbb{E}(Y)$. Est-il sans biais ?

Exercice 4.4

On considère une variable aléatoire X suivant une loi de Pareto de paramètres $\theta > 0$ et $k = 1$, dont on a observé n réalisations i.i.d. $x_1, \dots, x_n$; la densité de X est définie par

$$f_X(x) = \theta x^{-2} \mathbb{1}_{[\theta;+\infty[}(x) = \begin{cases} \theta x^{-2} & \text{si } x \geq \theta, \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer l'espérance de X ; à partir de ce résultat, appliquer la méthode des moments.
2. Calculer le moment d'ordre $\frac{1}{2}$. Appliquer la méthode des moments.

Exercice 4.5 *Non traité en TD.*

Soit $X_1, \dots, X_n$ un échantillon i.i.d. dont la v.a. parente X est une v.a. continue de loi de probabilité

$$f(x) = \begin{cases} \dfrac{\theta}{x^{\theta+1}} & \text{si } x > 1 \\ 0 & \text{sinon,} \end{cases}$$

où θ est un paramètre réel strictement supérieur à 1.

1. Trouver un estimateur $\hat{\theta}_1$ de θ par la méthode des moments.
2. L'inégalité de Jensen implique que si g est une fonction strictement convexe et Y une v.a. non dégénérée (de variance non nulle), alors

   $$\mathbb{E}[g(Y)] > g(\mathbb{E}[Y]).$$

   Montrer que l'estimateur calculé précédemment est biaisé.

3. Déterminer la fonction de répartition F de X. En déduire la fonction de répartition G de $Y = \ln(X)$ et montrer que Y suit une loi exponentielle.
4. Calculer par la méthode des moments un second estimateur $\hat{\theta}_2$ de θ. Cet estimateur est-il sans biais ?

Exercice 4.6 *Non traité en TD.*

Soit θ une grandeur physique à mesurer. À l'aide d'un dispositif de mesure, on réalise plusieurs mesures indépendantes de θ :

$$\forall i \in \{1, \dots, n_X\}, \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\theta, \sigma_X^2\right)$$

Puis, à l'aide d'un autre dispositif de mesure, on réalise plusieurs mesures indépendantes de θ :

$$\forall i \in \{1, \dots, n_Y\}, \qquad Y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\theta, \sigma_Y^2\right)$$

Les mesures X sont indépendantes des mesures Y.

1. En utilisant uniquement les mesures X, construire avec la méthode des moments un estimateur de θ noté $\hat{\theta}_X$. Calculer son biais et sa variance.
2. De même, construire θ_Y en utilisant exclusivement les mesures Y. Calculer le biais et la variance de $\hat{\theta}_Y$.
3. On cherche à combiner $\hat{\theta}_X$ et $\hat{\theta}_Y$ sous forme d'un BLUE (*Best Linear Unbiaised Estimator*, estimateur linéaire non-biaisé de variance minimale). C'est à dire, une combinaison linéaire de $\hat{\theta}_X$ et $\hat{\theta}_Y$:
$$\hat{\theta} = a\hat{\theta}_X + b\hat{\theta}_Y$$
Déterminer a et b de telle manière que $\hat{\theta}$ soit un BLUE.

Exercice 4.7 *Non traité en TD.*

L'objectif de cet exercice est de continuer l'exercice 4.3 sur les estimateurs de moments, de manipuler des outils de base en statistique, et de comparer les estimateurs, analytiquement et en simulation.

Il montre qu'un estimateur biaisé peut être préférable à un estimateur non-biaisé au sens de la précision.

Comme dans l'exercice 4.3 nous avons un échantillon $X_i \overset{i.i.d.}{\sim} \mathcal{U}_{[-\theta,\theta]}$, avec $i \in \{1, \cdots, n\}$.
Via la méthode des moments, l'exercice vous fait construire deux estimateurs de θ :

$$\hat{\theta}_1 = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2},$$

et

$$\hat{\theta}_2 = \frac{2}{n}\sum_{i=1}^{n} |X_i|.$$

Il vous est fortement conseillé de traiter cet exercice avant de commencer les questions suivantes qui sont la suite de cet exercice.

1. **Étude de $\hat{\theta}_2$**
   (a) Calculer $\mathrm{Var}\,\hat{\theta}_2$. Puis calculer l'erreur quadratique moyenne (risque quadratique) de $\hat{\theta}_2$: $\mathbb{E}\left(\left(\hat{\theta}_2 - \theta\right)^2\right)$.

2. **Étude de $\hat{\theta}_1$**
   Pour cette partie, certaines propriétés sont rappelés :
   — Si $W \sim \mathcal{N}(0, \sigma^2)$, alors $\mathbb{E}\,W = 0$, $\mathbb{E}\,W^2 = \sigma^2$, $\mathbb{E}\,W^3 = 0$, $\mathbb{E}\,W^4 = 3\sigma^4$.
   (a) En introduisant $Z_i = \frac{3}{\theta^2}X_i^2$, exprimer $\frac{\hat{\theta}_1}{\theta}$ en fonction des Z_i.
   (b) Poser $K_n = \frac{1}{n}\sum_{i=1}^{n} Z_i - 1$, exprimer $\frac{\hat{\theta}_1}{\theta}$ en fonction de K_n.
   (c) Calculer $\mathbb{E}\,Z_i$ et $\mathrm{Var}\,Z_i$.
   (d) Démontrer l'approximation asymptotique : $K_n \overset{\text{app}}{\sim} \mathcal{N}\left(0, \frac{4}{5n}\right)$.
   (e) En utilisant l'approximation asymptotique suivante $\sqrt{1+W} \approx 1 + \frac{1}{2}W - \frac{1}{8}W^2$ pour $W \xrightarrow{\mathbb{P}} 0$, en déduire $\mathbb{E}\,\frac{\hat{\theta}_1}{\theta}$. En déduire le biais de $\hat{\theta}_1$, vérifier la cohérence du résultat avec le résultat de l'exercice 4.3.
   (f) De manière similaire, calculer $\mathbb{E}\left(\frac{\hat{\theta}_1}{\theta} - 1\right)^2$, en déduire l'erreur quadratique moyenne de l'estimateur $\hat{\theta}_1$.

3. Comparer $\hat{\theta}_1$ et $\hat{\theta}_2$. Mettre cette comparaison en perspective par rapport à la comparaison des biais de l'exercice 4.3.

Exercice 4.8 *Non traité en TD.*
Soit $(X_1, \dots, X_n)$ un échantillon iid de taille n dont la loi parente est la loi de Bernoulli de paramètre p et soit $\theta = \mathrm{Var}\,X$.

1. Construire un estimateur de p par la méthode des moments.
2. En utilisant le fait que $\mathrm{Var}\,(X) = p(1-p)$, déduire un estimateur T de $\theta = \mathrm{Var}\,X$.
3. Montrer que T est un estimateur biaisé de θ.
4. Proposer un estimateur sans biais de θ.

Chapitre 5

Méthode du maximum de vraisemblance

Exercice 5.1 *À préparer intégralement.*

On a observé les nombres d'absences par mois dans une entreprise noté $K_1, \dots, K_n$ pendant n mois consécutifs.

1. Montrer que le modèle suivant est applicable :

$$\forall i \in \{1, \dots, n\} \qquad K_i \overset{\text{iid}}{\sim} \mathscr{P}(\lambda)$$

2. Calculer la fonction de vraisemblance en fonction de $\overline{k}$, la réalisation de $\overline{K} = \frac{1}{n}\sum_{i=1}^{n} K_i$.
3. Calculer la fonction de log-vraisemblance et en déduire l'estimateur du maximum de vraisemblance de λ, que l'on notera $\hat{\lambda}_{MV}$.
4. Calculer l'information de Fisher apportée par l'échantillon sur le paramètre λ.
5. Donner une approximation de la loi de $\hat{\lambda}_{MV}$ pour n grand.

Exercice 5.2

Soit $X_1, \dots, X_n$ un échantillon iid issu d'une population de densité

$$f(x) = \begin{cases} \frac{\theta+1}{2}(1-|x|)^{\theta} & \text{si } -1 < x < 1 \\ 0 & \text{sinon,} \end{cases}$$

où $\theta > -1$.

1. Tracer l'allure densité pour $\theta = -\frac{1}{2}$, $\theta = 0$, $\theta = 1$, $\theta = 2$ et $\theta = 4$.
2. Déterminer $\hat{\theta}_{MV}$ l'estimateur du maximum de vraisemblance de θ.

3. En imaginant observer uniquement des valeurs proches de 0, comment se situe l'estimation de θ ? De même avec des valeurs proches de 1 ou -1. Est-ce cohérent avec les tracés de la densité ?
4. Déterminer l'information de Fisher apportée sur le paramètre θ par l'échantillon.
5. En déduire une approximation de la loi de $\hat{\theta}_{MV}$ pour n grand.

Exercice 5.3

Soit $(X_1, \dots, X_n)$ un échantillon extrait de la loi uniforme sur $[\theta, \theta + 1]$ avec $\theta > 0$. On pose $S_n = \max_{1 \leq i \leq n} X_i$ et $I_n = \min_{1 \leq i \leq n} X_i$.

1. Exprimer la fonction de vraisemblance en fonction de s_n et i_n.
2. Montrer que tous les estimateurs de la forme

$$\hat{\theta}_n(\alpha) = \alpha(S_n - 1) + (1 - \alpha)I_n$$

avec $\alpha \in [0, 1]$ sont des estimateurs du maximum de vraisemblance de θ.
3. Calculer la fonction de répartition, puis la fonction de densité des variables aléatoires S_n et I_n.
4. Calculer les espérances de S_n et I_n.
5. Quelle est l'unique valeur α^* telle que $\hat{\theta}_n = \hat{\theta}_n(\alpha^*)$ soit un estimateur sans biais de θ ?

Exercice 5.4

Soit X une v.a. discrète à valeurs dans $V_X = \{0, 1, 2\}$, dont la loi de probabilité est définie en fonction d'un paramètre $\theta \in]0, 1[$ de la façon suivante :

$$\begin{aligned} \mathbb{P}(X = 0) &= 1/2, \\ \mathbb{P}(X = 1) &= \theta/2, \\ \mathbb{P}(X = 2) &= (1 - \theta)/2. \end{aligned}$$

Soit $X_1, \dots, X_n$ un échantillon i.i.d. de v.a. parente X. Pour tout $k \in \{0, 1, 2\}$, on note N_k le nombre d'observations de l'échantillon égales à k, c'est-à-dire le cardinal de $\{i \in \{1, \dots, n\} \mid X_i = k\}$, et on note n_k la réalisation de N_k.

1. Calculer l'expression de la fonction de log-vraisemblance, en fonction de n_0, n_1 et n_2.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ.
3. Calculer l'information de Fisher associée au paramètre θ.
4. En déduire la loi approchée de cet estimateur quand n est grand.

Exercice 5.5 *Non traité en TD.*

On considère une variable aléatoire X suivant une loi de Pareto de paramètres $\theta > 0$ et $k = 1$, dont on a observé n réalisations indépendantes $x_1, \dots, x_n$; la densité de X est définie par

$$f_X(x) = \theta x^{-2} \mathbb{1}_{[\theta;+\infty[}(x) = \begin{cases} \theta x^{-2} & \text{si } x \geq \theta, \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer la fonction de vraisemblance de θ étant donné l'échantillon $x_1, \dots, x_n$.
2. Quel est l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ ?

Exercice 5.6 *Non traité en TD.*

La direction à la formation et à la pédagogie d'une université de technologie de l'Oise décide de chercher à connaître la proportion d'étudiants ayant déjà triché à un examen dans leur formation.

Il apparaît comme immédiat qu'il faut que certaines précautions soient prises pour que les étudiants répondent honnêtement. Un processus dit confidentiel est mis en place.

Pour cela, la question suivante est posée : « *Lancez une pièce, si vous obtenez pile, répondez à la question A, si vous obtenez face, répondez à la question B.* »

Avec :

— A : Avez-vous déjà triché à un examen dans votre formation ?

— B : Lancez-à nouveau la pièce, avez-vous obtenu face ?

Sont collectées uniquement les réponses, sans savoir à quelle question les étudiants ont répondu.

Soit Y le nombre de « oui ». Soit n le nombre d'étudiants interrogés.

Soit le paramètre p la proportion d'étudiants tricheurs.

1. Si un étudiant a répondu « oui », pouvez-vous déterminer si il triche à un examen ?
2. Quelle est la loi de Y ?
3. Construire un estimateur du maximum de vraissemblance pour estimer p.
4. Quelle est le biais et la variance de cet estimateur ?
5. Si Y était la réponse à la question A et non ce processus confidentiel, quel estimateur auriez-vous obtenu ? Quel est son biais et sa variance ?
6. Quel facteur d'étudiant en plus devez-vous interroger pour avoir la même précision dans le pire des cas (quel que soit p) avec le processus « confidentiel » par rapport au processus ordinaire ?

Exercice 5.7 *Non traité en TD.*

Dans un supermarché, on mesure chaque jour le nombre de vols. De plus on connait pour chaque jour, le nombre de clients. On note x_i le nombre de clients d'un jour i, on considère que c'est une donnée qui n'est pas aléatoire, et on note y_i la réalisation de Y_i le nombre de vols sur la journée i.

Les données sont mesurés sur n jours.

Le premier choix de modélisation est de considérer que le nombre de vol est en moyenne proportionnel aux nombre de clients.

1. Quels sont les hypothèses à faire pour considérer Y_i comme suivant une distribution de Poisson ?
2. Écrire un modèle, avec Y_i suivant une distribution de Poisson, de telle manière de respecter le premier choix de modélisation. On introduira le paramêtre λ comme constante de proportionalité.
3. Calculer la log–vraisemblance du paramètre λ. On prêtera une attention particulière au fait que les variables ne sont pas identiquement distribuées.
4. En déduire l'expression de $\hat{\lambda}_{\mathrm{MV}}$, estimateur du maximum de vraisemblance de λ.
5. Calculer l'espérance et la variance de $\hat{\lambda}_{\mathrm{MV}}$.
6. On appelle estimateur des moindres carrés du paramètre λ l'estimateur $\hat{\lambda}$ qui minimise la fonction
$$F(\lambda) = \sum_{i=1}^{n} (Y_i - \lambda x_i)^2.$$
Donner l'expression de cet estimateur.
7. Montrer que $\hat{\lambda}$ est un estimateur sans biais de λ, et calculer sa variance.
8. On admet que, pour tout $x_1, \dots, x_n$
$$\left(\sum_{i=1}^{n} x_i^2\right)^2 \leq \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i^3\right). \tag{0.1}$$
On rappelle que le risque d'un estimateur peut s'exprimer comme la somme de sa variance et du carré de son biais
$$R\left(\hat{\lambda}\right) = b\left(n, \hat{\lambda}\right)^2 + \operatorname{Var} \hat{\lambda}$$
De $\hat{\lambda}$ et $\hat{\lambda}_{\mathrm{MV}}$, lequel a le plus petit risque ? Justifier.

Exercice 5.8 *Non traité en TD.*

Soit X_1, X_2, X_3, tel que :

— $\forall i \in \{1, 2, 3\} \qquad X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$

— (X_1, X_2, X_3) est gaussien

— $\operatorname{Cor}(X_1, X_2) = \operatorname{Cor}(X_2, X_3) = \rho$ et $\operatorname{Cor}(X_1, X_3) = \rho^2$.

σ^2 et ρ sont supposés connus. On suppose également les données positivement corrélés $\rho \geq 0$.

La loi jointe de X_1, X_2, X_3 est donnée par :

$$f : (x_1, x_2, x_3) \mapsto C \exp\left(-\frac{(x_1 - \mu)^2 + \left(1 - \rho^2\right)(x_2 - \mu)^2 + (x_3 - \mu)^2 - \rho((x_1 - \mu)(x_2 - \mu) + (x_2 - \mu)(x_3 - \mu))}{2(1 + \rho)(1 - \rho)\sigma^2} \right)$$

Où C est une constante dépendant que de ρ et σ^2.

1. Montrer que si $\rho = 0$ la loi jointe de X_1,X_2,X_3 est bien ce qu'il aurait été trouvé si les variables étaient indépendantes.
2. Construire $\hat{\mu}$ l'estimateur de μ par la méthode du maximum de vraissemblance.
3. Que remarquez-vous quand à l'influence respective des X_1, X_2, et X_3 dans $\hat{\mu}$ en fonction de ρ, comparez au cas indépendant.
4. Calculez le biais et la variance de $\hat{\mu}$. Étudiez suivant la valeur de ρ, commentez.

Chapitre 6

Estimation par intervalle de confiance

Exercice 6.1 *À préparer intégralement.*

Dans une usine de production mécanique, une machine produit en série des tiges métalliques dont la longueur, par suite de l'imperfection du procédé, peut être considérée comme une v.a. X.

Un client reçoit un lot de 10000 tiges. Il se propose d'estimer la valeur de la moyenne à partir d'un échantillon de n tiges prélevées aléatoirement dans ce lot. On note $X_1, \dots, X_n$ les longueurs correspondantes.

1. En supposant que l'imperfection du procédé est la résultante d'imperfections à chaque étapes de la fabrication, proposez un modèle sur l'échantillon $X_1, \dots, X_n$. Énoncez les hypothèses que vous utilisez.
2. Donner sans démonstration l'expression d'estimateurs sans biais de la moyenne et de la variance.
3. En supposant connue la valeur de la variance, donner l'expression d'une fonction pivotale sur la moyenne puis en déduire un intervalle de confiance bilatéral I_1 au niveau $1-\alpha$ pour la moyenne.
4. Calculer un intervalle de confiance bilatéral I_2 comme précédemment mais en supposant la variance inconnue.
5. Donner un intervalle de confiance unilatéral I_3 pour la variance de la forme $[T, \infty[$, au niveau $1-\alpha$ (on suppose que la moyenne reste inconnue).
6. A. N. : On a obtenu pour un échantillon de $n = 10$ pièces les résultats suivants : $\sum_{i=1}^{10} x_i = 229,9$ et $\sum_{i=1}^{10} x_i^2 = 5285,6$. En déduire une estimation de la moyenne et de la variance, puis calculer numériquement les réalisations des intervalles I_2 et I_3 calculés aux questions précédentes, avec $1-\alpha = 0,95$.
7. En supposant que $\sqrt{\operatorname{Var} X} = 0.1$, quelle valeur faudrait-il donner à n pour que la longueur de l'intervalle de confiance bilatéral I_1 sur μ au niveau 0.95 n'excède pas 0.05 ?

Exercice 6.2

Sur un réseau de parcelles réparties sur l'ensemble de la Bretagne, dans le cadre d'un projet de recherche commun entre l'INRA et la chambre d'agriculture de Bretagne, diverses mesures ont été réalisées sur un ensemble de parcelles, la variable d'intérêt est la teneur en argile de l'horizon de surface (0–30cm) exprimée en g/kg. Les agronomes ont besoin de connaitre la médiane théorique de cette valeur. Voici le début des données ainsi que les résumés :

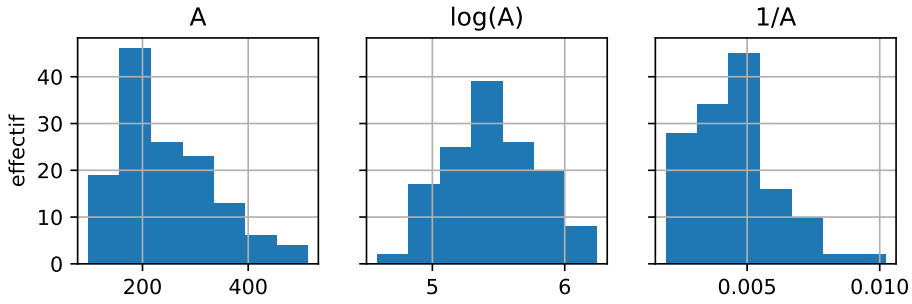
i	Parcelle	Commune	Dep	Arg (A_i)
1	220100B	Ploufragan	22	159
2	220100C	Ploufragan	22	152
3	220101	Ploufragan	22	158
4	220102	Trémuson	22	180
5	220200	Mur de Bretagne	22	199
6	220201	Mur de Bretagne	22	228
...	...	...	...	...

$$\sum_{i=1}^{137} a_i = 33950 \qquad \sum_{i=1}^{137} a_i^2 = 9454722$$

$$\sum_{i=1}^{137} \log a_i = 747.115 \qquad \sum_{i=1}^{137} (\log a_i)^2 = 4090.514$$

$$\sum_{i=1}^{137} \frac{1}{a_i} = 0.621923 \qquad \sum_{i=1}^{137} \frac{1}{a_i^2} = 0.00316082$$

Et quelques histogrammes ont été tracés sur les données :



Écrire un modèle sur les données et en déduire un intervalle de confiance autour de la médiane pour ce modèle. Une attention particulière sera portée à la justification de toutes les hypothèses que vous devrez ajouter, et à la justification du modèle que vous choisissez. Appliquer numériquement en fin de raisonnement.

Exercice 6.3

Un fabricant d'appareil d'instrumentation opératoire fabrique un bras robotisé destiné à tenir un bistouri afin de réaliser des incisions.

Afin de respecter le protocole opératoire, pour connaitre la profondeur du bistouri, notée ξ, le fabricant installe des capteurs pour mesurer cette profondeur. Afin d'assurer une certaine fiabilité, le fabricant choisit d'installer 7 capteurs.

À un instant de temps, on note $X_1, \dots, X_7$ les mesures de la profondeur.

Le fabricant désire avoir un intervalle de confiance bilatéral de ξ au niveau $1-\alpha$ avec $\alpha = 10^{-2}$ (en réalité bien moins, mais les valeurs des tables ne permettent pas de faire les applications numériques). De plus il considère que sous-estimer la profondeur du bistouri est une erreur bien plus grave que la sur-estimer, car la sous-estimation conduit l'automate à inciser plus profondément. Il impose donc que la probabilité que l'intervalle sous-estime soit 10 fois plus faible que la probabilité que l'intervalle sur-estime.

À un instant, on a observé les mesures suivantes en millimètres :

$$12.437 \quad 12.335 \quad 12.468 \quad 12.239 \quad 12.377 \quad 12.468 \quad 12.284$$

1. Proposez un modèle.
2. Construire un intervalle de confiance repectant les contraintes du fabricant.
3. Appliquez numériquement. Commentez l'asymétrie de l'intervalle.

Exercice 6.4

On note $X_1, \dots, X_n$ les nombres de plaintes au commissariat des riverains consécutives à n soirées étudiantes.

1. Proposez un modèle sur la variable X.
2. En utilisant le TCL, proposer une fonction asymptotiquement pivotale pour la moyenne notée λ.
3. En utilisant le théorème de Slutsky, montrer que $\frac{\overline{X}-\lambda}{\sqrt{\overline{X}/n}}$ est une fonction asymptotiquement pivotale pour λ.
4. En déduire l'expression d'un intervalle de confiance bilatéral approché au niveau $1-\alpha$.
5. A.N. Aux mois de mai et juin, 20 soirées (officielles) ont été organisées ; au total, 73 plaintes ont été déposées. Quelle est la réalisation de l'intervalle, au niveau de confiance de 95% ?
6. Reprendre le calcul de l'intervalle de confiance sans utiliser le théorème de Slutsky.

Exercice 6.5 *Non traité en TD.*

Soit X une v. a. suivant une loi de densité $f(x) = \theta^2 x e^{-\theta x} \mathbb{1}_{[0,+\infty[}(x)$ avec $\theta > 0$ et $X_1, \dots, X_n$ un échantillon i.i.d. de variable parente X. Une réalisation de cet échantillon pour $n = 100$ a donné des valeurs x_i vérifiant $\sum_{i=1}^{n} x_i = 398$.

1. Déterminer un estimateur du maximum de vraisemblance de θ, que l'on notera $\hat{\theta}$.
2. Quelle est l'information de Fisher associée à θ ? En déduire une fonction asymptotiquement pivotale pour θ et un intervalle de confiance bilatéral approché sur θ au niveau de confiance $1 - \alpha$.
3. A.N. Calculer la réalistation de l'intervalle de confiance pour $1 - \alpha = 95\%$.

Exercice 6.6 *Non traité en TD.*

Pour étudier la teneur en fer de 400kg de minerai, on a prélevé au hasard un échantillon de 20 prises de 1.6kg sur lesquelles on a mesuré la teneur en fer. $\overline{X}$ et S^* prennent les valeurs 0.2486 et 0.0028.

1. Modéliser le problème.
2. Déterminer l'intervalle de confiance bilatéral sur la moyenne à 99%.
3. Quelle est la taille minimale de l'échantillon qu'il faut prendre pour approcher la moyenne à 0.001 près dans un intervalle à 95% ?
4. Donner un intervalle de confiance inférieur ($[0, a]$) au niveau 0.99 pour la variance à partir des mesures effectuées.

Exercice 6.7 *Non traité en TD.*

Un formulaire supplémentaire est disponible à la fin de cet énoncé.

On considère une variable aléatoire X à valeurs dans $\{1, 2, 3\}$ de loi de probabilité

$$\mathbb{P}(X = k) = \begin{cases} 1 - \lambda - \lambda^2 & \text{si } k = 1, \\ \lambda^2 & \text{si } k = 2, \\ \lambda & \text{si } k = 3. \end{cases}$$

Soit $X_1, \dots, X_n$ un échantillon iid de variable parente X, et N_k le nombre de variables X_i dans l'échantillon égales à k :

$$N_k = \operatorname{card}\{i = 1, \dots, n \,:\, X_i = k\}.$$

On notera x_i la réalisation de X_i, et n_k la réalisation de N_k ; on a observé $(n_1, n_2, n_3) = (31, 24, 45)$.

1. Quel est l'intervalle de validité de λ ? Justifier.
2. Calculer l'espérance de X.
3. Calculer un estimateur $\hat{\lambda}_m$ de λ par la méthode des moments, et calculer sa réalisation avec les données de l'énoncé.

4. L'estimateur $\hat{\lambda}_m$ est-il sans biais ? Justifier.
5. Sachant que $\mathrm{Var}\, X = h(\lambda)$ avec $h(\lambda) = \lambda(4 - 3\lambda - 4\lambda^2 - \lambda^3)$, calculer la loi asymptotique de $\overline{X}$, et en déduire la loi asymptotique de $\hat{\lambda}_m$ (on conservera $h(\lambda)$ dans l'expression).
6. Proposer une fonction asymptotiquement pivotale pour λ fonction de $\hat{\lambda}_m$ et $h(\hat{\lambda}_m)$. En déduire un intervalle de confiance asymptotique de niveau $1 - \alpha = 0.95$; calculer sa réalisation.
7. Montrer que $\ln L(\lambda; x_1, \dots, x_n) = \ln L(\lambda; n_1, n_2, n_3) = n_1 \ln(1 - \lambda - \lambda^2) + (2n_2 + n_3)\ln(\lambda)$.
8. Calculer $\ln L(\lambda; n_1, n_2, n_3)$ avec les données observées, puis l'estimation de λ correspondante.
9. Calculer l'estimateur du maximum de vraisemblance $\hat{\lambda}_{MV}$ dans le cas général.
10. Quelle est la loi asymptotique de $\hat{\lambda}_{MV}$?

Calcul des racines d'un polynôme du second degré. Soit $g(x) = ax^2 + bx + c$. Le discriminant est $\Delta = b^2 - 4ac$. Les racines du polynôme dépendent de la valeur de Δ :

— si $\Delta > 0$, le polynôme admet deux racines simples x_{01} et x_{02} :

$$x_{01} = \frac{-b - \sqrt{\Delta}}{2a} \quad \text{et} \quad x_{02} = \frac{-b + \sqrt{\Delta}}{2a};$$

il se factorise donc sous la forme $g(x) = a(x - x_{01})(x - x_{02})$; il est alors du même signe que a en-dehors de ses racines, et du signe opposé entre ses racines.

— Si $\Delta = 0$, il admet alors une racine double $x_0 = -b/(2a)$ et s'écrit donc $g(x) = a(x - x_0)^2$; il est du même signe que a pour tout $x \neq x_0$.

— Si $\Delta < 0$, il n'admet alors pas de racine réelle, et est du même signe que a pour tout x.

Inégalités de Jensen. Si g est une fonction strictement convexe et Y est une variable aléatoire non dégénérée (de variance non-nulle), alors

$$\mathbb{E}g(Y) > g(\mathbb{E}Y);$$

si g est une fonction strictement concave et Y est une variable aléatoire non dégénérée, alors

$$\mathbb{E}g(Y) < g(\mathbb{E}Y).$$

Pour rappel, une fonction convexe (respectivement, concave) est telle que pour toute paire de points (A, B) du graphe de la fonction, le segment $[AB]$ est entièrement situé au-dessus (respectivement, en-dessous) du graphe.

Convergence d'une fonction d'une variable aléatoire asymptotiquement normale. Soit Y_n une suite de variables aléatoires non dégénérées telle que $\lim \mathrm{Var}\, Y_n = 0$ $(n \to +\infty)$, et une fonction g continument dérivable en $\lim \mathbb{E}Y_n$ $(n \to +\infty)$; si

$$\frac{Y_n - \mathbb{E}Y_n}{\sqrt{\mathrm{Var}\, Y_n}} \quad \xrightarrow{\mathscr{L}} \quad \mathcal{N}(0, 1),$$

alors

$$\frac{g(Y_n) - g(\mathbb{E}Y_n)}{|g'(\mathbb{E}Y_n)|\sqrt{\operatorname{Var} Y_n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

Exercice 6.8 *Non traité en TD.*
Un formulaire supplémentaire est disponible à la fin de cet énoncé.

Soit X une variable aléatoire suivant une loi de Laplace de paramètre $c \in \mathbb{R}$ et $\Delta > 0$ définie par la densité de probabilité suivante

$$f(x; c, \Delta) = \frac{1}{2\Delta} \exp\left(-\frac{|x-c|}{\Delta}\right).$$

On utilise souvent cette distribution pour modèliser des phénomènes où les queues de distributions sont plus lourdes que celles de la loi normale.
On dispose d'un échantillon iid $X_1, \dots, X_n$ de variable parente X.

1. Tracer la fonction de densité f en faisant apparaître c et Δ.
2. Méthode des moments.
   (a) Montrer que $\mathbb{E}(X) = c$ et $\operatorname{Var}(X) = 2\Delta^2$. En déduire $\mathbb{E}\left(X^2\right)$. On pourra admettre et utiliser le fait que $|X - c|$ suit une loi exponentielle de paramètre $1/\Delta$.
   (b) Construire un estimateur $\hat{c}$ pour c et $\widehat{\Delta}$ pour Δ à partir des moments $\mathbb{E}(X)$ et $\mathbb{E}\left(X^2\right)$.
   (c) Les estimateurs $\hat{c}$ et $\widehat{\Delta}$ sont-ils biaisés ?
   (d) En supposant Δ connu, donner une loi asymptotique de $\hat{c}$ et en déduire un intervalle de confiance asymptotique bilatéral sur c.
3. Méthode du maximum de vraisemblance.
   On souhaite maintenant appliquer la méthode du maximum de vraisemblance. On reprend l'échantillon $X_1, \dots, X_n$ et on introduit $X_{(1)}, \dots, X_{(n)}$ l'échantillon trié par ordre croissant.
   (a) Les variables $X_{(1)}$ et $X_{(2)}$ sont-elles indépendantes ? Justifier rapidement.
   (b) Écrire la fonction de log–vraisemblance des paramètres en fonction des observations et montrer que maximiser la vraisemblance pour c revient à minimiser la fonction

$$g(c; x_1, \dots, x_n) = \sum_{i=1}^{n} |x_{(i)} - c|. \tag{0.1}$$

   (c) On suppose dans cette question que $n = 3$. Tracer la fonction (0.1) correspondante en l'étudiant sur les intervalles $]-\infty; x_{(1)}]$, $]x_{(1)}; x_{(2)}]$, $]x_{(2)}; x_{(3)}]$ et $]x_{(3)}; +\infty[$ puis donner l'estimateur du maximum de vraisemblance dans ce cas.
   (d) On admet que pour tout n les maximums de vraisemblance sont les médianes. Calculer l'information de Fisher $I_1(c)$ en supposant que $n = 1$ et en utilisant uniquement la dérivée première de la log–vraisemblance.

(e) En utilisant le fait que $I_n(c) = nI_1(c)$, en déduire une loi asymptotique de $\hat{c}_{\mathrm{MV}}$. En déduire un intervalle de confiance asymptotique sur c en supposant Δ connu.

4. Comparer les estimateurs $\hat{c}$ et $\hat{c}_{\mathrm{MV}}$ en comparant les intervalles de confiance obtenus. Commenter.

Inégalités de Jensen Si g est une fonction strictement convexe et Y est une variable aléatoire non dégénérée (de variance non-nulle), alors

$$\mathbb{E}\left(g(Y)\right) > g(\mathbb{E}\left(Y\right)),$$

si g est une fonction strictement concave et Y est une variable aléatoire non dégénérée, alors

$$\mathbb{E}\left(g(Y)\right) < g(\mathbb{E}\left(Y\right)).$$

Pour rappel, une fonction convexe (respectivement, concave) est telle que pour toute paire de points (A, B) du graphe de la fonction, le segment $[AB]$ est entièrement situé au-dessus (respectivement, en-dessous) du graphe.

Dérivée de $x \in \mathbb{R}^\star \mapsto |x|$ On rappelle que la dérivée de la fonction valeur absolue vaut

$$\frac{\mathrm{d}\,|\cdot|}{\mathrm{d}\,x} = \operatorname{sign}(x) = \begin{cases} 1 & \text{si } x > 0 \\ -1 & \text{si } x < 0 \end{cases}$$

Chapitre 7

Régression linéaire

Exercice 7.1 *À préparer intégralement.*

On fait passer dans un résistor de résistance r un courant parfaitement connu noté x_i, et on mesure la différence de potentiel U_i aux bornes du résistor.

On suppose que la valeur U_i est bruitée, et on considère donc qu'il s'agit d'une variable aléatoire.

La mesure est repetée n fois. On a donc $x_1, \dots, x_n$ les courants, et $U_1, \dots, U_n$ les v.a. représentants les tensions, et $u_1, \dots, u_n$ les tensions mesurées.

On suppose que la variance de l'erreur ne dépend pas de i.

1. En utilisant la loi d'Ohm (Tension = Résistance × Courant), modéliser U_i. On notera σ^2 la variance de l'erreur.
2. Expliciter les estimateurs du maximum de vraisemblance des paramètres r et σ^2, que l'on notera respectivement $\hat{r}$ et $\widehat{\sigma^2}$.
3. Montrer que $\hat{r}$ est sans biais. Calculer sa variance. Quelle est la loi de $\hat{r}$?
4. En supposant que σ^2 connu, donner l'expression d'un intervalle de confiance de niveau $1-\alpha$ pour r.
5. On a obtenu les résultats suivants :

x_i (A)	1	2	3	4	5	7	10	13	16	20
u_i (V)	17.8	36.3	54.5	76.8	89.2	118.1	187.3	232.4	298.9	345.7

   En supposant $\sigma^2 = 1V^2$, calculer une estimation de r ainsi qu'un intervalle de confiance de niveau 95 % pour r puis tracer les points (x_i, u_i) ainsi que la droite de régression estimée.
6. Quelle tension U_0 peut-on s'attendre à mesurer pour un courant d'intensité $x_0 = 50A$? Toujours en supposant $\sigma^2 = 1V^2$, calculer un intervalle de confiance de niveau 95% pour $\mathbb{E}(U_0)$ ainsi qu'un intervalle de prédiction.

Exercice 7.2

On cherche à explorer le lien entre la consomation d'un véhicule et sa vitesse.

On place une Citroën Ami 8 sur un circuit de 10km, et ce circuit est effectué à différentes vitesses. Après chaque expérience, la quantité de carburant consommé est mesurée, et la consommation est calculée, on obtient les données suivantes :

Vitesse (km/h)	Consommation (L/100km)
30	0.24
40	0.20
50	0.38
60	4.58
70	7.12
80	2.49
90	6.17
100	1.80
110	16.15

Le constructeur spécifie que la voiture peut rouler jusqu'à 125 km/h. Toutefois pour des raisons de sécurité, aucun conducteur ne veut conduire cette expérience avec ce type de véhicule. L'objectif de l'exercice est donc d'obtenir un intervalle de confiance sur la consommation à cette vitesse.

1. Modèle simple.
   (a) Tracer la consommation en fonction de la vitesse.
   (b) Écrire un modèle linéaire simple pour expliquer la consommation en fonction de la vitesse. Discuter des hypothèses.
   (c) Pour le modèle de regression linéaire simple, on donne $\hat{a}_{\text{réa}} = 4.744$ et $\hat{b}_{\text{réa}} = .1299$, $R^2_{\text{réa}} = 0.482$, ainsi que le tableau de résultats suivant :

Vitesse (km/h)	Conso (L/100km)	Levier	Prédiction (L/100km)	Résidu (L/100km)	Résidu standardisé	Résidu studentisé
30	0.24	0.38	-0.85	1.09	0.35	0.33
40	0.20	0.26	0.45	-0.25	-0.07	-0.07
50	0.38	0.18	1.75	-1.37	-0.38	-0.36
60	4.58	0.13	3.05	1.53	0.42	0.39
70	7.12	0.11	4.35	2.77	0.75	0.72
80	2.49	0.13	5.65	-3.16	-0.86	-0.84
90	6.17	0.18	6.95	-0.78	-0.22	-0.20
100	1.80	0.26	8.24	-6.44	-1.90	-2.53
110	16.15	0.38	9.54	6.61	2.12	3.30

   Y-a-t'il des abscisses considérées comme extrèmes ? Lesquelles ?
   (d) Y-a t'il des points aberrants ?
   (e) La normalité est-elle une hypothèse acceptable ?
   (f) L'homoscédasticité est-elle une hypothèse acceptable ?

2. On suppose maintenant que la consommation est une fonction exponentielle d'une fonction linéaire de la vitesse. Ainsi on va chercher à expliquer $D_i = \ln C_i$ en fonction de v_i.

   (a) Écrire un modèle linéaire simple pour expliquer la log-consommation en fonction de la vitesse. Discuter des hypothèses.

   (b) Pour le modèle de regression linéaire simple, on donne $\hat{a}_{\text{réa}} = -2.692$ et $\hat{b}_{\text{réa}} = 4.732 \cdot 10^{-2}$, $R^2_{\text{réa}} = 0.655$, ainsi que le tableau de résultats suivant :

Vitesse	Conso	Log-conso	Levier	Prédiction	Résidu	Résidu standardisé	Résidu studentisé
30	0.24	-1.43	0.38	-1.27	-0.15	-0.19	-0.18
40	0.20	-1.61	0.26	-0.80	-0.81	-0.94	-0.93
50	0.38	-0.97	0.18	-0.33	-0.64	-0.70	-0.68
60	4.58	1.52	0.13	0.15	1.37	1.47	1.63
70	7.12	1.96	0.11	0.62	1.34	1.42	1.55
80	2.49	0.91	0.13	1.09	-0.18	-0.19	-0.18
90	6.17	1.82	0.18	1.57	0.25	0.28	0.26
100	1.80	0.59	0.26	2.04	-1.45	-1.68	-2.02
110	16.15	2.78	0.38	2.51	0.27	0.34	0.32

   Y-a-t'il des abscisses considérées comme extrèmes ? Lesquelles ?

   (c) Y-a t'il des points aberrants ?

   (d) La normalité est-elle une hypothèse acceptable ?

   (e) L'homoscédasticité est-elle une hypothèse acceptable ?

3. À partir du modèle le plus approprié, établir un intervalle de confiance sur la consommation à 125 km/h. Appliquer numériquement.

Exercice 7.3 *Non traité en TD.*

Le revenu R_t et l'épargne nette E_t ont été mesurés trimestre par trimestre pendant trois ans. Après correction des variations saisonnières, on dispose des indications suivantes :

$$\sum_{t=1}^{12} R_t = 19.7, \quad \sum_{t=1}^{12} R_t^2 = 4827, \quad \sum_{t=1}^{12} E_t = 6.1$$

$$\sum_{t=1}^{12} E_t^2 = 456, \quad \sum_{t=1}^{12} R_t E_t = 1480$$

1. Poser le modèle de régression linéaire et estimer les paramètres.
2. Quelle est la qualité de ce modèle ?
3. Quelle épargne peut-on attendre lorsque l'on a respectivement 20 et 50 de revenu ? Calculer un intervalle de prédiction pour ces valeurs.

Exercice 7.4 *Non traité en TD.*

On a relevé dans le tableau suivant les moyennes x au baccalauréat de 10 élèves, et leurs scores Y à un test de QI :

x	8.8	9.6	11.2	10.4	12.8	15.2	12.0	16.0	8.0	9.2
Y	108	112	115	118	121	125	122	130	96	113

On suppose que les Y_i sont des v.a. indépendantes avec $Y_i \sim \mathcal{N}(a + bx_i, \sigma^2)$, les x_i étant des nombres fixés.

1. Déterminer les estimateurs du maximum de vraisemblance des paramètres a, b et σ^2.
2. Construire un intervalle de confiance bilatéral sur a, puis une borne inférieure, au niveau de confiance 95 %.
3. Construire un intervalle de confiance bilatéral sur b, puis une borne supérieure, au niveau de confiance 95 %.
4. *Question à réaliser après les deux premiers chapitres sur les tests.* Tester l'hypothèse $H_0 : b = 0$ contre $H_1 : b > 0$ au niveau de signification $0,01$.

Exercice 7.5 *Non traité en TD.*

La masse volumique d'un moût en fermentation permet de connaître l'avancement de la fermentation alcoolique. Pour cela, on fait n prélèvements dont on mesure très précisément le volume, et dont on mesure la masse. On suppose que seule la mesure de masse est entâchée d'erreurs. On obtient les données suivantes :

Volume (L)	Masse (g)
0.636	637.8
1.486	1492.5
0.351	357.3
0.526	533.6
0.225	228.7
0.351	355.8
1.246	1257.3
0.521	528.1
0.387	393.2
2.570	2619.5

1. Modèle simple :

   Si on suppose l'erreur centrée, et la masse volumique constante, on a :

   $$\mathbb{E}\, M_i = \rho v_i$$

   Les mesures étant faite de manière indépendante, et en l'absence d'autres hypothèses, nous faisons l'hypothèse de normalité. Le modèle proposé est donc :

   $$\forall i \in \{1, \dots, n\}, \qquad M_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\rho v_i, \sigma^2\right)$$

(a) Tracer les réalisations de M_i en fonction de v_i, le modèle vous paraît-il adapté ?

(b) Écrire pour ce modèle l'estimateur du maximum de vraisemblance de ρ.

(c) Calculer la réalisation des résidus (r_i) et des résidus corrigés $\tilde{r}_i = r_i/(1-h_i)$ avec $h_i = \frac{v_i^2}{\sum_j v_j^2}$

(d) Tracer les résidus corrigés en fonction du volume. Le choix du modèle vous paraît-il pertinent ?

2. Après discussion avec l'opérateur de mesure, on aboutit à la conclusion que la variance de l'erreur sur la masse est proportionnelle au volume :

$$\operatorname{Var} M_i = cv_i$$

Le modèle obtenu est donc :

$$\forall i \in \{1, \dots, n\}, \qquad M_i \overset{\text{ind}}{\sim} \mathcal{N}(\rho v_i, cv_i)$$

(a) Est-ce un modèle linéaire usuel ?

(b) Soit $N_i = \frac{M_i}{\sqrt{v_i}}$, écrire la loi de M_i, que reconnaissez-vous ?

(c) Exhiber l'estimateur du maximum de vraisemblance de ρ, en fonction des N_i puis en fonction des M_i. Ce résultat vous paraît-il cohérent ?

(d) Calculer les résidus sur n_i, les résidus corrigés, les tracer en fonction de la variable explicative, conclure sur l'homoscédasticité.

Exercice 7.6 *Non traité en TD.*

Le but de cet exercice est de tester la réalité d'une relation linéaire entre le chiffre d'affaire et le nombre de salariés d'une entreprise. Les données recueillies sont les suivantes :

Année	Nombre de salariés	Chiffre d'affaire
1957	294	634
1959	314	728
1961	383	819
1963	402	938
1965	475	1136
1967	786	1317

1. Représenter le nuage de points.
2. Poser le modèle et estimer les paramètres.
3. Quelle est la qualité de ce modèle ?
4. Y-a-t-il vraiment une dépendance linéaire entre nombre de salariés et le chiffre d'affaire ?
5. Étudier les résidus de la régression : peut on admettre qu'ils soient la réalisation d'une variable aléatoire gaussienne centrée ?

Exercice 7.7 *Non traité en TD.*

L'échantillon suivant représente la masse et la taille de 13 étudiants tirées au hasard :

masse	70	63	72	60	66	70	74	65	62	64	67	65	68
taille	155	150	180	135	156	168	178	160	152	198	145	139	152

1. Représenter le nuage de points.
2. Poser le modèle et estimer les paramètres.
3. Calculer les résidus, les résidus corrigés. Étudier la variance des résidus corrigés en les représentant, étudiez également leur distribution au moyen d'un diagramme quantile-quantile.
4. S'il se trouve des points aberrants, les éliminer et recalculer les paramètres du modèle.
5. Estimer la masse moyen d'un étudiant mesurant 1 mètre 68. Quelle est la précision de cette prédiction ?

Chapitre 8

Tests : Principes et Théorème de Neyman-Pearson

Exercice 8.1 *À préparer intégralement.*

On dispose d'un échantillon de taille $n = 10$ de v.a. parente X normale de moyenne 0 et de variance σ^2 inconnue. On notera dans tout ce problème T la statistique $\sum_{i=1}^{n} X_i^2$. On considère le problème de test suivant :

$$H_0 : \sigma^2 = \sigma_0^2 \quad (= 1)$$
$$H_1 : \sigma^2 = \sigma_1^2 \quad (= 2).$$

1. Déterminer la région critique optimale pour $\alpha^* = 0.05$.
2. Calculer la puissance du test.

Exercice 8.2

Un fabricant de lave-linges assure que la durée de vie moyenne de ses produits est de 10 ans.

Dans le cadre de la lutte contre l'obsolescence programmée, une association de consommateur suspecte l'utilisation frauduleuse du circuit BZT1231 au sein de la carte électronique. Il est documenté sur d'autres produits que lorsque ce composant est utilisé alors la durée de vie moyenne est ramenée à 6 ans.

On supposera que les durées de vies peuvent être modélisées par des lois exponentielles.

L'étude porte sur $n = 9$ lave-linges dont les durées de vie ont été mesurées en années :

16.7 1.1 0.2 4.5 6.2 19.0 1.0 0.2 0.1

Peut-on accuser le fabricant de fraude ?

Exercice 8.3

Une v.a. aléatoire X suit une loi $\mathcal{N}(\mu, \sigma^2)$ d'écart-type connu $\sigma = 2$. Au vu d'un échantillon i.i.d. $X_1, \dots, X_n$ de v.a. parente X, on veut tester l'hypothèse $H_0 : \mu = 0$ contre l'hypothèse alternative $H_1 : \mu = \mu_1$ où $\mu_1 > 0$.

1. Soit $L(\mu; x_1, \dots, x_n)$ la fonction de vraisemblance. Donner l'expression du rapport $\frac{\mathscr{L}_1}{\mathscr{L}_0} = \frac{L(\mu_1; x_1, \dots, x_n)}{L(\mu_0; x_1, \dots, x_n)}$.
2. En déduire la région critique du test de Neyman-Pearson, en fonction du niveau de signification α^*.
3. Calculer la puissance de ce test en fonction de μ_1, n et α^*.
4. Pour $n = 100$ et $\alpha^* = 0.05$, calculer la puissance du test pour les valeurs $\mu_1 = 0.1,\ 0.2,\ 0.5,\ 0.8,\ 1$ et tracer la courbe de puissance en fonction de μ_1.
5. Pour $\mu_1 = 1$ et $\alpha^* = 0.05$, calculer la puissance du test pour les valeurs $n = 10,\ 20,\ 50,\ 80,\ 100$ et tracer la courbe de puissance en fonction de n.
6. Toujours pour $\mu_1 = 1$ et $\alpha^* = 0.05$, quelle doit être la taille minimale n_0 de l'échantillon pour que la puissance soit supérieure à 0.95 ?
7. On considère maintenant $\mu_1 = 1$ et $n = 100$. On a observé $\overline{x} = 0.4$, quel est le résultat du test d'hypothèse pour les valeurs $\alpha^* = 0.1,\ 0.05,\ 0.01$.
8. Calculer le degré de signification du test et conclure à nouveau à la question précédente.

Exercice 8.4 *Non traité en TD.*

Soit X une variable aléatoire discrète obéissant à la loi géométrique :

$$\mathbb{P}(X = x) = p(1-p)^{x-1},$$

pour $x = 1, 2, \dots, \infty$ et $p \in]0, 1[$. On rappelle que $\mathbb{E}[X] = 1/p$ et $\mathrm{Var}[X] = (1-p)/p^2$. On considère un échantillon iid $X_1, \dots, X_n$ de variable aléatoire parente X et le problème de test

$$\begin{aligned} H_0 &: \quad p = p_0 \\ H_1 &: \quad p = p_1 \end{aligned}$$

avec $p_1 > p_0$. Montrer que la région critique W du test optimal au niveau α^* s'exprime en fonction de $\overline{x}$, la réalisation de $\overline{X}$. Donner une approximation de W en supposant n grand.

Exercice 8.5 *Non traité en TD.*

La durée de vie de tubes fluorescents d'un certain modèle est supposée obéir à une loi normale d'écart-type égal à 150 heures et de moyenne μ inconnue. La durée de vie moyenne d'un échantillon de 100 tubes a été trouvée égale à 1570 heures.

1. Tester l'hypothèse $\mu = 1600$ heures relativement à l'hypothèse $\mu = 1500$ heures avec un niveau de signification de α^*. Appliquer pour $\alpha^* = 0.05$ et $\alpha^* = 0.01$, comparer les régions critiques. Conclure sur le test.
2. Calculer la *p-value* de ce test. Conclure sur le test.

Chapitre 9

Tests : UPP, rapport de vraisemblance, regression linéaire

Exercice 9.1 *À préparer intégralement.*

Soit $X_1, \dots, X_n$ un échantillon i.i.d. de variable parente X, de densité

$$f(x) = \frac{x}{\theta^2} \exp\left(-\frac{x}{\theta}\right) \mathbb{1}_{[0,+\infty[}(x),$$

θ étant un paramètre strictement positif.

1. Calculer un estimateur du maximum de vraisemblance $\hat{\theta}$ de θ.
2. Calculer l'espérance et la variance de X et en déduire une fonction asymptotiquement pivotale de θ.
3. On considère le problème de test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ avec $\theta_1 > \theta_0$. Montrer que la région critique W du test le plus puissant pour ce problème au niveau α^* s'exprime en fonction de $\hat{\theta}$, puis donner une approximation de W en supposant n grand.
4. On considère maintenant le problème de test suivant $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Existe-t-il un test UPP pour ce problème ?
5. Calculer la statistique du rapport de vraisemblance $\Lambda(X_1, \dots, X_n)$, exprimée en fonction de $\hat{\theta}$, pour le problème de test de la question précédente.
6. En utilisant la statistique $-2 \ln \Lambda(X_1, \dots, X_n)$ et en supposant que n est grand, proposer une région critique pour le test de la question (4). Quelle décision prendra-t-on si $\theta_0 = 2$, $n = 50$, $\sum_i x_i = 115$ et $\alpha^* = 0.05$.

Exercice 9.2

Ayant lancé 1000 fois une pièce de monnaie, on a obtenu 450 fois « face ». L'objet de cet exercice est de tester l'hypothèse selon laquelle la pièce est équilibrée.

1. Soit X le nombre de « faces » obtenu au cours de n lancers. On a observé une seule réalisation de X. Calculer l'estimateur du maximum de vraisemblance de p.
2. On considère le problème de test suivant : $H_0 : p = p_0$ contre $H_1 : p \neq p_0$. Calculer l'expression littérale de la statistique $\Lambda(X)$ du test du rapport de vraisemblance, en fonction de X, p_0 et n.
3. En utilisant l'approximation asymptotique $-2\ln\Lambda(X) \sim \chi^2_1$, donner une expression littérale de la région critique du test du rapport de vraisemblance, au niveau α^*.
4. Donner le résultat du test du rapport de vraisemblance avec les données de l'exercice, $p_0 = 1/2$, et $\alpha^* = 5\%$.

Exercice 9.3

Une entreprise alimentaire cherche à vérifier si les sacs de sucre qu'elle utilise dans sa production ont un poids au moins égal à la valeur annoncée. L'entreprise reçoit un lot de sacs et en pèse n. L'échantillon alors obtenu est considéré comme un échantillon i.i.d. dont la variable aléatoire parente X suit une loi normale d'espérance μ et de variance *connue* σ^2.

1. Première solution
   (a) Quelle est la région critique du test ?

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

   (b) A. N. : sachant que $(x_1, \dots, x_n)$ =(49.56, 48.33, 50.13, 50.29, 48.85, 51.19, 50.19, 49.96, 50.33, 50.17) ($\sum_{i=1}^{10} x_i = 499$), $\sigma^2 = 1$, $\mu_0 = 50$ et $\alpha^* = 0.05$, quelle décision prendra-t-on ?
   (c) Quelle est la puissance du test pour la valeur $\mu = 49$?
2. Deuxième solution : on désire prendre une décision en s'appuyant uniquement sur le nombre K de sacs de l'échantillon dont le poids est inférieur ou égal à $\mu_0 - 1$.
   (a) Montrer que K est une variable aléatoire binomiale $B(n, p)$ où p dépend de μ_0, μ et σ.
   (b) Que deviennent les hypothèses H_0 et H_1 ?
   (c) En utilisant les mêmes données que dans la première partie, déterminer la région critique de ce test. (On n'utilisera aucune approximation). Quelle décision prendra-t-on ?
   (d) Calculer la puissance du test pour la valeur $\mu = 49$, commenter.

Exercice 9.4 *Non traité en TD.*

Cet exercice prolonge l'exercice 8.5.

La durée de vie de tubes fluorescents d'un certain modèle est supposée obéir à une loi normale d'écart-type égal à 150 heures et de moyenne μ inconnue. La durée de vie moyenne d'un échantillon de 100 tubes a été trouvée égale à 1570 heures.

1. Refaire le test de l'exercice 8.5 en prenant pour hypothèse alternative $\mu < 1600$ heures, puis $\mu \neq 1600$ heures.

Exercice 9.5 *Non traité en TD.*

Cet exercice prolonge l'exercice 7.3. Les premières questions sont rappelées.

Le revenu R_t et l'épargne nette E_t ont été mesurés trimestre par trimestre pendant 3 ans. Après correction des variations saisonnières, on dispose des indications suivantes :

$$\sum_{t=1}^{12} R_t = 19.7 \quad \sum_{t=1}^{12} R_t^2 = 4827 \quad \sum_{t=1}^{12} E_t = 6.1$$

$$\sum_{t=1}^{12} E_t^2 = 456 \quad \sum_{t=1}^{12} R_t E_t = 1480$$

1. Poser le modèle et estimer les paramètres.
2. Quelle est la qualité de ce modèle ?
3. Y-a-t-il vraiment une dépendance linéaire entre le revenu et l'épargne ?

Exercice 9.6 *Non traité en TD.*

Le but de cet exercice est de tester la réalité d'une relation linéaire entre le chiffre d'affaire et le nombre de salariés d'une entreprise. Les données recueillies sont les suivantes :

Année	Nombre de salariés	Chiffre d'affaire
1957	294	634
1959	314	728
1961	383	819
1963	402	938
1965	475	1136
1967	786	1317

1. Représenter le nuage de points.
2. Poser le modèle et estimer les paramètres.
3. Quelle est la qualité de ce modèle ?
4. Y-a-t-il vraiment une dépendance linéaire entre nombre de salariés et le chiffre d'affaire ?
5. Étudier les résidus de la régression : peut on admettre qu'ils soient la réalisation d'une variable aléatoire gaussienne centrée ?

Exercice 9.7 *Non traité en TD.*

Cet exercice complète l'exercice 7.1. Toutes les notations sont reprises de cet exercice, et les questions sont supposées traités.

Rappel contexte :

La différence de potentiel mesurée aux bornes d'une résistance r traversée par un courant d'intensité x_i $(i = 1, \dots, n)$ est modélisée par une variable aléatoire

$$U_i = rx_i + \varepsilon_i$$

où ε_i est un bruit de mesure supposé suivre une loi normale d'espérance nulle et de variance σ^2. On considère un échantillon indépendant $U_1, \dots, U_n$ de n mesures réalisées pour des intensités $x_1, \dots, x_n$.

1. La lecture du marquage du résistor nous ammène à supposer qu'il s'agit d'une résistance de 18 ohm. En supposant $\sigma^2 = 1$, pouvez-vous infirmer cette hypothèse ? On raisonnera au niveau 1%.

Chapitre 10

Tests : Stratégie heuristique, conformité, regression linéaire

Exercice 10.1 *À préparer intégralement.*

Un industriel affirme que sa production de ballons n'a que 10% de défectueux et pas plus. Pour tester la qualité d'un ballon, on la gonfle jusqu'à la faire exploser. Si l'explosion se produit alors que le ballon a un diamère X inférieur ou égal à 20 cm, on considère que ce ballon est défectueux. On réalise un contrôle sur $n = 30$ ballons pris au hasard.

1. On considère le problème de test suivant :

$$H_0 : d_1 = d_0$$
$$H_1 : d_1 < d_0,$$

   où d_1 représente le premier décile de la loi de X, c'est à dire le fractile d'ordre 0.1. Quelle valeur doit prendre d_0 pour tester l'affirmation de l'industriel ?

2. Montrer qu'il est possible de reformuler le test pour se ramener à un test sur le paramètre θ d'une loi binomiale.

3. On a observé les données ci-dessous :
   28.1 26.1 22.2 24.8 23.5 31.1 28.9 26.7 17.5 26.2 21.5 24.5 22.5 32.3 31.1
   16.5 31.7 29.6 26.7 35.5 19.3 20.5 32.1 10.7 23.7 26.3 25.9 28.5 18.7 31.9
   Quel est le résultat du test d'hypothèse ?

Exercice 10.2

On cherche à déterminer si un analyseur en continu de SO_2 en sortie d'une usine d'incinération de déchets est bien étalonné. Pour cela, on a comparé les valeurs X fournies par l'analyseur aux valeurs Y déterminées en laboratoire pour 10 prélèvements :

analyseur x_i	33	24	18	41	37	28	21	43	30	8
laboratoire y_i	33	25	19	43	39	27	21	44	29	9
différence z_i	0	-1	-1	-2	-2	1	0	-1	1	-1

1. On suppose que $Z = X - Y \sim \mathcal{N}(\mu, \sigma^2)$ avec $\sigma = 1$. On désigne par H_0 l'hypothèse selon laquelle l'analyseur est bien étalonné, et par H_1 l'hypothèse contraire. Formuler H_0 et H_1 en fonction de μ.
2. On considère le test de région critique $W = \{|\overline{Z}| > k\}$. Calculer le seuil critique k pour $\alpha^* = 0.1$. Peut-on rejeter H_0 à ce niveau de signification ?
3. Calculer le degré de signification du test pour les données de l'exercice.

Exercice 10.3 *Non traité en TD.*

Chez un fabriquant de joints en caoutchouc, le département d'ingénierie de la qualité a mis en œuvre un plan d'échantillonnage pour vérifier le poids d'un joint d'étanchéité, poids qui est affecté par les variations d'écoulement du caoutchouc provenant de l'extrudeuse. La valeur cible du poids du joint est de 270 g.

On considère que le poids X est distribué normalement avec une espérance μ et un écart-type σ. Pour maîtriser le procédé, on prélève et pèse régulièrement $n = 5$ pièces de caoutchouc de l'extrudeuse.

1. Donner l'expression de la région critique W du test de l'hypothèse $H_0 : \mu = 270$ g contre $H_1 : \mu \neq 270$ g.
2. Lors d'un récent contrôle, on a obtenu, pour un échantillon de cinq pièces, les observations : 275.8, 270.1, 281.4, 279.2, 279.3. Calculer le degré de signification du test. Doit-on poursuivre ou arrêter la production ?
3. Indépendemment du paramètre de centrage, on souhaite que les variations de poids dues aux procédés de fabrication restent raisonnables, i.e. avec un écart-type $\sigma = 5$ g. Réaliser le test d'hypothèse :

$$H_0 : \sigma^2 = 25$$
$$H_1 : \sigma^2 > 25.$$

4. En admettant maintenant que $\sigma = 5$ g, reprendre la première question et calculer la probabilité β d'accepter l'hypothèse selon laquelle l'extrudeuse opère à 270 g alors qu'en réalité le procédé est centré à 275 g ?

Exercice 10.4 *Non traité en TD.*

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de variable parente X, de densité $f(x) = \frac{\theta^4}{6}\, x^3 e^{-\theta x} \mathbb{1}_{[0,+\infty[}(x)$ θ étant un paramètre strictement positif.

1. Donner l'expression de $\hat{\theta}$, estimateur du maximum de vraisemblance de θ.

2. Calculer l'information de Fisher $I_n(\theta)$ relative au paramètre θ. En déduire une fonction asymptotiquement pivotale pour θ que l'on exprimera en fonction de $\hat{\theta}$.
3. On considère le problème de test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ avec $\theta_1 > \theta_0$. Montrer que la région critique W du test le plus puissant pour ce problème au niveau α^* s'exprime en fonction de $\hat{\theta}$, puis donner une approximation de W en supposant n grand.
4. On considère maintenant le problème de test suivant $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Existe-t-il un test UPP pour ce problème ?
5. Calculer la statistique du rapport de vraisemblance λ, exprimée en fonction de $\hat{\theta}$, pour le problème de test de la question 4.
6. En utilisant la statistique $\ln \lambda$ et en supposant que n est grand, proposer une région critique pour le test de la question 4. Quelle décision prendra-t-on si $\theta_0 = 2$, $n = 50$, $\sum_i x_i = 115$ et $\alpha^* = 0.05$.

Exercice 10.5 *Non traité en TD.*

Un lion peut se trouver dans trois états : très actif (θ_1), peu actif (θ_2) et léthargique (θ_3). Le nombre X de personnes mangées en une nuit par un lion est une v.a. dont la loi dépend de l'état du lion $\theta \in \Theta = \{\theta_1, \theta_2, \theta_3\}$, selon le tableau suivant :

x	0	1	2	3	4
$p(x;\theta_1)$	.00	.05	.05	.80	.10
$p(x;\theta_2)$	.05	.05	.80	.10	.00
$p(x;\theta_3)$	.90	.08	.02	.00	.00

Ayant appris que x personnes ont été mangées par un lion au cours d'une nuit, on cherche à en déduire l'état du lion. Plus précisément, on considère les hypothèses $H_0 = \{\theta_3\}$ contre $H_1 = \{\theta_1, \theta_2\}$.

1. Donner dans un tableau les valeurs de la statistique $\Lambda(x)$ du rapport de vraisemblance, pour les différentes valeurs possibles de x.
2. Donner la loi de $\Lambda(X)$ sous l'hypothèse H_0.
3. En déduire la région critique du test du rapport de vraisemblance au niveau $\alpha^* = 2$ %.

Exercice 10.6 *Non traité en TD.*

Un cultivateur désire comparer les performances de deux types d'engrais. Pour cela, il observe les rendements $X_1, \ldots, X_n$ de n parcelles traitées par l'engrais I, et ceux $Y_1, \ldots, Y_m$ de m parcelles traitées par l'engrais II. On suppose que les rendements de chaque type suivent des lois normales d'espérances respectives μ_1 et μ_2, et de variance commune σ^2 connue.

1. On considère le problème de test :

$$\begin{array}{ll} H_0 : & \mu_1 = 2,\ \mu_2 = 4 \\ H_1 : & \mu_1 \neq 2 \text{ ou } \mu_2 \neq 4 \end{array}$$

   En utilisant le test du RV, proposer une solution en supposant n et m grands.

2. A.N. : $\overline{x} = 2.1$, $\overline{y} = 3.8$, $\sigma = 1$, $n = 40$, $m = 60$, $\alpha^* = 0.05$.

Chapitre 11

Tests d'homogénéité et de comparaison

Exercice 11.1 *À préparer intégralement.*

On cherche à comparer la durée de vie de deux types de pneus A et B. On dispose pour cela d'un échantillon de 41 durées de vie en milliers de km pour le type A et de 21 durées de vie pour le type B. Les résultats sont résumés dans le tableau suivant :

	n	$\sum x_i$	$\sum x_i^2$
A	41	1840	82996
B	21	828	32752

On admettra que les deux populations suivent les distributions normales $\mathcal{N}(\mu_A, \sigma_A^2)$ et $\mathcal{N}(\mu_B, \sigma_B^2)$ et dans tout cet exercice, on prendra comme niveau de signification des différents tests la valeur $\alpha^* = 0.05$.

1. Donner les estimations sans biais de μ_A, μ_B, σ_A^2 et σ_B^2.
2. Montrer que l'on peut admettre l'hypothèse d'égalité des variances des 2 populations.
3. En déduire une estimation sans biais de la variance commune σ^2.
4. Tester l'égalité des moyennes μ_A et μ_B.
5. On a reperé une erreur de saisie pour le type B, on a en réalité : $\sum x_i^2 = 32\,702$. Reprendre l'exercice avec cette nouvelle donnée.

Exercice 11.2

Afin de mesurer l'effet d'un traitement phytosanitaire sur le maïs on dispose de n parcelles, que l'on divise en deux, on traite la moitié sud de la parcelle, et on ne traite pas la moitié nord de la parcelle.

On mesure pour chaque moitié de parcelle le rendement, en t/Ha.

Parcelle	1	2	3	4	5	6	7	8	9
Sud (T : traité)	6.69	6.97	8.24	8.71	7.85	9.09	6.35	7.97	8.45
Nord (N : non traité)	6.46	6.68	7.82	8.22	7.34	9.03	6.56	7.74	8.38

1. Proposer un modèle ne tenant pas compte de la dépendance entre les deux moitiées de la même parcelle. Sous ce modèle, tenter de montrer que le traitement phytosanitaire à un effet.
2. Proposer un modèle plus réaliste, tenant compte de la dépendance entre les deux moitiées de la même parcelle. Sous ce modèle, tenter de montrer que le traitement phytosanitaire a un effet.
3. Quel est le défaut de ce plan d'expérience ?

Exercice 11.3

Une année, le taux national de réussite au baccalauréat dans une série donnée a été de 67 %.

1. Dans un centre d'examen A, il y a eu 216 reçus sur 300 candidats présentés. Les résultats de ce centre sont-ils conformes aux résultats nationaux ?
2. Dans un autre centre d'examen B, il y a eu 128 reçus sur 200 candidats. Les résultats des centres A et B sont-ils significativement différents ?

(Les tests seront réalisés au niveau de signification $\alpha^* = 5\%$).

Exercice 11.4 *Non traité en TD.*

On souhaite tester l'efficacité d'un additif destiné à diminuer la consommation d'essence. Pour ceci des essais sur 10 véhicules choisis au hasard ont été effectués et ont donné les résultats suivants :

véhicule	1	2	3	4	5	6	7	8	9	10
X (sans add.)	6.25	7.46	8.36	5.78	6.32	7.19	7.10	7.13	7.74	6.73
Y (avec add.)	6.10	7.30	8.20	5.90	6.30	7.20	6.80	7.00	7.60	6.70

1. Effectuer le test du signe un niveau de signification $\alpha^* = 0.05$. Attention : on ne fera pas l'approximation de la loi binomiale par la loi normale.
2. Faire un test de Wilcoxon signé.

Exercice 11.5 *Non traité en TD.*

Pour comparer la fiabilité de deux types T_1 et T_2 de machines à laver le linge, on choisit comme critère la durée de fonctionnement sans panne sur le programme « blanc », représentée comme une v.a. X pour T_1 et Y pour T_2. On suppose que X et Y suivent des lois normales de moyennes respectives μ_1 et μ_2 et de variances σ_1^2 et σ_2^2. Les observations, exprimées en jours de

fonctionnement, sur 10 machines de type T_1 et 13 machines de type T_2 ont donné les résultats suivants :

$$\sum_{i=1}^{10} x_i = 136.6 \quad \sum_{i=1}^{10} x_i^2 = 1923.2$$

$$\sum_{i=1}^{13} y_i = 187.0 \quad \sum_{i=1}^{13} y_i^2 = 2773.3$$

1. Tester l'égalité des variances, au niveau de signification $\alpha^* = 0.10$.
2. Les durées moyennes de fonctionnement sont-elles significativement différentes au seuil $\alpha^* = 0.05$?
3. Donner un intervalle de confiance à 95 % sur la différence $\mu_1 - \mu_2$ des durées de vie moyennes des deux types de machines.
4. En quoi cet intervalle de confiance est-il cohérent avec les résultats de la question 2 ?

Exercice 11.6 *Non traité en TD.*

On désire comparer les moyennes de 2 variables aléatoires supposées gaussiennes X et Y. On effectue pour cela 10 mesures indépendantes sur chaque variable aléatoire et on obtient les résultats suivants : $\sum_{i=1}^{10} x_i = 150$, $\sum_{i=1}^{10} x_i^2 = 2260$, $\sum_{i=1}^{10} y_i = 140$ et $\sum_{i=1}^{10} y_i^2 = 1974$. Tester l'hypothèse d'égalité des moyennes avec $\alpha^* = 0.05$.

Chapitre 12

Tests d'adéquation et d'indépendance

Exercice 12.1 *À préparer intégralement.*

La loi de Benford est une loi définie sur $\{1, 9\}$, s'exprimant comme :

$$\forall k \in \{1, \dots, 9\}, \qquad \mathbb{P}(X = k) = \log_{10}\left(1 + \frac{1}{k}\right).$$

Cette loi apparaît comme la loi sur le premier chiffre significatif de l'écriture décimale de phénomènes invariants à l'échelle.

On suppose par exemple, que les données financières d'une entreprise suivent très souvent la loi de Benford. Dans le cas contraire, cela pourrait correspondre à une anomalie explicable par entre autres, une tentative de manipulation des comptes afin de dissimuler une fraude.

L'adéquation a loi de Benford a été de nombreuses fois utilisée pour la détection automatique d'alertes sur les fraudes financières financières possible. Elle est utilisée par l'administration fiscale de certains pays.

Les données financières tous comptes confondus d'une entreprise ont été collectées. Elles sont les suivantes.

112.00	25448.93	19460.12	1710.36	21033.09
1157.33	4022.11	5037.10	893.57	2752.97
39244.80	3375.64	468.20	25445.81	1090.74
137.85	1911.97	118.27	1235.49	5018.95
2640.45	5108.76	8586.95	3355.68	21670.61
4210.36	3895.24	6154.03	5488.45	41605.07
2475.24	394.10	4814.88	4494.99	2655.40
13140.72	4996.00	4898.95	3750.07	1262.47
3148.84	15997.98	5017.32	323.77	4432.52
669.20	11380.35	4870.22	3442.10	1172.06

Sur la base de ce qui a été indiqué précédemment doit-on faire un contrôle approfondi de cette entreprise ?

On raisonnera avec un test asymptotique.

Exercice 12.2

Le tableau ci-dessous donne la répartition de la taille (en cm) de 2220 salariés français (hors salariés agricoles) et appartenant aux deux catégories socio-professionnelles (CSP) ouvriers et cadres supérieurs (source INSEE 1970).

taille \ CSP	ouvriers	cadres sup.	total
< 170	813	73	886
$[170, 175[$	636	123	759
≥ 175	451	124	575
total	1900	320	2220

Peut-on considérer, au niveau de signification de 5 %, que les deux facteurs taille et CSP sont indépendants dans la population totale de référence ?

Exercice 12.3

Cet exercice est en complément de l'exercice 11.2.

Nous rappellons le contexte.

Afin de mesurer l'effet d'un traitement phytosanitaire sur le maïs on dispose de n parcelles, que l'on divise en deux, on traite la moitié sud de la parcelle, et on ne traite pas la moitié nord de la parcelle.

On mesure pour chaque moitié de parcelle le rendement, en t/Ha.

Parcelle	1	2	3	4	5	6	7	8	9
Sud (T : traité)	6.69	6.97	8.24	8.71	7.85	9.09	6.35	7.97	8.45
Nord (N : non traité)	6.46	6.68	7.82	8.22	7.34	9.03	6.56	7.74	8.38

Dans cet exercice, le premier modèle supposait que N_i et T_i étaient indépendants. Le second modèle ne supposait pas cette indépendance.

Proposer et effectuer un test permetant justifier que le second modèle est à privilégier.

Exercice 12.4

On considère la réalisation suivante d'un échantillon iid de v.a. parente X :

$$9.1 \quad 7.4 \quad 17.2 \quad 10.7 \quad 15.5$$

1. Peut-on admettre au niveau $\alpha^* = 0.05$ que X suit une loi normale d'espérance 10 et de variance 4 ?
2. Peut-on admettre au niveau $\alpha^* = 0.05$ que X suit une loi normale ?

Exercice 12.5 *Non traité en TD.*

Une étude sur 320 familles ayant 5 enfants a donné les résultats suivants :

Nb de garçons	5	4	3	2	1	0	Total
Nb de filles	0	1	2	3	4	5	
Nb de familles	18	56	110	88	40	8	320

1. Sous l'hypothèse que la naissance d'un garçon et la naissance d'une fille sont des événements équiprobables, calculer les probabilités de chacun des six types de familles.
2. Peut-on admettre, au niveau de signification de 5% que les données obtenues sont compatibles avec cette hypothèse d'équiprobabilité ?
3. Calculer le degré de signification.

Exercice 12.6 *Non traité en TD.*

On examine 400 lots, chacun composé de 100 pièces. On compte dans chaque lot le nombre de pièces ne correspondant pas aux normes (pièces défectueuses). On construit le tableau suivant mettant en regard le nombre k de pièces défectueuses et le nombre n_k de lots présentant k pièces défectueuses.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
n_k	0	20	43	53	86	70	54	37	18	10	5	2	2

1. Donner, à partir de l'échantillon ci-dessus, la valeur estimée de la moyenne et de la variance de la variable aléatoire X, nombre de pièces défectueuses dans un lot.
2. Quelle est la distribution théorique qui vous paraît, compte tenu des résultats précédents, être la plus adaptée ? Tester cette hypothèse.

Exercice 12.7 *Non traité en TD.*

Dans l'exercice 1.2 du premier TD, nous avions conclu qu'il n'existait vraisemblablement pas de relation linéaire entre la quantité de poisson pêchée par un javelot et celle pêchée par un hameçon.

1. Construire un test d'hypothèses permettant de confirmer cette conclusion.
2. On souhaite maintenant savoir s'il existe une relation entre l'âge d'un pêcheur et son niveau. Construire un test pour répondre à cette interrogation.

Chapitre 13

Analyse de la variance

Exercice 13.1 *À préparer intégralement.*

Quinze veaux ont été répartis au hasard en trois lots, alimentés chacun de façon différente. Les gains de poids observés au cours d'une même période et exprimés en kg étant les suivants :

Lot 1	41.2	41.0	40.0	40.1	40.6
Lot 2	39.8	39.9	42.5	41.1	39.8
Lot 3	46.0	44.9	44.7	45.7	47.0

Le but de l'étude est de mettre en évidence une relation entre l'alimentation et la croissance des veaux. Les moyennes et les variance de ces 3 distributions sont $\bar{x}_1 = 40.58$, $\bar{x}_2 = 40.62$, $\bar{x}_3 = 45.66$, $s_1^{*\,2} = 0.282$, $s_2^{*\,2} = 1.407$ et $s_3^{*\,2} = 0.853$.

1. Tester la normalité des données correspondant au premier lot au niveau $\alpha^* = 0.05$. On supposera pour la suite que l'hypothèse de normalité peut être acceptée pour les deux autres lots.
2. Peut-on considérer que les variances des trois échantillons sont égales, au niveau de signification $\alpha^* = 0.05$?
3. Montrer que le type d'alimentation a un effet significatif sur la croissance des veaux. On prendra $\alpha^* = 0.05$.
4. Préciser pour quels types d'alimentation il existe des différences significatives.

Exercice 13.2

Les données suivantes représentent le nombre de problèmes arithmétiques simples (sur 85) résolus (de manière correcte ou non) en une heure par des sujets ayant reçu un médicament dépresseur, un stimulant et un placebo :

— dépresseur : 55, 0, 1 , 40
— stimulant : 75, 85, 51, 63

— placebo : 61, 54, 80, 47

Au vu des ces résultats, peut-on admettre que ces trois médicaments induisent des taux de performance différents ? (Faire une test de Kruskal-Wallis avec $\alpha^* = 0.10$.)

Exercice 13.3 *Non traité en TD.*

La composition en lipide de la myéline du système nerveux d'un échantillon provenant de quatre espèces différentes a donné les taux de cholestérol (en μ mol/mg lipide) suivants :

	Homme	boeuf	rat	grenouille
	0.70	0.58	0.68	0.64
	0.73	0.62	0.64	0.66
	0.68	0.65	0.71	0.69
	0.72	0.56	0.67	0.67
	0.69	0.59	0.65	0.65
$\overline{x}_k$	0.704	0.600	0.670	0.662
s_k^{*2}	$4.3\ 10^{-4}$	$12.5\ 10^{-4}$	$7.5\ 10^{-4}$	$3.7\ 10^{-4}$

On désire comparer, à l'aide d'une analyse de la variance, les taux de cholestérol entre les espèces.

1. Vérifier les conditions d'application de l'analyse de la variance, c'est-à-dire :
   (a) l'hypothèse de normalité de la distribution correspondant à l'espèce grenouille. On supposera pour la suite que les autres distributions vérifient la même propriété.
   (b) l'hypothèse concernant les variances des 4 populations.

   On prendra $\alpha^* = 5\%$.
2. Tester l'effet du facteur espèce en prenant un risque de première espèce de 5%.
3. Dans le cas d'un effet du facteur, préciser pour quelles espèces il existe des différences significatives.

Exercice 13.4 *Non traité en TD.*

Soient X_1 et X_2 deux variables aléatoires gaussiennes de même variance σ^2 et d'espérances μ_1 et μ_2. Les paramètres σ^2, μ_1 et μ_2 sont inconnus. On dispose d'un échantillon de taille n_1 de X_1 et d'un échantillon de taille n_2 de X_2. On note $\overline{X}_k$ et S_k^{*2} la moyenne et la variance empirique corrigée de chaque échantillon ($k = 1$ ou 2). On souhaite tester les hypothèses :

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 \neq \mu_2. \end{aligned}$$

Pour cela, on considère successivement deux méthodes.

1. Méthode 1 : *Test de Student.*

(a) Donner la région critique du test au niveau de signification α sous la forme :

$$f_1(\overline{X}_1, \overline{X}_2, S^*, n_1, n_2) > g_1(n_1, n_2, \alpha)$$

S^{*2} étant un estimateur sans biais de σ^2 dont on précisera l'expression en fonction de S_1^{*2} et S_2^{*2}.

(b) A. N. : Conclure avec $\overline{x}_1 = 5$, $\overline{x}_2 = 6$, $s_1^{*2} = 1$, $s_2^{*2} = 2$, $n_1 = n_2 = 10$, $\alpha = 0,05$.

2. Méthode 2 : *Analyse de la variance.*

(a) Préciser la relation entre MSW et S^*.

(b) Donner la région critique du test au niveau de signification α sous la forme :

$$f_2(\overline{X}_1, \overline{X}_2, S^*, n_1, n_2) > g_2(n_1, n_2, \alpha)$$

(c) A. N. : Conclure avec les mêmes valeurs numériques que précédemment.

3. On entreprend maintenant de comparer les deux méthodes précédentes.

(a) Soit T une variable aléatoire suivant une loi de Student à ν degrés de liberté. Montrer que T^2 suit une loi de Fisher à 1 et ν degrés de liberté.

(b) En déduire une relation entre $t_{\nu;1-\frac{\alpha}{2}}$ et $F_{1,\nu;1-\alpha}$. Comparer les fonctions g_1 et g_2 des questions 1 et 2.

(c) Montrer que :

$$MSB = \frac{n_1 n_2}{n_1 + n_2}(\overline{X}_1 - \overline{X}_2)^2.$$

(d) En déduire une relation entre les fonctions f_1 et f_2 des questions précédentes.

(e) Que peut-on en conclure concernant les procédures de Student et d'analyse de la variance dans le cas de deux populations ?

Exercice 13.5 *Non traité en TD.*

On désire tester l'influence de la turbine sur le débit d'air en sortie d'un turbocompresseur. Pour cela, on analyse le fonctionnement de 5 turbines montées sur le même "corps" de turbo et avec un même compresseur. Chaque ensemble ainsi formé ne diffère des autres que par la turbine. Ces 5 ensembles passent sur le même banc et on fait pour chacun 5 mesures de débit à la même vitesse de rotation (1750 Hz).

Turbine	1	2	3	4	5
	10.7651	10.8167	10.6752	10.7932	10.5463
	10.7716	10.8255	10.6279	10.7456	10.6305
	10.7246	10.8081	10.6667	10.7707	10.6036
	10.5752	10.7059	10.6072	10.7468	10.5878
	10.6960	10.6583	10.5253	10.7503	10.5887
$\bar{x}_i$	10.7065	10.7630	10.6205	10.7613	10.5914
s_i^{*2}	0.00633	0.00578	0.00357	0.00040	0.00090

1. Tester si les 5 échantillons ainsi obtenus proviennent d'une population gaussienne ($\alpha = 0.05$).
2. Tester l'égalité des variances des 5 populations ($\alpha = 0.05$).
3. Tester l'effet du facteur turbine ($\alpha = 0.05$).
4. Calculer la puissance du test.
5. Comparer les turbines 2 à 2 ($\alpha = 0.05$).

Annexe A

Indices

Exercice 2.7

2. La loi d'une somme de loi de Poisson indépendantes est une loi de Poisson dont le paramètre est la somme des paramètres.

4. Utiliser le TCL pour approcher le nombre total de passages.

Exercice 2.8

1. On a $\sum_{i=1}^{n+1} X_i = \sum_{i=1}^{n} X_i + X_{n+1}$

3. Utiliser $a^2 - b^2 = (a-b)(a+b)$ n fois dès le début.

5. Exprimer S_n sous forme d'une somme télescopique et utiliser la question (3)

6a. Utiliser la définition (0.3).

6d. Les Z_k sont gaussiens. Il suffit de montrer que les covariances des Z_i sont nulles.

Exercice 3.5

Traiter d'abord les ampoules A et B indépendamment. Utiliser une intégration par partie pour le calcul de l'espérance et la variance d'une loi exponentielle.

Exercice 3.6

1. Montrer que si $X \sim \mathscr{P}(\lambda)$ et $Y \sim \mathscr{P}(\mu)$ alors $X+Y \sim \mathscr{P}(\lambda+\mu)$. Pour cela, calculer la fonction de probabilité de $X+Y$, on devra utiliser la formule du binôme de Newton.

3. Pour calculer l'espérance $\mathbb{E}(X)$, il faut reconnaître le développement en série entière de la fonction exponentielle. Pour calculer $\mathbb{E}(X^2)$, il faut faire intervenir la dérivée du développement en série entière de λe^{λ}.

Exercice 3.8

Utiliser une loi de Poisson d'espérance 12.

Exercice 4.5

2. Utiliser la fonction $g(x) = x/(x-1)$ et montrer qu'elle est strictement convexe.
3. Calculer le fonction de densité g en dérivant G'.
4. Calculer $\mathbb{E}(Y)$. Utiliser à nouveau Jensen avec la fonction $x \mapsto 1/x$ sur $\mathbb{R}^{+*}$.

Exercice 4.6

3. Annuler le biais donne une relation entre a et b. Minimiser la variance en annulant la dérivée.

Exercice 4.7

1a. On connait $\mathrm{Var}(X_i)$ d'après l'exercice 4.3. Pour le risque, utiliser le fait que le biais est nul.

2c. Utiliser $Z_i = \frac{3}{\theta^2} X_i^2$ sachant que X_i suit une loi uniforme.

2d. Écrire K_n en fonction des $Z_i - 1$ et utiliser un TCL et la question précédente.

2e. Prendre $W = K_n$.

2f. Ramener à un calcul des moments de K_n.

Exercice 4.8

1. Le calcul de l'espérance suffit.
3. Exprimer l'espérance de T et transformer $\mathbb{E}(\overline{X}^2)$ avec la formule de la variance.

Exercice 7.6

4. Faire un test sur $b = 0$ en utilisant la statistique $\frac{\hat{b}}{\hat{\sigma}/\sqrt{nS_{XY}}} \overset{H_0}{\sim} \mathcal{T}_{n-2}$ et un test sur la régression en utilisant $\frac{M_{reg}}{M_{res}} \overset{H_0}{\sim} \mathcal{F}_{1,n-2}$.
5. Faire un test de Shapiro–Wilk pour la normalité suivi d'un test sur l'espérance.

Exercice 8.4

Utiliser le théorème de Neyman-Pearson.

Exercice 8.5

1. Utiliser le théorème de Neyman-Pearson.
2. Utiliser la définition du degré de signification.

Exercice 9.4

1. Montrer l'existence d'un test UPP dans le premier mais pas dans le second.

Exercice 9.5

3. Faire un test sur $b = 0$ en utilisant la statistique $\frac{\hat{b}}{\hat{\sigma}/\sqrt{nS_{XY}}} \overset{H_0}{\sim} \mathcal{T}_{n-2}$ et un test sur la régression en utilisant $\frac{M_{reg}}{M_{res}} \overset{H_0}{\sim} \mathcal{F}_{1,n-2}$.

Exercice 9.6

4. Faire un test sur $b = 0$ en utilisant la statistique $\frac{\hat{b}}{\hat{\sigma}/\sqrt{nS_{XY}}} \overset{H_0}{\sim} \mathcal{T}_{n-2}$ et un test sur la régression en utilisant $\frac{M_{reg}}{M_{res}} \overset{H_0}{\sim} \mathcal{F}_{1,n-2}$.
5. Faire un test de Shapiro–Wilk pour la normalité suivi d'un test sur l'espérance.

Exercice 10.3

1. On ignore σ. Il faut utiliser la statistique de Student.
2. Utiliser $\hat{\alpha} = 2\left(1 - F_{\mathcal{T}_{n-1}}(|t_{\text{obs}}|)\right)$ et discuter selon α^*.
3. Utiliser un test unilatéral sur la variance.
4. Il s'agit de calculer un risque de deuxième espèce pour une hypothèse $h_1 \in H_1$ précise.

Exercice 10.5

1. Comme l'échantillon est de longueur 1, la vraisemblance se lit directement dans le tableau.
2. Utiliser le tableau précédent.
3. Écrire la région critique et calculer directement le seuil en fonction de α^*.

Exercice 10.6

1. Écrire la vraisemblance de μ_1 et μ_2 en fonction des échantillons $x_1, \dots, x_n$ et $y_1, \dots, y_m$.

Exercice 11.4

1. Compter le nombre de différences négatives et le modéliser par une binomiale.

Exercice 11.5

1. Utiliser un test de Fisher.
2. Utiliser un test de Student.
3. Utiliser la fonction pivotale de la question précédente.

Exercice 11.6

Utiliser un test de Fisher suivi d'un test de Student.

Exercice 12.5

1. Nombre de cas favorables divisé par nombre de cas total.
2. Faire un test du χ^2 d'adéquation.

Exercice 12.6

1. Utiliser les formules pour un échantillon groupé par classe.
2. Un comptage suggère a priori une loi de Poisson. Faire un test du χ^2 d'adéquation.

Exercice 12.7

1. Utiliser une test de Pearson.
2. Utiliser un test de Spearman.

Exercice 13.3

1. Test de normalité + Bartlett
2. Utiliser une analyse de la variance.
3. Utiliser la procédure LSD de Fisher.

Exercice 13.4

1a. Utiliser un test de Student bilatéral.

Annexe B

Corrigés courts

Exercice 1.1

1. Les variables sont toutes les deux de types quantitatives.

2. $\widehat{f}_{0.1} = 1$, $\widehat{f}_{0.25} = 1$, $\widehat{f}_{0.5} = 2$, $\widehat{f}_{0.75} = 4$, $\widehat{f}_{0.9} = 5$.
$\overline{x} = 2.6$, $s = 1.51$.

Exercice 1.2

6. $\overline{x}_J = 13.86$, $s_J^{*2} = 318.45$, $s_J^2 = 297.22$, $s_J^* = 17.85$, $s_J = 17.24$, $\widehat{f}_{J,0.25} = 2.0$, $\widehat{f}_{J,0.5} = 8.1$, $\widehat{f}_{J,0.75} = 16.2$, $\widehat{f}_{J,0.75} - \widehat{f}_{J,0.25} = 14.2$
$\overline{x}_H = 9.65$, $s_H^{*2} = 67.88$, $s_H^2 = 63.36$, $s_H^* = 8.24$, $s_H = 7.96$, $\widehat{f}_{H,0.25} = 2.5$, $\widehat{f}_{H,0.5} = 8.0$, $\widehat{f}_{H,0.75} = 14.9$, $\widehat{f}_{H,0.75} - \widehat{f}_{H,0.25} = 12.4$

8. $r = 0.07$

Exercice 1.3

1. Il s'agit d'une variable qualitative

Exercice 1.4

3. Moy.=4.47,Var.=0.12, E-T=0.35, q1=4.20, Med=4.5, q3=4.70, IQR=0.50

5. Moy.=8.22,Var.=173, E-T=13.3, q1=4.20, Med=4.5, q3=4.70, IQR=0.50

Exercice 1.5

1. Diagramme en boîte donc quantitatif. Discret ou continu suivant justification.

2. $\overline{x} = 4.825$, $\overline{x}_1 = 4.7778$, $s^2 = 3.3069$, $s^{*2} = 3.4809$.

3. $M = 5$, $q_1 = 3$, $q_3 = 6.5$, $W = 6.5$, $H = 3.5$.

Exercice 1.6

Exercice 1.7

3. $\overline{x} = 3080.9$, $s = 724.65$.

Exercice 2.1

1. $a = \frac{2}{b^2}$ et $c = d$
2. 1. $\mathbb{E}[X] = \frac{2b}{3}$ et $\mathrm{Var}[X] = \frac{b^2}{18}$
   2. Pas d'espérance ni de variance

Exercice 2.2

2. $\mu = 163$ cg, $\sigma = 9.68$ cg.
3. $\mathbb{P}(X < 0)$ semble négligeable.
4. Environ 19 200 personnes.

Exercice 2.3

1. U suit la même loi que X.
2. $\mathbb{P}(X = -1 \text{ et } U = -1) = \frac{1}{8} \neq \frac{1}{16} = \mathbb{P}(X = -1)\mathbb{P}(U = -1)$. X et U ne sont pas indépendantes.
3. $\mathbb{P}(U = 1 \mid X = 1) = \frac{1}{2} \neq \frac{1}{4} = \mathbb{P}(U = 1)$. X et U ne sont pas indépendantes.

Exercice 2.4

Exercice 2.5

1. $G(x) = F(x)^n$
2. $g(x) = nf(x)F(x)^{n-1}$
3. $\mathbb{E}(Y) = \frac{n}{(n+1)}\theta$ et $\mathrm{Var}(Y) = \frac{n}{(n+2)(n+1)^2}\theta^2$.

Exercice 2.6

1. Loi de proba ssi $\theta \in [0, 1]$.
2. $\mathbb{E}(X) = \frac{1}{2} - \theta$, $\mathbb{E}(X^2) = \frac{1}{2}$, $\mathrm{Var}(X) = \frac{1}{4} + \theta - \theta^2$
3. Oui.
4. $N_1 \sim \mathscr{B}\left(n, \frac{1-\theta}{2}\right)$, $\mathbb{E}(N_1) = n\frac{1-\theta}{2}$, $\mathrm{Var}(N_1) = n\frac{1-\theta}{2}\frac{1+\theta}{2}$
5. 0.31744

Exercice 2.7

1. Découper la minute en n intervalles, faire apparaître $X = \sum_{i=1}^{n} \mathscr{B}(\cdots)$ et faire tendre n vers l'infini.
2. Loi de Poisson de paramètre 24.9.
4. $\mathbb{P}(X > 1500) \approx 0.4332263648$

Exercice 2.8

1. $\overline{X}_{n+1} = \frac{1}{n+1}\left(n\overline{X}_n + X_{n+1}\right)$

Exercice 2.9

Exercice 3.1

2. $S_n \overset{\text{app.}}{\sim} \mathcal{N}(n\mu, n\sigma^2)$
3. 0.0764.

Exercice 3.2

Exercice 3.3

3. $l > 41423.42$

Exercice 3.4

1. $\mathbb{E}\left(\overline{X}_n\right) = \mu$, $\mathrm{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$.
3. $\frac{1}{3}\frac{\sigma^2}{n}$, $\frac{3}{5}\frac{\sigma^2}{n}$, $\frac{\sigma^2}{n}$, $\frac{5}{3}\frac{\sigma^2}{n}$, $3\frac{\sigma^2}{n}$

Exercice 3.5

0.7486

Exercice 3.6

1. $Y_n \sim \mathscr{P}(n)$
2. 0.5591.
3. 0.5.
4. 0.5438, 0.5279, 0.5199 et 0.5160

Exercice 3.7

1.

$$\begin{aligned}\mathbb{P}(S \geq 60) = 1 - \mathbb{P}(S \leq 59) &= 1 - \Phi\left(\frac{59 + 0.5 - np}{\sqrt{np(1-p)}}\right) \\ &= 1 - \Phi(9.5/5) = 1 - 0.9713 = 0.0287.\end{aligned}$$

2.

$$\begin{aligned}\mathbb{P}(N \geq 10) = 1 - \mathbb{P}(N \leq 9) &= 1 - \Phi\left(\frac{9 + 0.5 - mq}{\sqrt{mq(1-q)}}\right) \\ &= 1 - \Phi(-1.3) = \Phi(1.3) = 0.9.\end{aligned}$$

Exercice 3.8

$1 - 0.1867 = 0.8133$

Exercice 4.1

1. $\mathbb{E}(X) = \theta/4$, $\theta = 4\mathbb{E}(X)$
2. $\hat{\theta}_m = 4\overline{X}$
3. $\mathbb{E}(\hat{\theta}_m) = \theta$, $\mathrm{Var}(\hat{\theta}_m) = (4 - 3\theta^2)/3n \underset{n\to+\infty}{\longrightarrow} 0$

Exercice 4.2

1. $\mathbb{E}(X) = p_1 + 2p_2$, $\mathbb{E}(X^2) = p_1 + 4p_2$, $\mathrm{Var}(X) = p_1 + 4p_2 - (p_1 + 2p_2)^2$
2. $\hat{p}_1 = 2\overline{X} - \widehat{m}_2$ et $\hat{p}_2 = (\widehat{m}_2 - \overline{X})/2$
3. $\mathbb{E}(\hat{p}_1) = p_1$ et $\mathbb{E}(\hat{p}_2) = p_2$

Exercice 4.3

1. $\hat{\theta}_1 = \sqrt{3}S$, estimateur biaisé.
2. $\mathbb{E}(Y) = \theta/2$, $\mathrm{Var}(Y) = \theta^2/12$
3. $\hat{\theta}_2 = 2\overline{Y}$, sans biais.

Exercice 4.4

2. $\hat{\theta} = \left(\frac{1}{2n}\sum_{i=1}^{n} X_i^{\frac{1}{2}}\right)^2$

Exercice 4.5

1. $\hat{\theta}_1 = \frac{\overline{X}}{\overline{X}-1}$
3. $F(x) = (1 - x^{-\theta})\mathbb{1}_{[1,+\infty[}$, $G(y) = (1 - e^{-y\theta})\mathbb{1}_{[0,+\infty[}$ et $g(y) = G'(y) = \theta e^{-y\theta}$
4. $\hat{\theta}_2 = 1/\overline{Y}$, estimateur biaisé.

Exercice 4.6

3.

$$\hat{\theta} = \frac{\frac{\sigma_Y^2}{n_Y}\overline{X} + \frac{\sigma_X^2}{n_X}\overline{Y}}{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

L'estimateur obtenu est une combinaison linéaire des estimateurs la pondération étant proportionnelle à l'inverse de leur variance.

Exercice 4.7

1a. $\mathbb{E}\left(\hat{\theta}_1 - \theta\right)^2 = \frac{1}{3n}\theta^2$

2a. $\frac{\hat{\theta}_1}{\theta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i}$

2b. $\frac{\hat{\theta}_1}{\theta} = \sqrt{1 + K_n}$

2f. $\mathbb{E}\left(\hat{\theta}_1 - \theta\right)^2 \approx \frac{1}{5n}\theta^2$

Exercice 4.8

1. $\hat{p} = \overline{X}$
2. $T = \overline{X}\left(1 - \overline{X}\right)$
3. $\mathbb{E}(T) = \frac{n-1}{n}p(1-p)$
4. $\frac{n}{n-1}T$.

Exercice 5.1

2. $L(\lambda; k_1, \dots, k_n) = \frac{e^{-n\lambda}\lambda^{n\overline{k}}}{\prod_{i=1}^{n} k_i!}$
3. $\ln L(\lambda; k_1, \dots, k_n) = -n\lambda + n\overline{k}\ln(\lambda) - \ln\left(\prod_{i=1}^{n} k_i!\right)$, $\hat{\lambda}_{MV} = \overline{K}$.
4. $I_n(\lambda) = \frac{n}{\lambda}$
5. $\mathcal{N}(\lambda, \lambda/n)$

Exercice 5.2

2. $\hat{\theta}_{MV} = -\frac{n}{\sum_i \log(1-|X_i|)} - 1$
4. $I_n(\theta) = \frac{n}{(\theta+1)^2}$
5. $\hat{\theta}_{MV} \overset{\text{app.}}{\sim} \mathcal{N}\left(\theta, \frac{(\theta+1)^2}{n}\right)$

Exercice 5.3

1. $L(\theta; x_1, \dots, x_n) = \mathbb{1}_{[s_n - 1, i_n]}(\theta)$
2. $\hat{\theta}_n(\alpha)$ représente tout le segment $[S_n - 1, I_n]$.
3. $f_{S_n}(x) = n(x-\theta)^{n-1}\mathbb{1}_{[\theta,\theta+1]}(x)$
$f_{I_n}(x) = n(1-x+\theta)^{n-1}\mathbb{1}_{[\theta,\theta+1]}(x)$.
4. $\mathbb{E}(I_n) = \theta + \frac{1}{n+1}$ et $\mathbb{E}(S_n) = \theta + \frac{n}{n+1}$.
5. $\mathbb{E}(\hat{\theta}_n) = \theta + \frac{1-2\alpha}{n+1} \Rightarrow \alpha^* = 1/2$.

Exercice 5.4

1. $\ln L(\theta; x_1, \dots, x_n) = n_0 \ln\frac{1}{2} + n_1 \ln\frac{\theta}{2} + n_2 \ln\frac{1-\theta}{2}$
2. $\hat{\theta}_3 = \frac{N_1}{N_1+N_2}$

3. $I_n(\theta) = \frac{n}{2\theta(1-\theta)}$

4. $\hat{\theta}_3 \overset{\text{app.}}{\sim} \mathcal{N}\left(\theta, \frac{2\theta(1-\theta)}{n}\right)$

Exercice 5.5

2. $\hat{\theta} = \min_{i=1,\dots,n} X_i = X_{(1)}$

Exercice 5.6

2. $Y \sim \mathcal{B}\left(n, \frac{1}{4} + \frac{p}{2}\right)$

3. $\hat{p} = 2\left(\frac{Y}{n} - \frac{1}{4}\right)$

4. $\frac{4}{n}\left(\frac{1}{4} + \frac{p}{2}\right)\left(\frac{3}{4} - \frac{p}{2}\right)$

6. Il faut interroger 4 fois plus d'étudiants.

Exercice 5.7

2. $\forall i \in \{1, \dots, n\} \qquad Y_i \overset{\text{ind}}{\sim} \mathcal{P}(\lambda x_i)$

4. $\hat{\lambda}_{\text{MV}} = \frac{\sum_i Y_i}{\sum_i x_i}$

6. $\hat{\lambda} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$

8. $\operatorname{Var} \hat{\lambda}_{\text{MV}} \leq \operatorname{Var} \hat{\lambda}$. Les deux estimateurs étant sans biais, l'estimateur $\hat{\lambda}_{\text{MV}}$ a donc un plus petit risque que $\hat{\lambda}$.

Exercice 5.8

2. $\hat{\mu} = \frac{(2-\rho)X_1 + \left(2-2\rho-2\rho^2\right)X_2 + (2-\rho)X_3}{(6-4\rho-2\rho^2)}$

4. $\mathbb{E}\,\hat{\mu} = \mu$, $\operatorname{Var} \hat{\mu} = \frac{6+6\rho-3\rho^2-7\rho^3}{2(1-\rho)(3+\rho)^2}$

Exercice 6.1

1.

$$\forall i \in \{1, \dots, n\}, \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$$

2. $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\widehat{\sigma^2} = S^{*2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$

3. $I_1 = \left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right]$

4. $I_2 = \left[\overline{X} - \frac{S^*}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}, \overline{X} + \frac{S^*}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]$

5. $I_3 = \left[\frac{(n-1)S^{*2}}{\chi^2_{n-1,1-\alpha}}, +\infty\right[$

6. $\overline{x} = 22.99$, $s^{*2} = 0.0221$, $i_2 = [22.8836, 23.0964]$ et $i_3 = [0.0118, +\infty]$

7. $n \geq 62$

Exercice 6.2

- À 95% : $[220.30, 247.59]$,
- À 99% : $[216.19, 252.31]$.

Exercice 6.3

1.
$$\forall i \in \{1, \dots, 7\}, \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\xi, \sigma^2\right)$$

2.
$$\left[\overline{X} - t_{n-1,1-\frac{10\alpha}{11}} \frac{S^*}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{11}} \frac{S^*}{\sqrt{n}}\right]$$

3. $[12.2621, 12.5547]$

Exercice 6.4

1.
$$\forall i \in \{1, \dots, n\} \qquad X_i \overset{\text{iid}}{\sim} \mathcal{P}(\lambda)$$

2. $\frac{\overline{X}-\lambda}{\sqrt{\lambda/n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

4. $\left[\overline{X} - u_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}; \overline{X} + u_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}\right]$.

5. $[2.81; 4.49]$.

6. $\overline{X} + \frac{1}{2n}u^2_{1-\alpha/2}\left(1 \pm \sqrt{1 + \frac{4n\overline{X}}{u^2_{1-\alpha/2}}}\right)$, A.N. : $[2.90; 4.59]$.

Exercice 6.5

1. $\hat{\theta} = \frac{2}{\overline{X}}$.

2. $I_n(\theta) = \frac{2n}{\theta^2}$, $\sqrt{2n}\left(\frac{\hat{\theta}}{\theta} - 1\right) \sim \mathcal{N}(0,1)$, $IC = \left[\frac{\hat{\theta}}{1+u_{1-\frac{\alpha^*}{2}}/\sqrt{2n}}, \frac{\hat{\theta}}{1-u_{1-\frac{\alpha^*}{2}}/\sqrt{2n}}\right]$

3. A.N. : $[0.44, 0.58]$.

Exercice 6.6

1. $\forall i \in \{1, \dots, n\} \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$

Exercice 6.7

Exercice 6.8

2b. $\hat{c} = \overline{X} \qquad \widehat{\Delta} = \sqrt{\frac{\widehat{m}_2 - \overline{X}^2}{2}}$

2d. $IC = \left[\hat{c} \pm \frac{u_{1-\alpha/2}}{\sqrt{n}}\Delta\sqrt{2}\right]$.

3a. Non.

3e. $IC = \left[\hat{c}_{\text{MV}} \pm \frac{u_{1-\alpha/2}}{\sqrt{n}}\Delta\right]$.

Exercice 7.1

1.

$$\forall i \in \{1, \dots, n\}, \qquad U_i \overset{\text{ind}}{\sim} \mathcal{N}\left(rx_i, \sigma^2\right)$$

2. $\hat{r} = \frac{\sum_{i=1}^n x_i U_i}{\sum_{i=1}^n x_i^2}$ et $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (U_i - \hat{r}x_i)^2$.

3. $\mathbb{E}(\hat{r}) = r$, $\text{Var}(\hat{r}) = \sigma^2 / \sum_{i=1}^n x_i^2$ et $\hat{r} \sim \mathcal{N}\left(r, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$.

4. $\hat{r} \pm \frac{u_{1-\alpha/2}\sigma}{\sqrt{\sum_{i=1}^n x_i^2}}$

5. $\hat{r}_{\text{réalisation}} = 17.9$, $ic = [17.84, 17.97]$.

6. $\widehat{U_0}_{\text{réalisation}} = 895.3$, $ic_{\mathbb{E}(U_0)} = [892.2, 898.4]$, $ic_{U_0} = [891.7; 899.0]$.

Exercice 7.2

1b.

$$\forall i, \qquad C_i = a + bv_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

1c. Aucune abscisse extrème.

1d. non.

2a.

$$\forall i, \qquad C_i = a + bv_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

2c. non.

3. $IC_{95\%} = [1.37, 460.84]$ L/100km.

Exercice 7.3

1. Modèle : $Y_i = a + bx_i + \varepsilon_i$ avec ε_i v.a. iid et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ où x_i représente le revenu et Y_i représente l'épargne.
$\hat{b} = 0.307$, $\hat{a} = 0.00502$ et $\hat{\sigma}^2 = 0.222$.

2. $R^2 = 0.995$

3. Pour $x_0 = 20$: $\widehat{Y_0} = 6.145$ et $ic = [5.02, 7.27]$.
Pour $x_0 = 50$: $\widehat{Y_0} = 15.355$ et $ic = [14.05, 16.67]$.

Exercice 7.4

1. $\hat{b} = 3.22$, $\hat{a} = 79.59$, $\hat{\sigma}^2_{MV} = S_{res} = 15.66$ et $\hat{\sigma}^2 = \frac{n}{n-2}S_{res} = 19.58$.

2. Intervalle bilatéral : $\hat{a} \pm t_{n-2;1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_x^2}}$, a.n. $[64.94, 94.25]$.

Intervalle unilatéral : $a > \hat{a} - t_{n-2;1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_x^2}} = 67.78$.

3. Intervalle bilatéral $\hat{b} \pm t_{n-2;1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{ns_x^2}}$, a.n. $[1.95, 4.48]$.

Intervalle unilatéral : $b < \hat{b} + t_{n-2;1-\alpha}\frac{\hat{\sigma}}{\sqrt{ns_x^2}} = 4.23$.

4. $\frac{\hat{b}}{\hat{\sigma}/\sqrt{ns_x^2}} = 5.87 > t_{8;0.99} = 2.90 : H_0$ rejetée.

Exercice 7.5

1b. $\hat{\rho} = \frac{\sum_i v_i M_i}{\sum_i v_i^2}$

2b.

$$\forall i \in \{1, \dots, n\}, \qquad N_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\rho\sqrt{v_i}, c\right)$$

Exercice 7.6

2. $Y_i = a + bx_i + \varepsilon_i$ pour tout i où les ε_i sont iid et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. $\overline{x} = 442.23$, $\overline{y} = 928.67$, $S_{XX} = 27132.22$, $S_{YY} = 55499.89$, $S_{YY} = 55499.89$, $S_{XY} = 35970.28$, $\hat{b} = 1.325740$, $\hat{a} = 342.25$, $S_{reg} = 47687.243$, $S_{res} = 7812.646$, $\hat{\sigma}^2 = 11718.975$

3. $R^2 = 0.859$

4. $W = \left\{|\hat{b}| > \frac{\hat{\sigma}}{\sqrt{nS_{xx}}}t_{10;0.995} = 0.745\right\}$, on rejette la nullité de b.
   $W = \left\{\frac{M_{reg}}{M_{res}} > F_{1,4;0.95} = 7.71\right\}$, $\frac{M_{reg}}{M_{res}} = 24.42$, la régression est significative.

5. Les résidus sont : $-98, -31, -31, 63, 164, -67$.
   Shapiro–Wilk : $W = 0.89601$, p–value $= 0.3509$. On peut supposer la normalité.
   t–test : $t = 0$. On garde H_0.

Exercice 7.7

Exercice 8.1

1. $W = \{t > \sigma_0^2 \chi^2_{n,1-\alpha^*}\}$ A.N. $t > 18,3$

2. $1 - \beta = 0,5$

Exercice 8.2

Exercice 8.3

1. $\frac{L_1}{L_0} = \exp\left(\frac{n}{2\sigma^2}(2\mu_1\overline{x} - \mu_1^2)\right)$

2. $W = \{\overline{x} > k\}$ avec $k = \frac{2}{\sqrt{n}}u_{1-\alpha^*}$

3. $\pi = 1 - \Phi\left(u_{1-\alpha^*} - \frac{\sqrt{n}}{2}\mu_1\right)$

6. $n_0 = 44$
7. On rejette H_0 pour $\alpha^* = 0.1$ et $\alpha^* = 0.05$, mais on ne rejette pas H_0 pour $\alpha^* = 0.01$.
8. $\hat{\alpha} = 0.0228$

Exercice 8.4

$W = \left\{\overline{x} < \frac{1}{p_0} - \sqrt{\frac{1-p_0}{np_0^2}} u_{1-\alpha^*}\right\}$

Exercice 8.5

Exercice 9.1

1. $\hat{\theta} = \frac{\overline{X}}{2}$
2. $\mathbb{E}(X) = 2\theta$, $\mathrm{Var}(X) = 2\theta^2$, $\frac{\hat{\theta}-\theta}{\theta/\sqrt{2n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$.
3. $W = \{\hat{\theta} > k\}$ et $k \approx \theta_0 \left(1 + \frac{u_{1-\alpha^*}}{\sqrt{2n}}\right)$.
4. Pas de test UPP.
5. $\Lambda(X_1, \dots, X_n) = \left(\frac{\hat{\theta}}{\theta_0}\right)^{2n} \exp\left(2n\left(1 - \frac{\hat{\theta}}{\theta_0}\right)\right)$.
6. $W = \{-2\ln\lambda \geq \chi^2_{1;0.95} = 3.84\}$ et $-2\ln\lambda = 25.67$: rejet de H_0.

Exercice 9.2

1. $\hat{p}_{\mathrm{MV}} = X/n$
2. $\Lambda(X) = \left(\frac{np_0}{X}\right)^X \left(\frac{n(1-p_0)}{n-X}\right)^{n-X}$
3. $W = \left\{-2\left[x\ln\left(\frac{np_0}{x}\right) + (n-x)\ln\left(\frac{n(1-p_0)}{n-x}\right)\right] > \chi^2_{1,1-\alpha^*}\right\}$
4. $-2\ln\lambda = 10.0167$, $\chi^2_{1,0.95} = 3.84$, rejet de H_0

Exercice 9.3

1a. $W = \left\{\bar{x} < \mu_0 - \frac{\sigma}{\sqrt{n}} u_{1-\alpha^*}\right\}$

1b. On conserve H_0

1c. $1 - \beta = 0.93$

2a. $K \sim \mathscr{B}(n, p)$ où $p = \mathbb{P}(X < \mu_0 - 1) = \Phi(\frac{\mu_0 - 1 - \mu}{\sigma})$

2b. $H_0 : p = p_0$, $H_1 : p > p_0$ avec $p_0 = \Phi(\frac{-1}{\sigma})$

2c. $W = \{k > 3\}$, on conserve H_0

2d. $\pi = 0.8281$, ce test est moins puissant que le précédent.

Exercice 9.4

Exercice 9.5

1. $\overline{x} = 1.642$, $\overline{y} = 0.508$, $S_{XX} = 399.55$, $S_{YY} = 37.74$, $S_{XY} = 122.50$, $\hat{\beta} = 0.307$, $\hat{\alpha} = 0.00502$, $S_{reg} = 37.557$, $S_{res} = 0.185$, $\hat{\sigma}^2 = 0.222$.

2. $R^2 = 0.995$

3. $W = \left\{|\hat{b}| > \frac{\hat{\sigma}}{\sqrt{nS_{xx}}} t_{10;0.995} = 0.0152\right\}$, on rejette la nullité de b.
   $W = \left\{\frac{M_{reg}}{M_{res}} > F_{1,10;0.95} = 4.96\right\}$, $\frac{M_{reg}}{M_{res}} = 2031$, la régression est significative.

Exercice 9.6

2. $Y_i = a + bx_i + \varepsilon_i$ pour tout i où les ε_i sont iid et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. $\overline{x} = 442.23$, $\overline{y} = 928.67$, $S_{XX} = 27132.22$, $S_{YY} = 55499.89$, $S_{YY} = 55499.89$, $S_{XY} = 35970.28$, $\hat{b} = 1.325740$, $\hat{a} = 342.25$, $S_{reg} = 47687.243$, $S_{res} = 7812.646$, $\hat{\sigma}^2 = 11718.975$

3. $R^2 = 0.859$

4. $W = \left\{|\hat{b}| > \frac{\hat{\sigma}}{\sqrt{nS_{xx}}} t_{10;0.995} = 0.745\right\}$, on rejette la nullité de b.
   $W = \left\{\frac{M_{reg}}{M_{res}} > F_{1,4;0.95} = 7.71\right\}$, $\frac{M_{reg}}{M_{res}} = 24.42$, la régression est significative.

5. Les résidus sont : -98, -31, -31, 63, 164, -67.
   Shapiro–Wilk : $W = 0.89601$, p–value $= 0.3509$. On peut supposer la normalité.
   t–test : $t = 0$. On garde H_0.

Exercice 9.7

1. Test sur la moyenne d'une v.a. gaussienne de variance connue : $W = \left\{\frac{|\hat{r} - r_0|}{\sqrt{1/\sum_{i=1}^n x_i^2}} > u_{1-\alpha^*/2}\right\}$.

Exercice 10.1

1. $d_0 = 20$ cm.

2. $Z = \text{Card}\{X_i | X_i \leq 20,\ i \in \{1, \dots, n\}\}$, $Z \sim \mathcal{B}(n, \theta)$ avec $\theta = \mathbb{P}(X \leq 20)$. $H_0 : \theta = 0.1$ contre $H_1 : \theta > 0.1$.

3. On observe $z = 5$, $\hat{\alpha} = \mathbb{P}_{\theta=0.1}(Z \geq 5) = 1 - \mathbb{P}_{\theta=0.1}(X \leq 4) \approx 0.1841$, l'affirmation de l'industriel est raisonnable.

Exercice 10.2

1. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.

2. $k = 0.52$. On rejette H_0.

3. $\hat{\alpha} = 0.058$.

Exercice 10.3

1. $W = \left\{\frac{|\overline{x} - 270|}{s^*/\sqrt{n}} > t_{n-1,1-\frac{\alpha^*}{2}}\right\}$

2. On a $\hat{\alpha} \in [0.02; 0.04]$. On devrait arrêter la production pour $\alpha^* = 0.05$, on accepterait de poursuivre pour $\alpha^* = 0.01$.

3. On accepte H_0.

4. $W = \left\{ \frac{|\overline{x}-270|}{5/\sqrt{5}} > u_{1-\frac{\alpha^*}{2}} \right\}$, $\beta = \mathbb{P}_{\mu_1}\left(\overline{W}\right) = 0.39$

Exercice 10.4

Exercice 10.5

1. Pour $x = 0, 1, 2, 3, 4$ on a resp. $\Lambda(x) = 1, 1, 0.025, 0.00, 0.00$

2. $\mathbb{P}_{H_0}(\Lambda(X) = 1) = 0.98$, $\mathbb{P}_{H_0}(\Lambda(X) = 0.025) = 0.02$ et $\mathbb{P}_{H_0}(\Lambda(X) = 0.00) = 0$.

3. $W = \{\Lambda(x) < 1\}$.

Exercice 10.6

1. $W = \{-2\log\lambda > \chi^2_{2,1-\alpha^*}\}$ avec $-2\log\lambda = \frac{n(\overline{x}-2)^2+m(\overline{y}-4)^2}{\sigma^2}$

2. A.N. : $-2\log\lambda = 2.8$ et $\chi^2_{2;0.95} = 5.99$: on ne rejette pas H_0.

Exercice 11.1

1. $\overline{x}_A \simeq 44.88$, $\overline{x}_B \simeq 39.43$, $s^{*2}_A \simeq 10.33$, $s^{*2}_B \simeq 5.14$

2. $W = \left\{ \frac{s^{*2}_A}{s^{*2}_B} < 0.483 \right\} \cup \left\{ \frac{s^{*2}_A}{s^{*2}_B} > 2.29 \right\}$, $\frac{s^{*2}_A}{s^{*2}_B} \simeq 2$

3. $s^{*2} \simeq 8.76$

4. $W = \left\{ \frac{|\overline{x}_A-\overline{x}_B|}{s^*\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} > 2 \right\}$, $\frac{|\overline{x}_A-\overline{x}_B|}{s^*\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} \simeq 6.86$: rejet de l'hypothèse d'égalité.

5. On ne peut plus admettre l'égalité des variances, on doit faire un test de Welch. $W = \left\{ |d|/\sqrt{s^{*2}_D} > t_{\nu,1-\alpha^*/2} \right\}$. On rejette l'hypothèse d'égalité des espérances.

Exercice 11.2

1. Fisher : $\hat{\alpha} = .88$. Student : $\hat{\alpha} = .6$.

2. $t_{\text{obs}} = 2.9919$. $\hat{\alpha} = .017$.

3. L'exposition nord/sud pourrait expliquer la différence mise en évidence.

Exercice 11.3

1. $W = \{|\hat{p}_A-p_0|\,(p_0(1-p_0)/n_A)^{-1/2} > u_{1-\alpha^*/2}\}$ avec $\hat{p}_A = X_A/n_A$. A.N. : $|\hat{p}-p_0|\,(p_0(1-p_0)/n_A)^{-1/2} = 1.84$ et $u_{0.975} = 1.96$: pas de rejet de H_0.

2. $W = \{|\hat{p}_A - \hat{p}_B|\,(\hat{p}(1-\hat{p})\,(1/n_A + 1/n_B))^{-1/2} > u_{1-\alpha^*/2}\}$ avec $\hat{p}_A = X_A/n_A$, $\hat{p}_B = X_B/n_B$, $\hat{p} = (X_A + X_B)/(n_A + n_B)$. A.N. : $1.89 < 1.96$: pas de différence significative.

Exercice 11.4

1. $W = \{z \leq c\}$ avec $Z \overset{H_0}{\sim} \mathscr{B}(n, 0.5)$. Pour $\alpha^* = 0.05$, on trouve $c = 2$ donc on rejette H_0.
2. $W = \{w^+ \leq B\}$ avec $w^+ = 1 + 4 = 5$ et $B = 10$, on rejette H_0.

Exercice 11.5

1. Test de Fisher : $s_X^{*2}/s_Y^{*2} < F_{n_1-1;n_2-1;\alpha/2}$ ou $> F_{n_1-1;n_2-1;1-\alpha^*/2}$, a.n. $s_X^{*2}/s_Y^{*2} < 0.33$ ou > 2.80, $s_X^{*2} = 6.36$, $s_Y^{*2} = 6.95$, $s_X^{*2}/sY*2 = 0.92$: égalité des variances.
2. Test de Student : $\frac{|\overline{x}-\overline{y}|}{s^*\sqrt{1/n_1+1/n_2}} > t_{n_1+n_2-2;1-\alpha^*/2}$, a.n. $t_{21;0.975} = 2.080$, $s^{*2} = 6.70$, $\frac{|\overline{x}-\overline{y}|}{S^*\sqrt{1/n_1+1/n_2}} = 0.66$: égalité des moyennes.
3. $\overline{X} - \overline{Y} \pm S^* t_{n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, a.n. $[-2.98, 1.54]$.

Exercice 11.6

AN : $A = t_{18;0.975} = 2.101$, $s^{*2} = 1.33$, $\frac{|\overline{y}-\overline{y}|}{s^*\sqrt{1/n_X+1/n_Y}} = \frac{|15-14|}{\sqrt{1.33}\sqrt{2/10}} = 1.93$.

On accepte donc l'égalité des moyennes.

Exercice 12.1

$d^2_{\text{obs}} = 12.97$, $\hat{\alpha} \in [0.1, 0.2]$, conclusion.

Exercice 12.2

$W = \{d^2 > 5.99\}$ et $d^2 = 53$: rejet de l'hypothèse d'indépendance.

Exercice 12.3

Dessiner le graphique pour comprendre le lien linéaire. Le test sur le coefficient de corrélation donne : $r_{TN} = .970$, $\hat{\alpha} < 2 \cdot 10^{-4}$. Le lien est très significatif.

Exercice 12.4

1. Test de K-S : $W = \{d_n > 0.563\}$ et $d_n = 0.40$: pas de rejet de H_0.
2. Test de Shapiro–Wilk : $\mathscr{W}(X) = 0.92$ et $w_{5,0.05} = 0.775$: pas de rejet de H_0.

Exercice 12.5

1. $1/32, 5/32, 10/32, 10/32, 5/32, 1/32$
2. $D^2 = 11.96 > F^{-1}_{\chi^2}(0.95) = 11.1$: rejet de l'hypothèse d'équiprobabilité.
3. $\hat{\alpha} \approx 0.0351$.

Exercice 12.6

1. Calcul de la moyenne et de la variance de l'échantillon

$$\overline{x} = 1/n \sum_{k=1}^{K} n_k k = 1872/400 = 4.68$$

$$s^{*2} = 1/(n-1) \sum_{k=1}^{K} n_k (k - \overline{x})^2 = 4.47$$

2. $d^2 = \sum_k \frac{(n_k - np_k)^2}{np_k} = 4.8$ et $W = \left\{D^2 >= 15.5\right\}$. On accepte l'hypothèse H_0.

Exercice 12.7

1. Test de Pearson : $W = \left\{|t| > t_{n-2,1-\alpha/2}\right\}$ avec $t = \frac{r\sqrt{n-2}}{1-r^2}$. A.N. $t = 0.25, t_{13,0.975} \approx 2.16, \hat{\alpha} = 0.81$. On accepte $H_0 \ : \ \rho = 0$ (hypothèse d'indépendance).

2. Test de Spearman : $W = \left\{|t| > u_{1-\alpha/2}\right\}$ avec A.N. $r_S = 0.33$, $t = 1.27$, $\hat{\alpha} = 0.22$. On accepte $H_0 \ : \ \rho = 0$ (hypothèse d'indépendance).

Exercice 13.1

1. Test de Shapiro-Wilk : $\mathscr{W}(X) = 0.91$ et $w_{5,0.05} = 0.775$: pas de rejet de H_0.

2. Test de Bartlett : $W = \{b > 5.99\}$ et $b = 2.35$: égalité des variances.

3. Test de l'analyse de la variance : $W = \{f > 3.885\}$ et $f = 50.20$: espérances significativement différentes.

4. LSD de Fisher : $t_{1,2} = 0.07$, $t_{1,3} = 8.73$ et $t_{2,3} = 8.66$ à comparer à 2.18 : différences significatives entre 1 et 3 et entre 2 et 3.

Exercice 13.2

$W = \{h > \chi^2_{2;0.9}\}$, $h = 5.35$ et $\chi^2_{2;0.9} = 4.61$: effet significatif.

Exercice 13.3

1. SW : $\hat{\alpha} > .5$. On ne rejette pas l'hypothèse de normalité.
Test de Bartlett : $W \ : \ B = (N-K)\ln(MSW) - \sum_{k=1}^{K}(n_k - 1)\ln(S_k^{*2}) > \chi^2_{K-1,1-\alpha}$, $MSW = 7\times10^{-4}$, $b = 1.90$ et $\chi^2_{3,0.95} = 7.81$. On ne rejette pas l'hypothèse d'égalité des variances.

2. $F = \frac{MSB}{MSW} = 13.419$ et $F_{K-1,N-K,1-\alpha} = F_{3,16,0.95} = 3.24$: effet espèce significatif.

3. bœuf, homme : $\underline{6.22}$, rat, homme : 2.03, grenouille, homme : $\underline{2.51}$, rat, bœuf : $\underline{4.18}$, grenouille, bœuf : $\underline{3.71}$, grenouille, rat : 0.48.

Exercice 13.4

Exercice 13.5

1. Test de normalité de Shaprio-Wilk.

2. Test de l'égalité des variances : test de Cochran

$$g = \frac{\max(s_k^{*2})}{\sum_k s_k^{*2}} = \frac{0.00634}{0.0171} = 0.372$$

Comme $g_{n,k,0.05} = g_{5,5,0.05} = 0.54$, on accepte l'hypothèse d'égalité des variances.

3. Test de l'effet du facteur turbine

source de variation	degrés de liberté	somme des carrés	moyenne des carrés
Inter-modalité	$k-1=4$	$SSB = 0.126$	$MSB = 0.0315$
Intra-modalité	$n-k=20$	$SSW = 0.0682$	$MSW = 0.00341$
Total	$N-1$	SST	

La région critique est définie par

$$\frac{MSB}{MSW} > F_{k-1,N-k,1-\alpha^*}.$$

Ici on a $\frac{MSB}{MSW} = 9.24$ et $F_{4,20,0.95} = 2.87$. Nous sommes donc dans la région critique et il y a un effet turbine.

4. Calcul de la puissance

$$1-\beta = \mathbb{P}(\frac{MSB}{MSW} > 2.87|H_1)$$

On en déduit

$$F_{k-1,N-k,\lambda,\beta} = 2.87$$

avec

$$\lambda = \frac{1}{\sigma^2}\sum_k n_k \alpha_k^2.$$

On estime λ :

$$\hat{\lambda} = (k-1)(\frac{N-k-2}{N-k}\frac{MSB}{MSW} - 1) = 4(\frac{25-5-2}{25-5}9.24 - 1) = 29.26$$

En utilisant les abaques pour $\nu_1 = 4$, $\nu_2 = 20$, $\alpha^* = 0.05$ et $\Phi = \sqrt{\frac{29.26}{5}} = 2.42$, on obtient $1-\beta = 0.98$.

5. Comparaison des différentes moyennes : Intervalles de Tukey

$$T = \frac{1}{2\sqrt{n}}\sum_{i=1}^{K}|c_i| q_{K,N-K,\alpha^*} = \frac{1}{\sqrt{n}} q_{K,N-K,\alpha^*} = \frac{1}{\sqrt{5}} q_{5,20,0.05} = \frac{4.23}{\sqrt{5}} = 1.89$$

$$T\sqrt{MSW} = 1.89\sqrt{0.034} = 0.11$$

Les estimations des contrastes $\hat{c}_{ij} = \overline{x}_i - \overline{x}_j$ sont les suivantes :

	1	2	3	4
2	−0.0565			
3	0.0860	0.1425		
4	−0.0548	0.0017	−0.1408	
5	0.1151	0.1716	0.0291	0.1699

On considère qu'il y a une différence significative entre deux turbines lorsque l'intervalle de confiance centré en $\hat{c}_{ij}$ et de largeur $T\sqrt{MSW} = 0.11$ ne contient pas 0, c'est-à-dire lorsque cette largeur est inférieur à $|\hat{c}_{ij}|$. On obtient alors les résultats suivants :

	1	2	3	4
2	identique			
3	identique	différente		
4	identique	identique	différente	
5	différente	différente	identique	différente