WORKBOOK - EXERCICES TD

Responsable : Carole Durand-Desprez

Site Web : Chamilo/STA401

Objectifs *:*

Chaque fiche de TD correspond à une séance. Chacune commence par un exercice entièrement corrigé pour modèle.

Avant chaque td, il est vivement conseillé d'avoir revu le cours, fait un résumé et d'avoir complété votre formulaire qui se trouve en fin de ce worbook.

Il est rapidement impossible de travailler sans calculatrice en Statistique. Vous devrez absolument connaître l'utilisation du mode 'STAT' de votre calculatrice depuis le lycée. Si vous avez des lacunes, vous pouvez consulter le fichier "aide-calculatrice" sur le site. Ce fichier contient aussi les nouvelles commandes pour mettre en oeuvre les méthodes de cette année.

Votre réussite passe par un travail régulier.

Bon travail à tous !

PLANNING

FICHE - 1 - Révisions

Exercice 1.1 *(Révisions de Statistique descriptive)*

Le nombre de pannes a été enregistrées par semaine dans un centre pendant une période de 100 semaines.

Pannes	0	1	2	3	4	5	6	7
Semaines	42	35	11	5	3	2	1	1

1. Quelle est la variable étudiée ? De quel type est-elle ? Quelles sont ses modalités ?
2. Calculer les fréquences empiriques des modalités. En faire une représentation graphique.
3. Calculer la moyenne, la variance et l'écart type empirique de l'échantillon.
4. Calculer les fréquences cumulées. Quelle est la fréquence empirique de l'intervalle [2 ;5] ?
5. Calculer la médiane. Faire la représentation graphique de la fonction de répartition. Retrouver la médiane à l'aide de ce graphique.

Solution :

1. La variable étudiée est X : "le nombre de pannes par semaine". C'est une variable quantitative (numérique) et discrète (elle ne peut prendre que les valeurs entières, et plusieurs indivudus prennent la même modalité). Ses modalités sont : {0,1,2,3,4,5,6,7}. L'effectif (n_i) de chaque modalité est donné dans le tableau.

2. Pour calculer la fréquence empirique de chaque modalité il suffit de diviser l'effectif correspondant par la taille de l'échantillon total ($f_i = \frac{n_i}{n}$)

Fréquences	0,42	0,35	0,11	0,05	0,03	0,02	0,01	0,01

Le diagramme des fréquences est donné ci-dessous :

3. La moyenne emprique de cet échantillon est : $\overline{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum n_i c_i = \frac{1}{100}(0+35+22+15+12+10+6+7) = 1,07$.

La variance emprique de cet échantillon est : $s_x^2 = \frac{1}{n}\left(\sum x_i^2\right) - \overline{x}^2 = \frac{1}{n}\left(\sum n_i c_i^2\right) - \overline{x}^2 = \frac{1}{100}(0 + 35 + 44 + 45 + 48 + 50 + 36 + 49) - 1,07^2 = 1,9251$. *L'écart type est* $s_x = \sqrt{1,9251} = 1,3874797$ *Retrouver ces résultats en utilisant le mode 'STAT' de votre calculatrice.*

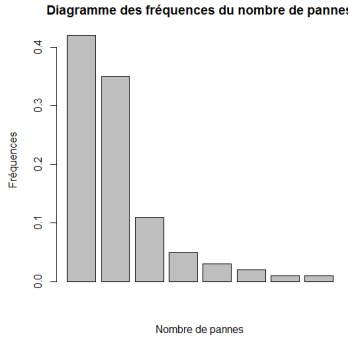
4. Les fréquences cumulées empiriques représentent les valeurs de la fonction de répartition :

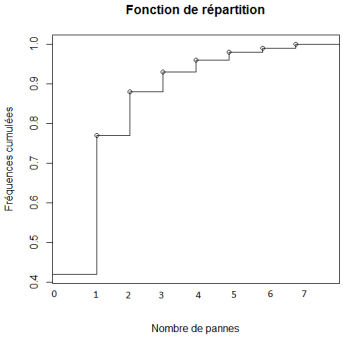
Fréquences cumulées	0,42	0,77	0,88	0,93	0,96	0,98	0,99	1

La fréquence de [2 ;5] est la somme des fréquence de chacune des modalités : 0,11+0,05+0,03+0,02 = 0,21. C'est aussi la différence des valeurs de la fonction de répartition empirique : F(5)-F(1) = 0,98-0,77=0,21. Donc, pour 21% des semaines, il y a entre 2 et 5 pannes.

5. La médiane correspond à la modalité pour laquelle exactement 50% des effectifs sont inférieurs. Dans les fréquences cumulées on voit que 0,5 correspond à la modalité 1. La médiane du nombre de pannes est de 1.

La fonction de répartition est donnée ci-dessous. On voit aussi sur le graphique que 0,5 correspond à la modalité 1. La médiane est de 1 panne.





Exercice 1.2

Le poids d'un fichier transformé en pdf varie selon les différentes versions d'un même logiciel utilisé. On a relevé ces poids sur un échantillon de 12 (mesurés en ko).

poids	101	104	102	99	101	98	97	104	100	96	103	101

1. Quelle est la variable étudiée ? De quel type est-elle ?
2. Calculer la médiane.
3. Calculer la moyenne, la variance et l'écart type empirique de l'échantillon. (S'entraîner sur la calculatrice)
4. Regrouper ces données en trois classes de même longueur (3ko). Faire un histogramme.
5. Calculer l'échantillon centré et réduit associé. Vérifier la moyenne et la variance de cet échantillon.

Exercice 1.3

On considère deux sites web A et B concurrents. On compte le nombre de clics sur chacun des sites. Le nombre de clics moyen sur A est de 19,6 par heure avec un écart type de 1,8. Pour le site B, le nombre de clics moyen est de 20,1 par heure avec un écart type de 2,1. Sur quel site le nombre de clics est le plus important ? Sur quel site le nombre de clics est le plus régulier ?

Exercice 1.4

On compte le nombre d'appels sur une hotline pour l'aide en ligne d'un nouveau logiciel de gestion :

Nb jours	2	3	4	5	6	7	8	9
Nb appels	3	3	5	38	39	75	26	1

Lors de la dernière mise à jour du logiciel, cette ligne avait enregistré 115 appels en tout, une moyenne de 6 jours et un écart type de 1,2083. Que dire depuis cette mise à jour ? Comment peut-on comparer ?

Avec les méthodes des chapitres suivants, la comparaison de telles données sera beaucoup plus "efficace".

Exercice 1.5

On dispose de 2 ordianteurs A et B. Des simulations de tailles de fichiers pour des calculs de tris en mémoire ont donné les valeurs suivantes sur les fichiers temporaires (par exemple, il y a eu 6 fois des fichiers de taille 3 Meg avec l'ordinateur A) :

Taille (Meg)	3	4	8	11
A	6	2	5	9
B	8	3	11	7

Calculer les moyennes et variances des tailles des fichiers avec les deux ordinateurs. Déduire quel est l'ordinateur pour lequel la taille des fichiers est la moins importante.

Avec les méthodes des chapitres suivants, la comparaison de telles données sera beaucoup plus "efficace".

Exercice 1.6 *(Toutes les questions sont indépendantes)*

1. Dans une classe de 30 élèves, la moyenne des notes des vingt filles est de 12 et la moyenne des dix garçons est de 8. Quelle est la moyenne générale de la classe ?

2. Déterminer la médiane, les quartiles et l'écart interquartile de la série de données suivantes : 11, 12, 12, 13, 15, 16, 16, 17, 17, 18, 19, 20, 22, 23.

3. On s'intéresse à la variable "Nature du lieu d'habitation". Le tableau ci-dessous résume les informations prises sur un échantillon d'individus :

Nature	Centre ville	Banlieue	Village	Cité	Autre
Effectif	87	30	32	30	10

   Quel est le type de cette variable ? Quelles sont ses modalités ? Calculer les fréquences empiriques. Faire un diagramme représentatif. Peut-on calculer la médiane, la moyenne et la variance ?

FICHE - 2

Exercice 2.1 *(Révisions de Probabilité)*

PARTIE A :

On sait que dans une certaine boîte mail, 5% des courriers sont des spams. On met au point un test permettant de détecter si un courrier est un spam ou pas (détection de certains mots dans le mail). Lorsqu'un mail est réellement un spam la probabilité pour que le test soit positif (spam détecté) est de 0,9. En revanche, lorsque le mail n'est pas un spam, la probabilité que le test soit négatif (spam non détecté) est de 0,82. On note les évènements : S "le mail est un spam" et T "le test est positif". [Vous pouvez vous aider d'un arbre pondéré.]

1. Calculer les probabilités $P[T \cap S]$, $P[T \cap \overline{S}]$. En déduire $P[T]$.
2. Calculer les probabilités dans les deux cas suivants :
   (a) Probabilité qu'un mail soit un spam lorsqu'un test s'est révélé positif.
   (b) Probabilité qu'un mail ne soit pas un spam alors que le test était négatif.
3. Les évènements S et T sont-ils indépendants ? Justifier.

PARTIE B :

Dans votre boîte mail, la probabilité d'apparition d'un spam est de 8%. Hier vous avez reçu 10 courriers que vous n'avez pas encore ouverts.

1. Soit X la variable aléatoire égale au nombre de spams reçus hier. Quelle est la loi de X ?
2. Quelle est la probabilité qu'il n'y ait aucun spam ?
3. La probabilité qu'il y ait au moins 7 spams est-elle inférieure à 0,1% ?

Solution :

PARTIE A :

1. D'après l'énoncé : $P[S] = 0,05 \quad P[T \mid S] = 0,9 \quad P[\overline{T} \mid \overline{S}] = 0,82 \quad$ *et* $\quad P[T \mid \overline{S}] = 1 - P[\overline{T} \mid \overline{S}] = 0,018$

$P[T \cap S] = P[T \mid S]P[S] = 0,9 * 0,05 = 0,045 \quad P[T \cap \overline{S}] = P[T \mid \overline{S}]P[\overline{S}] = 0,18 * 0,95 = 0,171$. *On déduit donc :* $P[T] = P[T \cap S] + P[T \cap \overline{S}] = 0,216$.

2. (a) $P[S \mid T] = \dfrac{P[T \cap S]}{P[T]} = 0,045/0,216 \approx 0,2083$

(b) $P[\overline{S} \mid \overline{T}] = \dfrac{P[\overline{T} \cap \overline{S}]}{P[\overline{T}]}$. *De plus, on a :* $(\overline{T} \cap \overline{S}) \cup (T \cap \overline{S}) = \overline{S}$, *donc* $P[(\overline{T} \cap \overline{S})] + P[(T \cap \overline{S})] = P[\overline{S}]$

D'où : $P[\overline{T} \cap \overline{S}] = P[\overline{S}] - P[T \cap \overline{S}] = 0,779$. *On déduit alors :* $P[\overline{S} \mid \overline{T}] = 0,779/(1 - 0,216) = 0,9936$

3. $P[T \cap S] \neq P[T]P[S]$ *donc les évènements ne sont pas indépendants. (ou bien :* $P[T \mid S] \neq P[T]$ *donc non indépendants.)*

PARTIE B :

1. Pour chaque mail X_i *suit une loi Bernoulli de paramètre* $p = 0,08$ *; tous indépendants.* X *suit une loi Binomiale de paramètres* $n = 10$ *et* $p = 0,08$. *(Somme de 8 variables indépendantes de Bernoulli(p))*

2. $P[X = 0] = \begin{pmatrix} 10 \\ 0 \end{pmatrix} 0,08^0 * 0,92^{10} = 0,4344$

3 $P[X \geq 7] = P[X = 7] + P[X = 8] + P[X = 9] + P[X = 10] = 1 - P[X \leq 6] = 2,02.10^{-6}$. *La probabilité est donc bien inférieure à 0,1%.*

Exercice 2.2

Dans un lot de coques de téléphones portables, il y a 5% de coques non conformes. On contrôle la fabrication de ces coques, mais le mécanisme de contrôle est aléatoire. Si la coque est conforme, elle est acceptée avec une probabilité égale à 0,96. Si la coque est non conforme, elle est rejetée avec probabilité 0,98. On choisit au hasard une coque que l'on contrôle.

1. Quelle est la probabilité que cette coque soit rejetée ?
2. Quelle est la probabilité que cette coque soit conforme, sachant qu'elle a été rejetée ?
3. Quelle est la probabilité que cette coque soit non conforme sachant qu'elle est acceptée ?
4. Quel est le pourcentage d'erreur dans le contrôle (soit la coque est conforme et rejetée, soit la coque est non conforme et acceptée) ?

Exercice 2.3

1. Vous êtes invité chez une personne dont vous savez qu'elle a exactement 2 enfants. Un garçon vient vous ouvrir la porte. Quelle est la probabilité que le deuxième enfant soit un garçon ?
2. Vous êtes invité chez une personne dont vous savez qu'elle a exactement 2 enfants. Un garçon vient vous ouvrir la porte. Vous entendez un bébé pleurer. Quelle est la probabilité que le deuxième enfant soit un garçon ?

Exercice 2.4 *Tous les exercices sont indépendants*

1. Un jury est formé de 6 personnes prises au hasard dans un groupe composé de 5 hommes et 4 femmes ; Soit X la variable aléatoire 'nombre de femmes dans le jury'. Quelle est la loi de X ? Calculer P(X = 3).

2. Il y a 1% des trèfles qui ont 4 feuilles. On cueille 100 trèfles dans un pré ; X est la variable aléatoire 'nombre de trèfles à 4 feuilles'. Calculer P(X = 3).

3. Dans une certaine compagnie d'assurance, il arrive en moyenne 2 déclarations d'accidents par heure dans la France entière. X est la variable aléatoire 'nombre de déclarations d'accidents arrivées dans la journée'. On considèrera que la journée compte 8 heures travaillées. Calculer P(X = 3).

4. La belote se joue avec un jeu de 32 cartes et quatre joueurs reçoivent chacun huit cartes ; X est le 'nombre d'as reçu' par un joueur donné. Quelle est la loi de X ? Calculer P(X = 3).

5. Au loto, vous devez cocher 6 numéros sur une grille qui en compte 49 ; X est le 'nombre de bons numéros sur votre grille'. Quelle est la loi de X ? Calculer P(X = 3).

6. *Les mésaventures du Chevalier de Méré* : Il est connu pour avoir échangé avec Blaise Pascal une correspondance régulière sur les jeux de hasard, alors que la théorie des probabilités était en pleine naissance au milieu du XVIIème siècle. Il aurait perdu des sommes d'argent importantes en pariant aux dés, en passant du premier jeu ci-dessous au second jeu.

   Premier jeu : le parieur lance quatre fois de suite un dé, à six faces. Il gagne s'il obtient au moins une fois un six.

   Second jeu : le parieur lance 24 fois de suite une paire de dés. Il gagne s'il obtient au moins une fois un double six.

   Modéliser les deux jeux. Expliquer alors pourquoi le Chevalier de Méré était forcément perdant.

Exercice 2.5

Un serveur/concentrateur dessert 1000 postes via 50 lignes à haut débit. Aux heures de pointe, chaque poste est occupé en moyenne pendant 2,5 secondes par minute. Quelle est la probabilité de saturation du réseau pendant une minute en heures de pointe ?

[Indications : Modéliser X la variable ' nombre de postes occupés ', calculer P($X \geq 51$)]

Exercice 2.6

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

1. On suppose que les 6 témoins choisissent au hasard le coupable. Quelles sont vos chances :
   (a) de n'être jamais désigné ?
   (b) d'être désigné exactement une fois ?
   (c) d'être désigné deux fois ou plus ?
2. On suppose que deux des 6 témoins vous ont désigné comme coupable. En utilisant la question 1 (c), pensez-vous que le juge pourra attribuer cela au hasard ?
3. On suppose maintenant que quatre des 6 témoins vous ont désigné. Le juge peut-il attribuer cela au hasard ?

Exercice 2.7

PARTIE A

Les mots de passe sur une certain site sont très vulnérables. On suppose qu'un hacker qui chercherait à trouver un mot de passe sur ce site aurait une chance sur 10 de le découvrir. Quatre hacker s'attaquent indépendamment au même mot de passe d'un compte.

1. Quelle est la loi de probabilité du nombre de hacker qui ont trouvé le mot de passe ?
2. Quelle est la probabilité que ce compte soit piraté ? Quelle est la probabilité que les 4 pirates trouvent le mot de passe ?

PARTIE B

Un site fait des recherches pour sécuriser ses mots de passe. En moyenne deux comptes sont piratés par semaine.

1. Quelle est la loi de probabilité du nombre de comptes piratés dans un mois ?
2. Quelle est la probabilité pour que 10 comptes soient piratés en un mois ? Il y a 5 ouvertures de comptes par semaine, quelle est la probabilité que le nombre de comptes piratés par mois soit supérieur au nombre de comptes ouverts par mois ?

Exercice 2.8 *Utilisations de l'inégalité de Bienaymé Thebychev*

1. Soit X le nombre aléatoire d'avions arrivant en une heure sur un aéroport. On suppose que E[X]=16 et V[X]=16.
   a) À l'aide de l'inégalité de Bienaymé Tchebychev, donner une minoration de P $(10 < X < 22)$
   b) Comparer votre résultat avec celui que l'on obtiendrait si on supposait que X avait pour loi $\mathcal{P}(16)$

2. En utilisant l'inégalité de Bienaymé Tchebychev, combien de fois doit-on lancer une pièce de monnaie pour que l'on ait une probabilité supérieure à 0,9 que la fréquence du nombre de « pile » obtenu soit compris entre 0,4 et 0,6 ?

Exercice 2.9 *Quelques proriétés de cours sur la loi de Poisson*

1. Montrer qu'une loi Binomiale de paramètre n et p peut être approchée par une loi de Poisson de paramètre $\lambda = np$, pour n assez grand (convergence en loi).

   [Indication : Partir de la probabilité P(X=k) d'une loi Binomiale, développer et faire apparaître le terme $\lambda = np$, trouver des équivalents en $+\infty$, puis passer à la limite]

2. Montrer que la somme de deux lois de Poisson indépendantes de paramètres λ_1 et λ_2 est une loi de Poisson de paramètre $\lambda_1+\lambda_2$

   [Indication : Partir de la probabilité P(X+Y = k) puis développer, utiliser ensuite la formule du binôme

   $(\lambda_1 + \lambda_2)^k = k\,! \sum_{i=0}^{k} \left(\frac{\lambda_1^i \lambda_2^{k-i}}{i!\,(k-i)!} \right)$, *puis conclure]*

FICHE - 3

Exercice 3.1

1. Montrer que la fonction de densité de probabilité $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ vérifie : $E(X) = 0$ et $V(X) = 1$
2. En notant F_1 la fonction de répartition de la loi $\mathcal{N}(\mu; \sigma^2)$ et F_2 la fonction de répartition de la loi $\mathcal{N}(0; 1)$, montrer que $F_1(x) = F_2(\frac{x-\mu}{\sigma})$. Il suffit donc de connaître la fonction de répartition de la loi centrée réduite pour déduire les autres (ce qui justifie que seule la table de la loi $\mathcal{N}(0; 1)$ suffit).
3. Application : On étudie X la v. a. représentant le temps de décharge d'une certaine batterie de téléphone portable en cours de fabrication. On sait que X suit une loi $\mathcal{N}(7; 4)$.

   a) Quelle est la loi de $\frac{X-7}{2}$?

   b) Quelle est la probabilité qu'une de ces batteries tienne plus de 10 h ?

   c) Le fabricant affirme que plus des 2 tiers des batteries tiennent entre 5 et 9 h. Est-ce vrai ?

   d) On améliore ce temps de décharge selon le modèle $5X - 2$, quelle est la nouvelle loi ? Reprendre la question b) pour ce nouveau modèle.

Solution :

1. $E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x\, e^{-\frac{x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{x^2}{2}}\right]_{-\infty}^{+\infty} = 0$

   $V(X) = \int_{-\infty}^{+\infty} x^2\, f(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2\, e^{-\frac{x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \left[-x\, e^{-\frac{x^2}{2}}\right]_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}\, dx$ *[IPP]*

   $= 0 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}\, dx = 1$ *(c'est une proba sur $\mathbb{R}$!)*

2. *Soit X une v. a. de loi* $\mathcal{N}(\mu; \sigma^2)$,

   $F_1(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{z^2}{2}}\, dz$ *[changement de variable* $z = \frac{t-\mu}{\sigma}$*]*

   $= P(Y \leq y = \frac{x-\mu}{\sigma}) = F_2(\frac{x-\mu}{\sigma})$ *en notant F_2 la fonction de répartition de la loi* $\mathcal{N}(0; 1)$

3. *a.* $\frac{X-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0; 1)$, *donc* $Z = \frac{X-7}{2} \rightsquigarrow \mathcal{N}(0; 1)$

   b. $P[X > 10] = P[\frac{X-7}{2} > \frac{10-7}{2}] = P[Z > 1,5] = 1 - P[Z < 1,5] = 1 - 0,9332 = 0,0668$

   c. $P[5 < X < 9] = P[\frac{5-7}{2} < \frac{X-7}{2} < \frac{9-7}{2}] = P[-1 < Z < 1] = P[Z < 1] - P[Z < -1] = 2P[Z < 1] - 1$

   $= 2 * 0,8413 - 1 = 68,26\%$ *le fabricant dit vrai.*

   d. $X \rightsquigarrow \mathcal{N}(7; 4) \Rightarrow 5X \rightsquigarrow \mathcal{N}(5 * 7; 5^2 * 4) \Rightarrow, Y = 5X - 2 \rightsquigarrow \mathcal{N}(33; 100)$.

   $P[Y > 10] = P[\frac{Y-33}{10} > \frac{10-33}{10}] = P[Z > -2,3] = P[Z < 2,3] = 0,98928$

Exercice 3.2 *Quelques petites démonstrations*

1. Soit X une v. a. réelle continue. Montrer que $V(X) = E(X - E(X))^2 = E(X^2) - E(X)^2$
2. Soit a un réel (scalaire), soit X une variable réelle de loi continue. Montrer que $E(a) = a$ et $V(a) = 0$. En déduire que $E(aX + b) = aE(X) + b$ et $V(aX + b) = a^2V(X)$
3. Soit a un réel positif, soit X une variable de loi continue symétrique, centrée sur 0.
   Montrer que $P(\mid X \mid < a) = 2P(X < a) - 1$

Exercice 3.3

La durée de vie d'un robot (exprimée en années), jusqu'à ce que survienne la première panne, est une variable aléatoire X qui suit une loi exponentielle de paramètre λ, $\lambda > 0$. Ainsi, la probabilité qu'un robot tombe en panne avant l'instant t est : $P(X \leq t) = \int_0^t \lambda\, e^{-\lambda x} dx$

1. Calculer λ pour que $P(X > 6) = 0,3$.

   Dans la suite de l'exercice, on prendra λ = 0,2.

2. A quel instant t la probabilité qu'un robot tombe en panne pour la première fois est-elle de 0,5 ?
3. Calculer la probabilité qu'un robot n'ait pas eu de panne au cours des deux premières années.
4. Sachant qu'un robot n'a pas eu de panne au cours des deux premières années, quelle est la probabilité qu'il soit encore en état de marche au bout de six ans ?
5. On considère un lot de dix robots fonctionnant de manière indépendante. Déterminer la probabilité que, dans ce lot, il y ait au moins trois robots qui n'ait pas eu de panne au cours des deux premières années.

Exercice 3.4

Une centrale de dépannage informatique emploie des ingénieurs en informatique. Un ingénieur au hasard part en intervention dès que survient un incident informatique. Soit T la variable représentant le temps d'attente de cet ingénieur jusqu'à ce qu'il parte en intervention. La variable T suit une loi exponentielle de paramètre $\lambda = 1/82$.

1. Calculer la probabilité que le temps à attendre une intervention soit compris entre 50 et 100.
   Calculer la probabilité que le temps à attendre une intervention soit supérieur à 300
2. Sachant qu'un ingénieur a déjà attendu 200, quelle est la probabilité qu'il attende encore 75 ?
3. Calculer le temps moyen attendu par un ingénieur avant d'intervenir.
4. On considère N ingénieurs dans cette entreprise. On suppose l'indépendance des variables associées au temps d'attente de chacun d'eux, on suppose que toutes suivent la même loi exponentielle précédente. On note X_d le nombre d'ingénieurs n'ayant eu aucun incident (donc pas d'intervention) pendant le temps d.
   a) Quelle est la loi de X_d ?

   b) Déduire le nombre moyen d'ingénieurs n'ayant jamais été appelés pendant ce temps d. Combien doit-on embaucher d'ingénieurs au total pour qu'il y ait en moyenne 3 ingénieurs qui n'ont jamais été appelés pendant un temps de 60 ?

Exercice 3.5

On suppose que la durée de vie d'un disque dur est distribuée selon une loi exponentielle. Le fabricant veut garantir que le disque dur a une probabilité inférieure à 0,001 de tomber en panne sur un an.

Quelle valeur maximale pour λ doit-il prendre ? Quelle durée de vie moyenne minimale devrait avoir le disque dur ?

Exercice 3.6

1. Soit X une variable aléatoire de loi $\mathcal{N}(0;1)$.
   a) Calculer les probabilités suivantes : $P[X < 1.45]$, $P[X < 2.01]$, $P[-1.65 < X < 1.34]$, $P[|X| < 2.05]$

   b) Calculer la valeur de u dans les cas suivants : $P[X < u] = 0.63$, $P[X > u] = 0.63$, $P[|X| < u] = 0.63$

   c) Soit $Y = 2X + 3$. Quelle est la loi de Y ? Calculer les probabilités : $P[Y < 4]$, $P[-2 < Y < 1]$.

2. Soit X une variable aléatoire de loi $\mathcal{N}(3;25)$.
   a) Calculer les probabilités suivantes : $P[X < 6]$, $P[X > -2]$, $P[-1 < X < 1.5]$

   b) Déterminer la valeur de u dans les cas suivants :
   $P[X < u] = 0.63$, $P[X > u] = 0.63$, $P[|X - 3| < u] = 0.63$

Exercice 3.7

Un ingénieur informatique met au point une certaine formule pour mesurer la complexité d'un mot de passe (tenant compte de la longueur, des caractères spécifiques ...). Il teste sa formule sur un générateur de mots de passe. On note X la v.a. correspondant à la complexité d'un mot de passe, on trouve une moyenne de 23 et un écart type de 6.

1. En utilisant l'inégalité de Bienaymé Tchebychev, donner une minoration de la probabilité qu'un mot de passe soit d'une complexité comprise entre 10 et 36.

2. On suppose maintenant que X suit la loi $\mathcal{N}(23;36)$.
   a) Calculer la probabilité qu'un mot de passe généré par ce logiciel soit de complexité supérieure à 40.

   b) Un mot de passe sera refusé si sa complexité est trop faible. Trouver cette valeur limite c_{lim} pour que la probabilité de refuser un mot de passe ne dépasse pas 0,005

   c) Déterminer le nombre positif h pour que 95 % des complexités soient dans l'intervalle $[23-h\,;\,23+h]$

   d) Reprendre maintenant la question 1) en calculant exactement la valeur de la probabilité qu'un mot de passe soit d'une complexité comprise entre 10 et 36. Comparer avec la minoration de Bienaymé Tchebychev.

FICHE - 4

Exercice 4.1

Une certaine boîte mail affirme que la probabilité qu'un courriel soit classé en tant que "spam" alors qu'il ne l'est pas réellement est p=1%.

On prend un échantillon de 1000 mails. On note X la variable aléatoire égale au nombre de mails classés spam à tort dans cet échantillon.

1. Quelle est la loi de X ? Par quelle loi peut-on approcher la loi de X (précisez les paramètres) ? Justifier cette approximation.
2. Calculer la probabilité pour qu'il y ait au plus 9 mails classés spam à tort.
3. Quel nombre minimum de faux-spam doit compter cet échantillon pour que la probabilité atteigne 20% (seuil limite pour lequel la boîte mail devra revoir sa procédure de détection de spam) ?
4. Calculer l'intervalle de fluctuation de p de niveau 95%.
5. Dans l'échantillon utilisé, on dénombre 14 faux-spam. Cet échantillon est-il cohérent ?

Solution :

1. La variable X : "nombre de mails classés spam à tort" suit une loi Binomiale de paramètres $n = 1000$ *et* $p = 0,01$. *Le théorème Central Limite permet de justifier l'approximation par une loi Gaussienne, puisque n est très grand,* $n = 1000$, $np = 10 > 5$, $n(1-p) = 990 > 5$. *On calcule les paramètres de la loi Normale :* $\mu = np = 10$ *et* $\sigma^2 = np(1-p) = 9,9$. *On déduit que la loi de X est approchée par la loi* $\mathcal{N}(10; 9, 9)$.

2. $P(X \leq 9) = P[\frac{X-10}{\sqrt{9,9}} < \frac{9-10}{\sqrt{9,9}}] = P(Y \leq -0,3178) = 1 - P(Y \leq 0,3178) \approx 1 - 0,625 \approx 0,375$ *[car Y est la variable centrée réduite associée à X, donc Y suit la loi* $\mathcal{N}(0; 1)$]

3. On note N le nombre minimum de faux-spam voulu, il doit vérifier :

$P[X \geq N] = 0,2$. *Donc,* $P[\frac{X-10}{\sqrt{9,9}} \geq \frac{N-10}{\sqrt{9,9}}] = P[Y \geq u] = 0,2 \Rightarrow P[Y < u] = 0,8 \Rightarrow u = 0,8416$.

On déduit alors : $N = 0,8416\sqrt{9,9} + 10 \approx 12,65$

Il faut au moins 13 faux-spams pour déclarer la procédure de détection mauvaise.

4. L'intervalle de fluctuation est : $\left[p \pm u_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}}\right] = \left[0,01 \pm 1,96\sqrt{\frac{0,01 * 0,99}{1000}}\right] = [0,00383; 0,01617]$

5. L'échantillon serait déclaré conforme si $X \in [1000 * 0,00383\,;\, 1000 * 0,01617]$. *Ici,* $14 \in [3,83\,;\, 16,17]$, *l'échantillon est cohérent.*

Exercice 4.2

Soit $(X_i)_i$ un échantillon de taille n de la loi exponentielle de paramètre λ. Montrer que si n est suffisamment grand, $Y = \sum X_i$ est une variable aléatoire de loi $\mathcal{N}(\frac{n}{\lambda}\ ;\ \frac{n}{\lambda^2})$.

Exercice 4.3

On sait par expérience qu'un certain logiciel antivirus bloque et élimine 90% des virus et logiciels malveillants. On met cet antivirus sur 400 ordinateurs d'une entreprise (que l'on suppose indépendants). Soit X le nombre d'ordinateurs infectés par un virus dans cette entreprise.

1. Quelle est la loi de X ? Donner une approximation de cette loi et calculer les paramètres.
2. Calculer la probabilité que l'entreprise ait au maximum 20 ordinateurs infectés.

3. Montrer que l'intervalle de fluctuation d'un pourcentage au niveau 99% est : $[p \pm 2,5758\sqrt{\frac{p(1-p)}{n}}]$. Calculer les valeurs des bornes de cet intervalle sur l'échantillon des 400 ordinateurs.
4. La semaine dernière, il y a eu 48 ordinateurs infectés. Est-ce dans la norme ?

Exercice 4.4

On considère deux sites web A et B concurrents. Sur le site A, on compte 19 clics en moyenne par heure. Sur le site B il a 21 clics en moyenne par heure. Soient X et Y les variables qui représentent le nombre de clics par jour se produisant sur le site A et B. On supposera que ces deux variables sont indépendantes.

1. Quelles sont les lois de X et Y ? Donner des approximations de ces lois, en justifiant.
2. Quelle est la probabilité qu'un jour il y ait plus de clics sur A que sur B ?
3. On considère un échantillon de X de taille n=30 (étude sur un mois), ainsi qu'un échantillon de Y de taille n=30. On suppose l'indépendance de ces variables. Calculer l'intervalle de fluctuation du nombre de clics moyen sur A au niveau 95%, puis calculer celui de B.
4. Pour bénéficier de certains avantages, il faut dépasser 485 clics par jour. Les deux sites annoncent qu'ils dépassent cette limite, est-ce crédible ? Expliquer

Exercice 4.5

Soit X la variable aléatoire modélisant le temps de compilation d'une page de codes avec un certain logiciel. On relève ce temps sur un échantillon de 90 pages compilées. Le tableau ci-dessous donne la répartition par classe :

Temps de compilation (en seconde)	[0,12 ;0,18[	[0,18 ;0,24[	[0,24 ;0,3[	[0,3 ;0,36[	[0,36 ;0,42[
Nombre de pages observées	10	22	29	20	9

On veut savoir si le temps de compilation est distribuée selon une loi Normale de moyenne 0,27 et d'écart type 0,07. Faire une étude graphique pour répondre à cette question (droite de Henry).

Exercice 4.6

Soit X une variable aléatoire de loi $\mathcal{N}(0;1)$. Soit Z la variable alétoire définie par $Z = X^2$.

1. Quelle est la loi de Z ?
2. Calculer $P(Z \leq 2,706)$ en utilisant uniquement la table de la loi Normale. Vérifier votre résultat avec la table de la loi de Z.

Exercice 4.7

1. Soit X une variable aléatoire de loi $\mathcal{T}_9$.
   a) Calculer les probabilités suivantes : $P[X < 1,833]$, $P[X < -2,821]$.
   b) Trouver la valeur de x dans les cas suivants : $P[X > x] = 0,90$; $P[|X| < x] = 0,95$
2. Soit X une variable aléatoire de loi $\mathcal{X}^2_7$.
   a) Calculer les probabilités suivantes : $P[X < 12,02]$, $P[2,833 < X < 6,346]$
   b) Trouver la valeur de x dans le cas suivant : $P[X > x] = 0,90$

FICHE - 5

Exercice 5.1

1. Soit X une variable aléatoire de loi continue uniforme sur [0,θ].
   a) Rappeler la fonction de densité de cette loi, son espérance et sa variance.
   b) Déterminer l'estimateur du maximum de vraisemblance de θ.
   c) Cet estimateur est-il biaisé ? En déduire un estimateur sans biais.

2. Soit X une variable aléatoire de loi de Poisson de paramètre θ.
   a) Rappeler la fonction de densité de cette loi, son espérance et sa variance.
   b) Déterminer l'estimateur du maximum de vraisemblance de θ.
   c) Cet estimateur est-il biaisé ?

Solution :

1. *a) La fonction de densité : f(x)=*$\frac{1}{\theta}$ *sur l'intervalle* $[0,\theta]$*, et 0 ailleurs.*

   $E(X) = \int_0^\theta \frac{x}{\theta}\,dx = \left[\frac{x^2}{2\theta}\right]_0^\theta = \frac{\theta}{2}$ *et* $V(X) = \int_0^\theta \frac{x^2}{\theta}\,dx - \left(\frac{\theta}{2}\right)^2 = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}$

   b) La vraisemblance : $V(x_1,...,x_n,\theta) = \frac{1}{\theta^n}$ *si* $\theta > max\{x_1,...,x_n\}$*, 0 sinon. La fonction V (fonction de* θ*) est nulle entre 0 et* $max\{x_1,...,x_n\}$*, puis elle est clairement décroissante, mais positive. Le maximum est donc atteint au point* $\hat{\theta} = max\{X_1,...,X_n\}$*. C'est le maximum de vraisemblance (chercher le point qui annulle la dérivée est inutile dans ce cas).*

   c) On déduit la loi de l'estimateur $\hat{\theta}$ *à partir de celle de* θ *:* $P(X \leq t) = t/\theta$ *si* $t < \theta$*, 0 sinon.*

   Donc, $F(t) = P(\hat{\theta} \leq t) = \prod P(X_i \leq t) = [P(X \leq t)]^n = \left(\frac{t}{\theta}\right)^n$ *si* $t < \theta$*, et F(t)=0 sinon.*

   On conclut : $E(\hat{\theta}) = \int_0^\theta t\,f(t)\,dt = \int_0^\theta t\,n\,t^{n-1}/\theta^n\,dt = \int_0^\theta n\left(\frac{t}{\theta}\right)^n dt = \frac{n\theta}{n+1}$*. L'estimateur est donc biaisé.*

   Son biais est : $E(\hat{\theta}) - \theta = \frac{-\theta}{n+1}$*. Un estimateur sans biais serait* $\frac{n+1}{n}\,max\{X_1,...,X_n\}$

2. *a) La loi de Poisson est discrète :* $P(X = x_i) = \frac{e^{-\theta}\theta^x}{x!}$ *. De plus,* $E(X) = \theta$ *et* $V(X) = \theta$

   b) La vraisemblance : $V(x_1,...,x_n,\theta) = P(X_1 = x_1, ..., X_n = x_n) = \prod \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\Sigma x_i}}{\prod x_i!}$*. De plus,*

   $Ln(V(x_1,...,x_n,\theta)) = -n\theta + (\Sigma x_i)\,ln(\theta) - \sum ln(x_i!)$*. On déduit :* $\frac{\partial Ln(V)}{\partial\theta} = -n + \frac{1}{\theta}\sum x_i = 0 \quad \Rightarrow$

   $\widehat{\theta} = \frac{1}{n}\sum X_i$*. De plus, on a bien* $\frac{\partial^2 Ln(V)}{\partial\theta^2} = -\frac{1}{\theta^2}\sum x_i \leq 0$*, ce qui justifie le maximum.*

   c) $E(\hat{\theta}) = E(\overline{X}) = \theta$*. Estimateur sans biais.*

Exercice 5.2

Soit X une variable aléatoire de loi exponentielle de paramètre $\lambda = 1/\theta$.

a) Rappeler la fonction de densité de cette loi, son espérance et sa variance.

b) Déterminer l'estimateur du maximum de vraisemblance de θ.

c) Cet estimateur est-il biaisé ? Est-il de variance asymptotiquement nulle ?

Exercice 5.3

Soit X une variable aléatoire de loi de Bernoulli de paramètre p.

a) Rappeler cette loi de probabilité , son espérance et sa variance.

b) Déterminer l'estimateur du maximum de vraisemblance de p.

c) Cet estimateur est-il biaisé ?

Exercice 5.4

Les durées de vie des batteries d'ordinateurs portables ont été testées. On note X la variable aléatoire modélisant cette durée. On suppose que X suit une loi Normale de moyenne μ et de variance σ^2, toutes deux inconnues. Sur un échantillon de 20 batteries, on a relevé ces durées de vie (exprimées en mois). Les mesures ont donné : $\sum x_i = 605$ et $\sum x_i^2 = 20415$.

1. Calculer les estimateurs du maximum de vraisemblance de μ et σ^2 pour la loi Normale.
2. Montrer que $\overline{X}$ est un estimateur sans biais, convergent et de variance asymptotiquement nulle.
3. Calculer la moyenne empirique et la variance empirique de cet échantillon.
4. Donner une estimation sans biais de la moyenne ainsi qu'une estimation sans biais de la variance de cet échantillon.

Exercice 5.5

Soit X une variable de loi $\mathcal{N}(\mu\ ;\ \sigma^2)$. On note S'^2 l'estimateur sans biais de la variance vu précédemment.

1. Montrer que $\dfrac{(n-1)S'^2}{\sigma^2}$ suit une loi du Khi-deux à n-1 degrés de liberté.
2. Montrer que $\dfrac{\overline{X}-\mu}{S'/\sqrt{n}}$ suit une loi de Student à n-1 degrés de liberté.
3. Sur les données de l'exercice précédent, montrer que μ est dans l'intervalle [30,003 ; 30,497] avec une probabilité de $0,95$.

FICHE - 6

Exercice 6.1

Un échantillon de 100 batteries de téléphones portables ont été testées. On a relevé 57 batteries dont les durées de vie correspondaient à ce qu'annonçait le fabricant. On notera p la probabilité qu'une batterie ait une durée de vie adéquat à ce qu'annonce le fabricant.

1. Donner un estimateur de p, puis une estimation. Par quelle loi peut-on approcher cet estimateur ? Justifier précisément.
2. Montrer que l'intervalle de confiance de p au niveau 95% est : $[F \pm 1,96\sqrt{\frac{F(1-F)}{n}}]$. Calculer les bornes de cet intervalle sur l'échantillon des 100 batteries. Interpréter.
3. Quel niveau de confiance faut-il prendre pour avoir une précision sur ce pourcentage de batteries analogues à ce qu'annonce le fabricant de $\pm 5\%$?

Solution :

1. *Un estimateur de p est donné par $\hat{p} = F = X/n$, où X représente le nombre de batteries adéquat dans l'échantillon des 100. Sur cet échantillon, l'estimation est de f=57/100=0,57. De plus, X suit une loi Binomiale de paramètres n=100 et p. D'après le théorème central limite (n grand), on a l'approximation : $X \rightsquigarrow \mathcal{N}(np; np(1-p))$. On déduit que $F \rightsquigarrow \mathcal{N}(p\,;\, \frac{p(1-p)}{n})$.*

2. $P(p \in [a;b]) = 0,95 \quad \Leftrightarrow \quad P\left(\frac{F-b}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{F-a}{\sqrt{\frac{p(1-p)}{n}}}\right) = 0,95$ *après centrage et réduction.*

   En estimant p par F, et n étant grand, on obtient : $P\left(\frac{F-b}{\sqrt{\frac{F(1-F)}{n}}} \leq \frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} \leq \frac{F-a}{\sqrt{\frac{F(1-F)}{n}}}\right) = 0,95$ *;*

   sur la table de la loi $\mathcal{N}(0;1)$, on lit la valeur $u_{(0,975)} = 1,96$ à une probabilité de $1-\alpha/2 = 0,975$. En égalisant $u_{(0,975)} = \frac{F-a}{\sqrt{\frac{F(1-F)}{n}}}$, *on déduit l'intervalle de confiance de p :* $\left[F - u_{1-\alpha/2}\,\frac{\sqrt{F(1-F)}}{\sqrt{n}}\,;\, F + u_{1-\alpha/2}\,\frac{\sqrt{F(1-F)}}{\sqrt{n}}\right]$.

   Pour cet échantillon, les bornes sont : $\left[0,57 \pm 1,96\sqrt{\frac{0,57*0,43}{100}}\right] = [0,473; 0,667]$. *En conséquence, il y a entre 47,3% et 66,7% des batteries dont la durée de vie est adéquate à ce qu'annonce le fabricant.*

3. *Pour avoir un intervalle de précision $\pm 5\%$, il faut que* $u_{1-\alpha/2}\,\frac{\sqrt{f(1-f)}}{\sqrt{n}} = 0,05$, *soit* $u_{1-\alpha/2} = \frac{0,05}{0,0495} = 1,01$. *Sur la table de la loi Normale centrée réduite, on lit $1-\alpha/2 = 0,8438$. On déduit : $\alpha = 0,3124$. Pour avoir un encadrement de p entre 52% et 62%, le niveau de confiance sera seulement de 68,8%.*

Exercice 6.2

On étudie les connexions des internautes à un certain site web. Ce site propose trois versions de son contenu selon la taille de l'écran utilisé pour se connecter. On notera les versions "p" pour les petits écrans tels que les téléphones portables, "m" pour les écrans moyens tels que les tablettes et "n" pour les écrans normaux des ordinateurs. Sur un échantillon de 6000 connexions observées par ce site, la moitié des internautes utilisaient un ordinateur (ils utilisaient donc la version "n"), et il y avaient deux fois plus d'utilisateurs de téléphones portables que de tablettes.

1. Pour cet échantillon, déterminer les fréquences d'utilisation de chacune des 3 versions.
2. Donner un intervalle de confiance de la probabilité d'utilisation de la version "m" au niveau de confiance de 99%.
3. On utilise maintenant un autre échantillon de taille 600 connexions sur lequel on dénombre 99 connexions avec la version "m".

   a) Donner l'intervalle de confiance d'utilisation de la version "m" au niveau de confiance de 99%. Comparer avec la question 2).

   b) Quel niveau de confiance devrait-on prendre pour avoir la même amplitude (précision) d'intervalle qu'à la question 2) ?

Exercice 6.3

Dans une usine fabricant une certaine pièce métallique, des études ont montré que la masse de ces pièces peut être considérée comme une variable aléatoire X de loi Normale de moyenne et variance inconnues. On prend un échantillon de 40 pièces fabriquées dont on mesure les masses, en grammes. On a les résultats suivants : $\sum x_i = 2204$ et $\sum x_i^2 = 121721$.

1. Montrer que l'intervalle de confiance de la variance au niveau 95% est $\left[\frac{nS^2}{58,1} ; \frac{nS^2}{23,65}\right]$. Donner les bornes de cet intervalle pour cet échantillon.
2. Donner un intervalle de confiance de la masse moyenne au niveau de confiance 99% puis 95%.
3. Le fabricant précise que ses machines sont réglées pour fabriquer des pièces avec une dispersion de 2,7g. Que pouvez-vous conclure ?
4. Refaire alors les intervalles de confiance de la masse moyenne de cette pièce au niveau de confiance 99% puis 95% en tenant compte de l'information donnée par le fabricant. Comparer avec les résultats de la question 2).

Exercice 6.4

Une compagnie d'assurance réalise une étude pour couvrir les risques d'une clinique. Celle-ci propose une nouvelle opération chirurgicale. Il y a eu 49 échecs sur 400 tentatives. On note p le pourcentage de réussite de cette nouvelle opération.

1. Quelle estimation de p proposez-vous ? En utilisant et justifiant l'approximation par la loi Normale , donner un intervalle de confiance pour p de niveau de confiance 0,95.
2. Combien d'opérations la clinique devrait-elle réaliser pour encadrer le pourcentage de réussite avec une précision de plus ou moins 1%, au niveau de confiance 0,95 ?
3. L'assurance accepte de couvrir un certain nombre d'opérations ratées N à condition que ce nombre n'ait que 1% de risque d'être dépassé. Quel est ce nombre N ?

Exercice 6.5

Afin de ne pas confondre intervalle de fluctuation et intervalle de confiance :

1. a) Pour faire un intervalle de fluctuation, quels sont les paramètres connus ? Que cherche-t-on à savoir ?

   b) Pour faire un intervalle de confiance, quels sont les paramètres connus ? Que cherche-t-on à savoir ?
2. a) Lors d'une élection, on fait un sondage d'intention de votes sur un échantillon de taille n = 1000. Les intentions de votes pour le candidat X sont de 51%. Quel intervalle permet d'encadrer le pourcentage de votes qu'obtiendra X au niveau de 95% ?

   b) Lors d'une élection, le candidat X a obtenu 51% de votes. On prend un échantillon de taille n = 1000. Quel intervalle permet d'encadrer le pourcentage de votes obtenu par X dans l'échantillon au niveau 95% ?

FICHE - 7

Exercice 7.1

Soit X la variable aléatoire représentant l'indice de radoactivité de l'eau (en picocuries par litre). On modélise X par une loi $\mathcal{N}(\mu;\sigma^2)$. On admet que l'écart-type est connu, et vaut 1. Les normes fixent à 5 l'indice moyen de radioactivité maximal pour une eau potable. La compagnie d'assainissement et de traitement de l'eau déclare que la radioactivité de son eau n'est pas supérieure à 5.

1. La compagnie des eaux souhaite montrer que son eau est aux normes. Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ doit-elle tester ? Etablir la règle de décision de ce test aux seuils de 5% et 1%.
2. Une association de consommateurs veut démontrer que l'eau n'est pas aux normes (taux de radioactivité trop importante). Quelles hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ doit-elle tester ? Etablir la règle de décision de ce test aux seuils de 5% et 1%.
3. Des experts contrôlent cet indice sur un échantillon de n=49, l'indice de radioactivité moyen vaut 5,2. Quelles sont les conclusions des deux tests pour un seuil de 5% ?

Solution :

1. *Par définition, le risque de première espèce est le risque minimisé dans un test : $\alpha = P_{\mathcal{H}_0}(\mathcal{H}_1)$ =P(accepter $\mathcal{H}_1$ alors qu'en réalité $\mathcal{H}_0$ est vraie). Si on pose les hypothèses suivantes : $\mathcal{H}_0 : \mu = 5$ contre $\mathcal{H}_1 : \mu > 5$. On obtient α =P(déclarer hors normes la compagnie non polluante). C'est le risque que veut évidemment minimiser la compagnie des eaux.*

   La statistique du test sur la moyenne, avec σ connu, est : $T = \dfrac{\overline{X}-5}{\sigma/\sqrt{n}}$. Elle suit une loi $\mathcal{N}(0;1)$. La règle de décision pour le test unilatéral supérieur est : { Rejet de $\mathcal{H}_0 \iff T > u_{1-\alpha}$ }. Pour $\alpha = 5\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff T > 1,6449$ }. Pour $\alpha = 1\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff T > 2,3263$ }.

   Ou alors : { Rejet de $\mathcal{H}_0 \iff \overline{X} > \mu_0 + u_{1-\alpha}\,\sigma/\sqrt{n}$}. Pour $\alpha = 5\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff \overline{X} > 5 + 1,6449/\sqrt{n}$ }. Pour $\alpha = 1\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff \overline{X} > 5 + 2,3263/\sqrt{n}$ }.

2. *Par le même raisonnement, on pose les hypothèses suivantes : $\mathcal{H}_0 : \mu = 5$ contre $\mathcal{H}_1 : \mu < 5$. On obtient $\alpha = P_{\mathcal{H}_0}(\mathcal{H}_1)$ =P(accepter $\mathcal{H}_1$ alors qu'en réalité $\mathcal{H}_0$ est vraie) = P(déclarer aux normes une compagnie des eaux polluante). C'est le risque que veut évidemment minimiser l'association des consommateurs.*

   La statistique du test sur la moyenne, avec σ connu, est : $T = \dfrac{\overline{X}-5}{\sigma/\sqrt{n}}$. Elle suit une loi $\mathcal{N}(0;1)$. La règle de décision pour le test unilatéral inférieur est : { Rejet de $\mathcal{H}_0 \iff T < -u_{1-\alpha}$ }. Pour $\alpha = 5\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff T < -1,6449$ }. Pour $\alpha = 1\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff T < -2,3263$ }.

   Ou alors : { Rejet de $\mathcal{H}_0 \iff \overline{X} < \mu_0 - u_{1-\alpha}\,\sigma/\sqrt{n}$}. Pour $\alpha = 5\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff \overline{X} < 5 - 1,6449/\sqrt{n}$ }. Pour $\alpha = 1\%$, on obtient : { Rejet de $\mathcal{H}_0 \iff \overline{X} < 5 - 2,3263/\sqrt{n}$ }.

3. *Ici, on fait un test sur un échantillon de taille n=49, $\overline{x}$ = 5,2. La valeur prise par la statistique est $\dfrac{5,2-5}{1/7} = 1,4$.*

   Pour le test fait par la compagnie des eaux : la règle est { Rejet de $\mathcal{H}_0 \iff T > 1,6449$ } ce qui n'est pas validée. On accepte donc $\mathcal{H}_0$, on déclare alors la compagnie des eaux aux normes.
   Pour le test fait par l'association des consommateurs : la règle est : { Rejet de $\mathcal{H}_0 \iff T < -1,6449$ } ce qui n'est pas validée. On accepte $\mathcal{H}_0$, on déclare donc la compagnie hors-normes.

   Ou alors, pour le test fait par la compagnie des eaux, la règle est : { Rejet de $\mathcal{H}_0 \iff \overline{X} > 5+1,6449/\sqrt{49}$ =5,235} ce qui n'est pas validée. On accepte donc $\mathcal{H}_0$, on déclare alors la compagnie des eaux aux normes.
   Pour le test fait par les consommateurs, la règle est : { Rejet de $\mathcal{H}_0 \iff \overline{X} < 5 - 1,6449/\sqrt{49} = 4,765$} ce qui n'est pas validée. On accepte donc $\mathcal{H}_0$,on déclare donc la compagnie hors-normes.

Exercice 7.2

La législation en vigueur impose aux aéroports certaines normes concernant les bruits émis par les avions au décollage et à l'atterrissage. Dans les zones habitées proches d'un aéroport, la limite tolérée se situe à 80 décibels, au-delà de cette limite, l'aéroport doit indemniser les riverains.

Les habitants d'un village voisin d'un aéroport assurent que le bruit atteint la valeur limite de 80 décibels. L'aéroport affirme qu'il n'est que de 78 décibels. Afin de trancher entre les deux parties, on mesure X l'intensité du bruit sur un échantillon de 100 avions. On admet que X suit une loi normale de moyenne μ et de variance $\sigma^2 = 49$ (décibels²).

1. On choisit de faire le test : $\mathcal{H}_0 : \mu = 80$ contre $\mathcal{H}_1 : \mu = 78$ [Soit le test : $\mathcal{H}_0 : \mu = 80$ contre $\mathcal{H}_1 : \mu < 80$]
   a) Donner l'interprétation du risque de première espèce de ce test. Qui a intérêt à faire ce test ?
   b) Déterminer la règle de décision pour un risque première espèce $\alpha = 5\%$.
   c) Que décide-t-on si le bruit moyen émis par les 100 avions est de 79,1 db ?
2. On choisit maintenant de faire le test : $\mathcal{H}_0 : \mu = 78$ contre $\mathcal{H}_1 : \mu = 80$. [Soit le test : $\mathcal{H}_0 : \mu = 78$ contre $\mathcal{H}_1 : \mu > 78$].
   a) Reprendre les trois questions du 1.
   b) Comparer les décisions prises dans les deux tests. Que constatez-vous ?
   c) Sur quels paramètres peut-on jouer pour sortir de cette contradiction ?

Exercice 7.3

Un radar aérien de surveillance en rotation envoie 1200 impulsions par seconde. S'il y a présence d'une cible, 20 impulsions la toucheraient en un seul balayage et seraient réfléchies. On analyse des résultats et on note $z_1, ..., z_{20}$ les divers bruits parasites sur ces 20 impulsions, indépendants. On suppose que $Z \rightsquigarrow \mathcal{N}(0; \sigma^2)$ avec $\sigma = 0,6$. Lors du traitement d'images, la transformation $X_i = 1 + Z_i$ traduit la présence d'une cible alors que $X_i = 0 + Z_i$ signifie la non présence d'une cible.

1. Déterminer la loi de X et interpréter la signification du test suivant : $\mathcal{H}_0 : \mu = 0$ contre $\mathcal{H}_1 : \mu = 1$
2. Lors d'un balayage on trouve une moyenne empirique de X de 0,45. Faire le test précédent pour les risques $\alpha = 10^{-3}$, $\alpha = 10^{-4}$ et $\alpha = 10^{-7}$
3. On suppose maintenant σ inconnu. On trouve une variance empirique de X de 0,5. Refaire la question 2.

Exercice 7.4

La construction d'une autoroute provoque une division de l'opinion parmi les habitants des agglomérations environnantes. Les uns y sont favorables, les autres s'y opposent pour des causes écologiques, de danger et de nuisance. Les politiques de la région pensent qu'il est raisonnable de construire cette autoroute si 65 % des avis sont favorables.

1. On considère le test : $\mathcal{H}_0 : p = 0,65$ contre $\mathcal{H}_1 : p < 0,65$
   a) On prend $\alpha = 5\%$. Expliciter précisément le risque de première espèce de ce test.
   b) Un sondage est organisé sur un échantillon de 200 habitants, on trouve que 62 % des personnes interrogées sont favorables à la construction. Faire le test et conclure que cette autoroute pourrait quand même être construite.
2. Après discussion avec les habitants mécontents, les politiques font appel à des statisticiens. Ils conviennent de changer la démarche : on descend la limite à 63% d'avis favorables, on fixe à 3% le risque de refuser à tort la construction de l'autoroute, et on prend un échantillon de taille supérieure à 200.
   a) Construire les hypothèses du test correspondant.
   b) Quelle doit-être la taille d'échantillon à interroger pour être sûr de refuser la construction de l'autoroute avec cette nouvelle procédure ? (On suppose qu'il y aura toujours 62% des personnes interrogées favorables à la construction).
   c) Calculer la taille de l'échantillon à interroger pour être sûr de refuser la construction de l'autoroute dans la première procédure. A qui profite-t-elle le plus la nouvelle procédure ?

Exercice 7.5

On considère le test paramétrique bilatéral de la moyenne d'une loi Normale de variance inconnue.

Montrer qu'on rejette $\mathcal{H}_0$ si et seulement si $\mu \notin \left[\overline{x} - t^{n-1}_{1-\alpha/2} \frac{s}{\sqrt{n-1}} ; \overline{x} + t^{n-1}_{1-\alpha/2} \frac{s}{\sqrt{n-1}} \right]$.

FICHE - 8

Exercice 8.1

On reprend l'exercice 7.1. On rappelle que les normes fixent à 5 l'indice moyen de radioactivité maximal pour une eau potable. Suite à un nouveau traitement, la compagnie des eaux déclare que la radioactivité de son eau est de 4. Sur un échantillon de taille 9, on trouve une moyenne de 4,5. On suppose toujours la variance connue et égale à 1.

1. Faire le test suivant : $\mathcal{H}_0 : \mu = 5$ contre $\mathcal{H}_1 : \mu = 4$ au seuil de 5% et conclure.
2. Calculer la p_{valeur} de ce test.
3. Calculer la puissance de ce test.

Solution :

1. *a) Pour ce test : { Rejet de* $\mathcal{H}_0 \iff T < -u_{1-\alpha}$ *} avec* $T = \dfrac{\overline{X} - 5}{\sigma/\sqrt{n}}$. *La valeur calculée de T est -1,5 et* $u_{0,95} = 1,6449$, *on accepte donc* $\mathcal{H}_0$. *(Ou alors :* c*=5-1,6449/*$\sqrt{9}$*=4,45* $< \overline{x} = 4,5$ *donc on accepte* $\mathcal{H}_0$*). On conclurait donc que les eaux seraient toujours polluées avec une probabilité 95%.*
2. *Ici,* $p_{valeur} = P(T < -1,5) = 1 - P(T < 1,5)$. *La table* $\mathcal{N}(0;1)$ *donne* $P(T < 1,5) \simeq 0,9332$, *donc* $p_{valeur} \simeq 0,0668$. *En conséquence, pour tous les risques* $\alpha < p_{valeur}$, *on accepterait* $\mathcal{H}_0$, *on conclurait que le traitement ne serait pas assez efficace (eaux polluées). Si on prend des risques grands* ($\alpha > 6,68\%$), *on conclurait que les eaux ne sont plus polluées.*
3. L*e risque de seconde espèce* $\beta = P($*accepter* $\mathcal{H}_0 |$ $\mathcal{H}_1$ *vraie*$) = P(\overline{X} > c \mid \mu = 4) = P(\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \dfrac{c - \mu}{\sigma/\sqrt{n}} \mid \mu = 4) = P(\dfrac{\overline{X} - 4}{\sigma/\sqrt{n}} > \dfrac{c - 4}{\sigma/\sqrt{n}}) = P(T > 1,35) \simeq 1 - 0,9115 \simeq 0,0885$. *La puissance de ce test est donc de 0,9115 c'est la probabilité d'accepter* $\mathcal{H}_1$ *avec raison (probabilité de déclarer les eaux non polluées avec raison).*

Exercice 8.2 On reprend l'exercice 7.2.

Pour chacun des deux tests faits dans les questions 1 et 2, calculer les p_{valeur}.

Calculer ensuite les risques de seconde espèce de ces deux tests. Qu'en concluez-vous ?

Exercice 8.3

On désigne par p la probabilité de réussite d'un certain évènement A. On veut tester : $p = \dfrac{9}{16}$ contre $p = \dfrac{9}{15}$. Pour cela, on observe sur un échantillon de taille 2400 le nombre d'individus pour lesquels il y a eu réussite de cet évènement. Si ce nombre est inférieur ou égal à 1400 on acceptera $p = \dfrac{9}{16}$, mais s'il est supérieur on acceptera $p = \dfrac{9}{15}$.

1. Justifier le principe de ce test.
2. Calculer les risques de première et seconde espèce.
3. On observe 1405 individus sur 2400 pour lesquels l'évènement s'est réalisé.
   (a) Calculer la p_{valeur} de ce test par un encadrement avec les tables, puis donner la valeur exacte avec la calculatrice.
   (b) Quelle est la probabilité d'accepter $p = \dfrac{9}{15}$ avec raison ?

Exercice 8.4

A l'occasion d'une élection, un sondage sur un échantillon représentatif de la population est effectué. Au second tour de cette élection, les électeurs devront choisir entre les candidats "A" et "B". Sur 3500 personnes, 700 ne voteront pas ou voteront "blanc", 1500 déclarent qu'ils voteront pour "A" et 1300 voteraient pour "B". Au seuil de 1%, pouvez-vous déclarer que le candidat "A" serait élu lors de l'élection ? Donner la p_{valeur} de ce test.

Exercice 8.5

Le temps de téléchargement d'un certain film est une variable aléatoire notée X (elle varie selon votre ordinateur, le moment dans la journée, ect). On modélise X par une loi normale $\mathcal{N}(\mu, \sigma^2)$. Le but est de déterminer les deux paramètres de cette loi. On dispose de l'échantillon de taille $n = 15$ suivant :

X (mn)	19,8	22,1	21,5	20,9	22	21	22,3	21	20,3	20,9	22	20,8	21,2	22	21

Indications numériques : $\sum x_i$= 318,8 et $\sum x_i^2$= 6782, 78.

1. Calculer la moyenne empirique et la variance empirique de cet échantillon.
2. Faire le test de la variance suivant : $\mathcal{H}_0 : \sigma^2 = 0,5$ contre $\mathcal{H}_1 : \sigma^2 \neq 0,5$ au seuil de 5%.
3. Donner un encadrement de la p_{valeur}, puis la valeur exacte. Conclure quant à la valeur de σ^2.
4. Faire le test de la moyenne suivant : $\mathcal{H}_0 : \mu = 21$ contre $\mathcal{H}_1 : \mu > 21$ au seuil de 5%.
5. Donner un encadrement de la p_{valeur} pour le test précédent, puis la valeur exacte.
6. Donner une conclusion à toute cette étude.

Exercice 8.6

Soit X une variable aléatoire de loi Normale $\mathcal{N}(600; 100^2)$. Des chercheurs prétendent avoir augmenté l'espérance de cette variable et qu'elle serait maintenant égale à 650. On prend donc un échantillon de taille 9, on trouve une moyenne de 610,2.

1. Faire le test suivant : $\mathcal{H}_0 : \mu = 600$ contre $\mathcal{H}_1 : \mu = 650$ au seuil de 5% et conclure (Vous pourrez calculer la p_{valeur} de ce test).
2. Calculer la puissance de ce test. Qu'en pensez-vous ?
3. En conservant le même risque $\alpha = 5\%$, calculer la taille de l'échantillon qu'il faudrait prendre pour avoir une puissance de 95%.

Exercice 8.7

Soit X une variable aléatoire de loi P. Soit (X_1, ..., X_n) un échantillon de taille n. On veut savoir si la médiane m_e de X est égale à une valeur réelle donnée m_0. On construit un test de la médiane au seuil α.

1. Quelle est la loi du nombre de valeurs de l'échantillon inférieures à m_0 ? On notera Y cette variable aléatoire. Donner la loi approchée de Y lorsque n est suffisamment grand.
2. Construire la statistique du test : $\{\mathcal{H}_0 : m_e = m_0$ contre $\mathcal{H}_1 : m_e \neq m_0 \}$, ainsi que la règle de décision.
3. Que représente ce test pour $m_0 = 0$?

Exercice 8.8

Une compagnie d'assurance sur la vie étudie la proportion de clients décédés dans les 10 années suivant la souscription d'un certain contrat. Sur un échantillon de 979 assurés pour ce contrat, la compagnie d'assurance dénombre 48 décès. La compagnie d'assurance souhaite savoir si la proportion de décès est trop importante, c'est à dire que si cette proportion est au moins de 5%, il faudrait augmenter les cotisations. Faire le test adéquat au seuil de 1%.

Exercice 8.9

La proportion de voyelles dans les textes anglais, estimée à partir de la liste des mots anglais contenus dans un dictionnaire informatisé est de 40,4%. On a compté les lettres dans une chanson des Beatles Yesterday. On a observé 223 voyelles sur 483 lettres.

1. La proportion de voyelles dans l'échantillon constitué par la chanson Yesterday est-elle conforme à la proportion théorique estimée à partir de la concaténation de tous les mots du dictionnaire informatisé ?
2. Dans le livre Alice's Adventures in Wonderland de Lewis Carroll, on a dénombré 49871 voyelles sur122990 lettres. Que dire sur la conformité de la proportion des voyelles de ce recueil par rapport à la proportion théorique ?
3. Quelles hypothèses pouvez-vous faire pour expliquer ces résultats ? (Attention aux conclusions trop rapides et hasardeuses !!)

FICHE - 9

Exercice 9.1

On étudie le temps de compilation d'une page de codes avec un certain logiciel. On relève ce temps sur un échantillon de taille $n = 11$ pages de codes.

On utilise ensuite une nouvelle version de ce logiciel sur ces mêmes pages. On note X et Y les temps de compilation avec l'ancienne version du logiciel et la nouvelle.

Y (ancienne)	3,28	2,94	3,5	3,07	3,04	3,12	2,85	3,05	3,32	2,89	3,11
X (nouvelle)	1,65	1,5	1,18	2,19	1,11	1,48	1,75	2,07	1,44	1,85	2,21

1. Mettre en oeuvre toute la méthode qui permet de savoir si la nouvelle version du logiciel est plus rapide au seuil de 5%. (Justifiez votre choix de méthode, précisez les conditions nécessaires, faire les calculs et conclure explicitement).
2. Calculez un encadrement de la p_{valeur}. Conclure explicitement.

Solution :

1. *Il s'agit ici d'un seul échantillon d'individus (11 pages) sur lequel on mesure une variable, c'est donc un test de comparaison de deux échantillons appariés . Pour appliquer le test, il est nécessaire que Y-X suive une loi normale. De plus, pour tester la rapidité, on fait le test sur les moyennes. Si ($\mu_y > \mu_x$) alors il y aura efficacité de la nouvelle version, donc on pose, $\mathcal{H}_1$: $\mu_y - \mu_x > 0$. Le test est donc un test unilatéral supérieur : $\mathcal{H}_0$: $\mu_y - \mu_x = 0$. contre $\mathcal{H}_1$: $\mu_y - \mu_x > 0$.*
   La statistique est : $T = \dfrac{\bar{D}\sqrt{n-1}}{s_d} \rightsquigarrow \mathcal{T}_{n-1}$. La valeur prise par la statistique sur l'échantillon est $9,7$. La lecture de la table donne : $t^{10}_{0,95} = 1,812$.
   La règle de décision du rejet de $\mathcal{H}_0$ est validée, donc on accepte $\mathcal{H}_1$. Le temps de compilation diminue avec la nouvelle version, au seuil de 5%.
2. *Ici, $p_{valeur} = P(T > 9,7) = 1 - P(T < 9,7)$. La table $\mathcal{T}_{10}$ donne $P(T < 9,7) \gg 0,9995$, donc $p_{valeur} \ll 0,0005$. En conséquence, pour $\alpha > p_{valeur}$, on accepte $\mathcal{H}_1$. On conclut, avec un risque quasiment nul de se tromper, que cette nouvelle version est efficace et que la compilation est plus rapide.*

Exercice 9.2

Dans une étude faite dans le cadre de la prévention routière, on veut comparer le temps de réaction nécessaire pour arrêter une automobile chez des sujets sous influence d'alcool (environ 0,09 litre) et chez les mêmes sujets avant ingestion d'alcool. Le temps de réaction est mesuré en centième de seconde sur 15 individus.

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Avant	33	29	26	23	21	36	27	38	22	33	42	35	22	39	37
Après	46	41	37	37	30	43	38	47	33	42	54	48	33	54	50

1. Justifier la méthode adéquate qui permet de comparer le temps moyen de réaction entre ces deux séries de données. Donner les conditions nécessaires. Poser les hypothèses du test. Faire ce test au seuil de 5%. Conclure.
2. Donner un encadrement de la p_{valeur}, puis calculer la valeur exacte. Conclure sur l'influence de 0,09 litre d'alcool.

Exercice 9.3

On mesure la force statique par dynamométrie manuelle (exprimée en kg) chez 10 enfants atteints de trisomie 21. Un premier test est realisé en Septembre. Puis, sur une période de six mois, ces enfants essaient de développer, sous forme de jeux, leur force statique. Un second test est alors réalisé en Février. On obtient les résultats suivants :

Septembre	33	42,5	54	60	61	68	68	69	72	86
Février	38	45	52	63	61	75	66	70	81	90

On désire savoir si le nouveau programme suivi par ces enfants a permis d'améliorer leur résultat.

1. Poser les hypothèses nécessaires et faire le test permettant de conclure au seuil de 5%.
2. Donner un encadrement de la p_{valeur}, puis la valeur exacte. Conclure sur l'efficacité de cette méthode.

Exercice 9.4

L'étude suivante consiste à analyser un dossier contenant 10001 chevaux de Troie, virus et spywares avec deux antivirus A et B sélectionnés. Toutes ces menaces ont été collectées entre le 23 novembre 2013 et le 20 janvier 2014. Les deux antivirus sont réglés avec les paramètres par défaut.

On installe ce dossier sur 9 ordinateurs. Un antivirus mobilise notamment la mémoire vive de l'ordinateur. On mesure la quantité totale de mémoire vive occupée sur notre système lors d'une analyse complète lancée avec chaque antivirus, pour chaque ordinateur.

On suppose que la différence des quantités de mémoire occupée est modélisée par une variable de loi Gaussienne.

Ordinateur	1	2	3	4	5	6	7	8	9
A (Mo)	2530	2723	3802	2952	3525	2888	2662	3637	2124
B (Mo)	2411	2372	3521	2036	3536	2996	2125	3548	2005

On désire comparer ces deux antivirus et vérifier s'il y a une différence signicative de l'occupation de la mémoire pendant l'analyse.

1. Quelle méthode est appropriée pour répondre au problème ? Justifier, donner les conditions nécessaires et poser les hypothèses du test.
2. Mettre en oeuvre ce test, et conclure pour un seuil de 5%.
3. Donner un encadrement de la p_{valeur} et la valeur exacte, puis conclure sur l'antivirus qui utiliserait le moins de mémoire pendant l'analyse.

Exercice 9.5

On reprend l'étude précédente avec les deux antivirus A et B. On prend cette fois-ci douze dossiers contenant des chevaux de Troie, virus et spywares que l'on analyse chacun par les deux antivirus. On s'intéresse alors au nombre de menaces détectées pour chacun. On suppose que la différence du nombre de menaces peut-être modélisée par une variable de loi Gaussienne.

Antivirus A	65	80	89	64	68	68	86	54	91	77	77	86
Antivirus B	53	63	62	52	64	50	75	35	72	59	63	55

Quel antivirus choisiriez-vous pour une protection optimale ?

Justifier le choix de la méthode appropriée, donner les conditions nécessaires, poser les hypothèses du test, mettre en oeuvre ce test au seuil de 5%, calculer la p_{valeur} et conclure cette étude.

FICHE - 10

Exercice 10.1

Dans une étude sur une certaine variable X, on a mesuré cette variable sur deux échantillons différents mais constitué chacun de 11 individus. On considère que X suit une loi normale, de moyenne μ et de variance σ^2.

X_1	3,28	2,94	3,5	3,07	3,04	3,12	2,85	3,05	3,32	2,89	3,11
X_2	1,65	1,5	1,18	2,19	1,11	1,48	1,75	2,07	1,44	1,85	2,21

On veut vérifier qu'en moyenne la variable X ***diminue*** dans le second échantillon.

1. Préciser et justifier la méthode appropriée qui permet de répondre à ce problème.
2. (a) Faire le test de comparaison des variances au seuil de α= 1%. Vous préciserez la statistique de ce test et sa loi, puis vous conclurez.
   (b) Montrer que la p_{valeur} de ce test est $0,048$. Discuter selon les valeurs possibles de α. Que concluez-vous pour α= 1% ? [Vérifier la cohérence avec (a)].
3. Posez les hypothèses du test des moyennes qui répond au problème. Précisez la statistique de ce test et sa loi. Calculez un encadrement de la p_{valeur}. Conclure votre test pour un risque α= 1%, et donner votre conclusion de façon explicite.

Solution :

1. *On note X_1 resp. X_2 les variables correspondantes aux deux échantillons différents, il s'agit donc d'échantillons indépendants. Les variables X_1 et X_2 suivent des lois normales. On veut comparer les moyennes des deux échantillons. On veut savoir si X diminue, donc ($\mu_1 > \mu_2$). Il faut donc faire un test de comparaison des moyennes de Student de deux échantillons indépendants (test unilatéral). Condition nécessaire à ce test : les variances des deux échantillons doivent être égales. On doit donc le vérifier au préalable par un test de Fisher.*
2. *(a) Le test : $\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2$ contre $\mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2$. La statistique : $T = \dfrac{S_2'^2}{S_1'^2} \rightsquigarrow \mathcal{F}_{(n_2-1;n_1-1)}$*
   Calculs : $\overline{x}_1 = 3,106$ $\overline{x}_2 = 1,675$ s_1' =0,195 $s_2' = 0,378$. La valeur prise par T : 3,758.
   La règle de décision : Rejet de $\mathcal{H}_0 \Longleftrightarrow T > f_{(1-\alpha/2)}^{(n_2-1;n_1-1)}$. Ici, $f_{(0,995)}^{(10;10)} = 5,85$. On ne rejette donc pas $\mathcal{H}_0$, on conclut que les variances sont égales avec une probabilité de 99%.
   (b) Pour ce test : $p_{valeur} = 2 * Min\left\{P_{\mathcal{H}_0}(T < T_{calc})\ ;\ P_{\mathcal{H}_0}(T > T_{calc})\right\} = 2 * P_{\mathcal{H}_0}(T > 3,758)$
   $p_{valeur} = 2 * [1 - P_{\mathcal{H}_0}(T < 3,758)] \simeq 2 * (0,024) \simeq 0,048$ (calcul fait avec calculatrice ou ordinateur). Alors, pour tout risque $\alpha > 0,048$ on conclut que les variances sont différentes, pour $\alpha < 0,048$ on conclut que les variances sont égales. Donc, pour α= 1%, on accepte l'égalité des variances, ce qui est cohérent avec (a).
3. *On pose le test : $\mathcal{H}_0 : \mu_1 = \mu_2$ contre $\mathcal{H}_1 : \mu_1 > \mu_2$. On est dans le cas : $\sigma_1^2 = \sigma_2^2$, avec de petits échantillons :*
   la statistique du test est : $T = \sqrt{\dfrac{n_1+n_2-2}{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \rightsquigarrow \mathcal{T}_{n_1+n_2-2}$.
   La règle de décision est : Rejet de $H_0 \Leftrightarrow T > t_{(1-\alpha)}^{(n_1+n_2-2)}$. La valeur prise par la statistique est 11,15.
   Ici, $p_{valeur} = P(T > 11,15) = 1 - P(T < 11,15)$. La table $\mathcal{T}_{20}$ donne $P(T < 11,15) \gg 0,9995$, donc $p_{valeur} \ll 0,0005$ (avec calculatrice ou ordinateur, $p_{valeur} \simeq 2\,10^{-10}$). En conséquence, pour tout risque $\alpha > p_{val}$ (en particulier pour $\alpha = 1\%$), on accepte $\mathcal{H}_1$. On conclut donc, avec un risque quasiment nul de se tromper, qu'il y a bien diminution de X en moyenne dans le second échantillon.
   Remarque : Compte tenu des questions 2.b), et 3) on peut même prendre des risques α jusqu'à $0,048$ pour conclure qu'il y a bien diminution de X.

Exercice 10.2

Dans une étude sur la durée de vie X d'un composant électronique, on mesure cette variable sur 10 composants "témoins", c'est-à-dire qui n'ont eu aucun traitement. Un deuxième échantillon de 10 composants ont subi un traitement dans l'espoir de voir une augmentation de la durée de vie. les mesures sont les suivantes :

Témoins : $n_1 = 10$; $\overline{x}_1 = 9,32$; $\widehat{\sigma}_1^2 = 0,91$

Traités : $n_2 = 10$; $\overline{x}_2 = 10,56$; $\widehat{\sigma}_2^2 = 1,10$

1. On veut vérifier que la durée de vie des composants est plus variable avec le traitement.
   a) Poser les conditions nécessaires sur les variables pour faire cette étude.
   b) Poser les hypothèses du test qui répond à ce problème. Donner la statistique du test et sa loi. Mettre en oeuvre ce test pour un seuil de 2,5%.
   c) Quels risques faut-il prendre pour répondre par l'affirmative à la question ?
2. On veut maintenant vérifier que la durée de vie moyenne des composants est prolongée par le traitement.
   a) Poser les hypothèses du test qui répond à ce problème (justifier). Donner la statistique du test et sa loi. Mettre en oeuvre ce test, calculez la p_{valeur} par encadrement, puis la valeur exacte.
   b) Quels risques permettent de répondre par l'affirmative à la question (vérifier la cohérence avec la question1).

Exercice 10.3

On étudie le DAS, indice de débit d'absorption spécifique, indice indiquant la quantité d'énergie véhiculée par les radiofréquences émises vers l'usager par un appareil radioélectrique (téléphone portable, par exemple). Plus le DAS d'un appareil radioélectrique est faible, moins cet appareil est dangereux pour la santé. Cet indice peut être modélisé par une variable aléatoire X de loi Gaussienne. L'unité est le watt par kilogramme (W/kg). On procède à neuf mesures pour deux "types" de téléphones portables différents. On donne les moyennes et variances empiriques des deux échantillons :

$\overline{x} = 1,707 \quad \overline{y} = 1,685 \quad s_x^2 = 0,04329 \quad s_y^2 = 0,01827.$

1. Préciser toutes les hypothèses de modélisation permettant de décider si les deux types de téléphones ont des DAS en moyenne significativement différents.
2. Montrer que selon ces hypothèses, la statistique $T = \sqrt{\dfrac{n_1+n_2-2}{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\dfrac{\overline{X_1}-\overline{X_2}}{\sqrt{n_1S_1^2+n_2S_2^2}}$ suit une loi de Student.
3. Tester au seuil de 5% l'hypothèse d'égalité des variances.
4. Construire le test qui permet de décider si les deux types de téléphones ont des DAS en moyenne significativement différents. Donner un encadrement de la p_{valeur} pour ce test, puis la valeur exacte. Quelle est votre conclusion au seuil de 5%?

Exercice 10.4

Une compagnie d'assurance sur la vie étudie la proportion de clients décédés dans les 10 années suivant la souscription d'un contrat. Il existe deux types de contrat : le premier C_1 avec une prime fixe, le second C_2 avec une prime variable en fonction de l'état de santé du malade (dans ce cas, la prime est plus élevée pour un client s'il est mauvaise santé).

On prélève un échantillon de 979 assurés pour C_1, on trouve 48 décès.

On prélève un échantillon de 140 assurés pour C_2, on trouve 13 décès.

La compagnie d'assurance souhaite savoir si les proportions de décès correspondants aux types de contrats sont significativement différentes, au seuil de 5%.

Justifier le test à mettre en oeuvre, préciser la statistique de ce test ainsi que sa loi, calculer la p_{valeur} de ce test, puis conclure.

FICHE - 11

Exercice 11.1

On étudie deux logiciels qui permettent de "nettoyer" des PC. Un échantillon de 200 ordinateurs sont passés par un de ces deux logiciels L1 et L2. On a noté les résultats de ces analyses : nettoyage efficace, ou moyennement efficace, ou peu efficace.

On désire savoir si les deux logiciels ont la même efficacité, donc si le résultat dépend du logiciel choisi.

Résultat / Logiciel	Peu efficace	Efficacité moyenne	Très efficace
L1	18	18	44
L2	22	42	56

1. Quel test permet de répondre au problème posé (justifier). Poser les hypothèses de ce test.
2. Construire le tableau croisant ces deux variables sous condition d'indépendance.
3. Mettre en oeuvre ce test (calculer la valeur prise par la statistique sur ces données). Conclure pour un risque de première espèce de 5%. Donner une conclusion explicite.
4. Donner un encadrement de la p_{valeur} de ce test. Conclure selon le risque encouru.

Solution :

1. *Les données sont sous forme d'un tableau de contingence (car effectifs croisant deux variables). On cherche une dépendance ou indépendance entre les 2 variables, donc test du Khi-deux d'indépendance. Les hypothèses du test sont : $\mathcal{H}_0$: ' Indépendance des 2 variables ' contre $\mathcal{H}_1$: ' Non indépendance des 2 variables '.*
2. | | *Peu efficace* | *Efficacité moyenne* | *Très efficace* |
   |---|---|---|---|
   | *L1* | *16* | *24* | *40* |
   | *L2* | *24* | *36* | *60* |

   Tableau obtenu en calculant tous les $\left(\frac{n_{i\bullet}n_{\bullet j}}{n}\right)$.
3. *La statistique du test :* $T = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} = n\left(\sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 1\right)$. *Ici, la valeur prise par la statistique est de* $3,5834$. *La statistique suit une loi* $\mathcal{X}_2^2$. *La règle de décision : Rejet de* $\mathcal{H}_0$ *ssi :* $T > z_{1-\alpha}^{(2)}$ *et* $z_{0,95}^{(2)} = 5,991$. *On ne peut donc pas rejeter* $\mathcal{H}_0$. *On conclut avec une probabilité de 95% l'indépendance entre le nettoyage et le logiciel utilisé.*
4. *Ici,* $p_{valeur} = P(T > 3,5834) = 1 - P(T < 3,5834)$. *La table* $\mathcal{X}_2^2$ *donne* $0,5 < P(T < 3,5834) < 0,9$. *On déduit que* $0,1 < p_{valeur} < 0,5$. *En conséquence, pour des seuils* $\alpha < p_{valeur}$, *donc inférieurs à 10%, on conclura l'indépendance entre le traitement et le logiciel.*

Exercice 11.2

Le logiciel R est un langage interprété. On peut se rendre compte (vérifiez-le en TP) que R est un peu lent lorsqu'on implémente explicitement des boucles (ex. la boucle for). Il faut donc les éviter autant que possible. Cependant, un nouveau package « compiler » inclus dans la distribution standard de R 2.14, permet d'optimiser les temps de compilations de boucles.

On a donc programmé une ACP (méthode d'analyse des données, statistique, complexe), sous R, avec trois traitements différents : le premier avec l'utilisation de boucles, le second avec boucles mais en utilisant la fonction « cmpfun » du package "compiler", le troisième sans boucle.

On a analysé 226 gros jeux de données (tous de tailles comparables) par l'un de ces trois traitements. Si le temps d'exécution est supérieure à 30 secondes, on considère que la compilation est "lente", sinon elle sera considérée comme "performante". On a relevé les résultats suivants :

Exécution	Performante	Lente
P1	30	30
P2	42	35
P3	58	31

On souhaite déterminer si le temps d'exécution dépend du programme utilisé, avec un risque de première espèce de 5%.

1. Quel test permet de répondre à la question posée ? Justifier. Calculer la valeur prise par la statistique de ce test.
2. Donner un encadrement de la p_{valeur}, puis la valeur exacte. Conclure

Exercice 11.3

Pendant un an on a observé 240 individus qui ont acheté un ordinateur portable ou une tablette tactile. Parmi ceux-ci : 110 ont acheté une tablette, 25 ont acheté un ordinateur portable et ont eu un problème, 78 ont acheté une tablette et n'ont eu aucun problème.

1. Construire le tableau de contingence de ces données. Calculer la valeur prise par la statistique de ce test.
2. Donner un encadrement de la p_{valeur}, puis la valeur exacte. Peut-on dire que les problèmes rencontrés sont indépendants du support informatique acheté ?

Exercice 11.4

On dispose d'un dé à six faces. On effectue une série de 200 lancers afin d'étudier si ce dé est parfaitement équilibré. On obtient les résultats suivants :

Numéro	1	2	3	4	5	6
Effectif	38	30	30	36	37	29

Le dé est-il parfaitement équilibré ou pas ? Sur quelle hypothèse pariez-vous ? (Autrement dit, est-on en présence d'une loi uniforme discrète ou pas ?)

Exercice 11.5 *Armez-vous de vos calculatrices !!!*

Soit X la variable aléatoire modélisant le temps de téléchargement d'un certain film. Celui-ci dépend de plusieurs facteurs dont l'ordinateur utilisé, la connexion, etc. Sur un échantillon de 90 ordinateurs, on relève ce temps de téléchargement (exprimé en mn). Le tableau suivant donne la répartition par classe :

Temps	[12 ;16[	[16 ;20[	[20 ;24[	[24 ;28[	[28 ;32[	[32 ;36[	[36 ;40[	[40 ;44[
Nombre d'ordinateurs observés	5	11	16	21	15	12	8	2

On veut savoir si le temps de téléchargement est distribuée normalement, au seuil de 5%.

1. Justifier la méthode utilisée pour répondre à la question. Estimer la moyenne et la variance de cet échantillon. Poser les hypothèses du test.
2. Calculer les effectifs théoriques de chaque classe (selon la loi définie en 1.)
3. Regrouper les classes de trop faibles effectifs. Donner la statistique du test. Déterminer le degré de liberté de la loi du Khi-deux de la statistique.
4. Finir le test et conclure au seuil de 1%.
5. Quelle autre méthode vue en début d'année permet également de vérifier la normalité d'une variable ?

Université Grenoble Alpes
L2 Licence Sciences et Technologies - STA401/2016-17

Formulaire

Carole Durand

Ce fascicule de formules sera complété tout au long des cours d'Amphi.

---

1. Données et Modèles

Calculs statistiques :

Un échantillon de données $(x_1, ..., x_n)$ est une série de valeurs prises par une variable X sur n individus.

La moyenne empirique : $\overline{x} = \frac{1}{n}\sum x_i$

La variance empirique : $s_x^2 = \frac{1}{n}\sum(x_i - \overline{x})^2 = \frac{1}{n}(\sum .....) - \overline{x}^2$

L'écart-type empirique : s_x.

La fréquence empirique : $f = \frac{k}{n}$ k est le nombre de réalisations d'un évènement parmi les n individus

Un échantillon centré réduit a une moyenne de 0 et une variance de 1.

La médiane est la plus petite valeur prise par X telle qu'au moins la moitié des effectifs soit inférieur.

Le premier quartile est la plus petite valeur prise par X telle qu'au moins le quart des effectifs soit inférieur.

Le 3ème quartile est la plus petite valeur prise par X telle qu'au moins les 3/4 des effectifs soit inférieur.

Probabilités :

La probabilité d'un évènement est la proportion d'individus qui réalisent cet évènement (dans la population).

$P[\Omega] = 1$; $P[\emptyset] = 0$; $0 \leq P[A] \leq 1$; $P[\overline{A}] = 1 - P[A]$; $P[A \cup B] = P[A] + ...$

La probabilité conditionnelle de A sachant B est $P[A \mid B] = \frac{.....}{P[B]}$

Deux évènements A et B sont indépendants ssi : $P[A \mid B] =$..... ou $P[A \cap B] =$.....

En notant $\overline{B}$ l'évènement contraire de B, on a : $P[A] = P[A \mid B]P[B] + P[A \mid \overline{B}]P[\overline{B}]$

Formule de Bayes : $P[A \mid B] = \frac{.....}{P[B]} = \frac{.....}{P[B \mid A]P[A] + P[B \mid \overline{A}]P[\overline{A}]}$

Lois :

X suit la loi Bernoulli $\mathcal{B}(p)$: $P[X = 1] = p$; $P[X = 0] = 1 - p$; $E[X] = p$ *et* $Var[X] =$ .....

X suit la loi Binomiale $\mathcal{B}(n, p)$ *: [n expériences indépendantes ou tirages avec remise]*

$P[X = k] = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}$ *avec* $E[X] =$ ..... *et* $Var[X] =$ .....

X suit la loi Hypergéométrique $\mathcal{H}(N, m, n)$ *: [n tirages sans remise]*

$$P[X = k] = \frac{\begin{pmatrix} m \\ k \end{pmatrix}\begin{pmatrix} N-m \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad \textit{avec} \quad E[X] = ... \quad \textit{et} \quad Var[X] = n\left(\frac{m}{N}\right)\left(\frac{N-m}{N}\right)\left(\frac{N-n}{N-1}\right)$$

X suit la loi de Poisson $\mathcal{P}(\lambda)$: $P[X=k]=\dfrac{\lambda^k e^{-\lambda}}{k\,!}$ *avec* $E[X]=$ ..... *et* $Var[X]=$ .....

Pour les lois continues, on lit sur les tables | le quantile x_p d'ordre p : $P[X\leq x_p]=F(x_p)=$ ..... |

La loi Normale $\mathcal{N}(0,1)$, la loi de Student $\mathcal{T}_n$ de paramètre n, la loi du Khi-deux $\mathcal{X}^2_n$ de paramètre n, la loi de Fisher-Snédécor $\mathcal{F}_{(n_1,n_2)}$ de paramètres (n_1,n_2).

Notations : u_p est le quantile d'ordre p de $\mathcal{N}(0,1)$; t^n_p est le quantile d'ordre p de $\mathcal{T}_n$; z^n_p est le quantile d'ordre p de $\mathcal{X}^2_n$; $f^{n_1;n_2}_p$ est le quantile d'ordre p de $\mathcal{F}_{(n_1,n_2)}$.

Si $X \rightsquigarrow \mathcal{N}(\mu,\sigma^2)$, alors : $Y=(X-\mu)/\sigma \rightsquigarrow$ .....

Si $X \rightsquigarrow \mathcal{N}(\mu,\sigma^2)$, alors : $Y=\sqrt{n}(\overline{X}-\mu)/\sigma \rightsquigarrow$ .....

T.C.L. (conditions) : la loi $\mathcal{B}(n,p)$ peut être appochée par la loi.....

Autre approximation (conditions) : la loi $\mathcal{B}(n,p)$ peut être appochée par la loi $\mathcal{P}(\lambda)$ avec $\lambda=$ .....

Autre approximation (conditions) : la loi $\mathcal{P}(\lambda)$ peut être appochée par la loi .....

Inégalité de Markov : X v.a. positive, $a>0$, $P(X\geq a)\leq \dfrac{E(X)}{.....}$

Inégalité de Bienaymé Tchebychev : X v.a. positive, $a>0$, $P(\mid X-E(X)\mid \geq a)\leq \dfrac{.....}{a^2}$

Intervalle de fluctuation de p pour une loi $\mathcal{B}(n,p)$ **avec un niveau de confiance** $1-\alpha$:

$I(p,\alpha)=\left[..... - u_{1-\alpha/2}\,\dfrac{.....}{\sqrt{n}}\,;\ ..... + u_{1-\alpha/2}\,\dfrac{.....}{\sqrt{n}}\right]$ où $u_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{N}(0,1)$

Intervalle de fluctuation de μ **pour une loi** $\mathcal{N}(\mu,\sigma^2)$ **avec un niveau de confiance** $1-\alpha$:

$I(\mu,\alpha)=\left[..... - u_{1-\alpha/2}\,\dfrac{.....}{\sqrt{n}}\,;\ ..... + u_{1-\alpha/2}\,\dfrac{.....}{\sqrt{n}}\right]$ où $u_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{N}(0,1)$

---

2. Estimation Statistique

Estimateurs ponctuels :

Un estimateur d'un paramètre est une fonction de $(X_1,...,X_n)$ qui 'approche' ce paramètre.

Un 'bon' estimateur est ..... , ... et .....

Estimateur du maximum de vraisemblance .....

Une estimation est la valeur de l'estimateur prise sur un échantillon de données.

Notation :

$\widehat{p}=F$	$\widehat{\mu}=\overline{X}=\frac{1}{n}\sum X_i$	$\widehat{\sigma^2}=S'^2=\frac{n}{n-1}S^2=\frac{1}{n-1}\sum$ .....

Si $(X_1,...,X_n)$ est un échantillon gaussien de loi $\mathcal{N}(\mu,\sigma^2)$ alors :

$\overline{X} \rightsquigarrow \mathcal{N}(\mu,\dfrac{\sigma^2}{n})$ $\qquad$ $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow$ ..... $\qquad$ $\dfrac{\overline{X}-\mu}{S/\sqrt{n-1}} \rightsquigarrow$ ..... $\qquad$ $\dfrac{nS^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2_{n-1}$

Pour n grand , $\dfrac{F-p}{\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1)$

Intervalles de confiance pour un niveau de confiance $1-\alpha$:

$I(\mu,\alpha)=\left[\overline{X}-u_{1-\alpha/2}\ \frac{.....}{\sqrt{n}}\,;\ \overline{X}+u_{1-\alpha/2}\ \frac{.....}{\sqrt{n}}\right]$ si σ^2 est connue ; $u_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{N}(0,1)$

$I(\mu,\alpha)=\left[\overline{X}-t^{n-1}_{1-\alpha/2}\ \frac{....}{\sqrt{n}}\,;\ \overline{X}+t^{n-1}_{1-\alpha/2}\ \frac{....}{\sqrt{n}}\right]$ si σ^2 est inconnue ; $t^{n-1}_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ .....

$I(\sigma^2,\alpha)=\left[\frac{.....}{z^{n-1}_{1-\alpha/2}};\ \frac{.....}{z^{n-1}_{\alpha/2}}\right]$ si σ^2 et μ sont inconnues ; $z^{n-1}_{1-\alpha/2}$ est le quantile d'ordre .....

Cas de grands échantillons :

$I(\mu,\alpha)=\left[\overline{X}-u_{1-\alpha/2}\ \frac{....}{\sqrt{n}}\,;\overline{X}+u_{1-\alpha/2}\ \frac{....}{\sqrt{n}}\right]$ où $u_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{N}(0,1)$

$I(p,\alpha)=\left[F-u_{1-\alpha/2}\ \frac{.....}{\sqrt{n}}\,;\ F+u_{1-\alpha/2}\ \frac{.....}{\sqrt{n}}\right]$

---

3. Tests Statistiques

Le principe :

- Désicion à prendre entre deux hypothèses incompatibles : $\mathcal{H}_0$ contre $\mathcal{H}_1$
- Risque de première espèce ou seuil du test : $\alpha = P[.....] = P_{\mathcal{H}_0}[.....]$
- Risque de deuxième espèce : $\beta = P[.....]$. La puissance du test est $1-\beta$
- Règle de décision : $\{$ *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T \in W$ $\quad$ (*la statistique du test* T *est dans la région critique*)$\}$

Test unilatéral inférieur	Test unilatéral supérieur	Test bilatéral
Schémas		

- La p_{valeur} est le seuil limite à partir duquel $\mathcal{H}_0$ est rejetée ($\mathcal{H}_1$ est acceptée) compte tenu des données observées.

Si on connait la p_{valeur}, alors $\{$ *Rejet de* $\mathcal{H}_0 \Longleftrightarrow$ .....$\}$

Schéma

Tests de paramètres : X suit une loi $\mathcal{N}(\mu, \sigma^2)$

$\mathcal{H}_0$: $\mu = \mu_0$ et σ^2 connue, alors $T = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$

$\mathcal{H}_0$: $\mu = \mu_0$ et σ^2 inconnue, alors $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \rightsquigarrow$ .....

$\mathcal{H}_0$: $\mu = \mu_0$ et n grand, alors $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \rightsquigarrow$ .....

$\mathcal{H}_0$: $\sigma^2 = \sigma_0^2$ et σ^2 et μ inconnues, alors $T = \dfrac{nS^2}{\sigma_0^2} \rightsquigarrow$ .....

$\mathcal{H}_0$: $p = p_0$ et n grand, alors $T = \dfrac{F - p_0}{\dfrac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}} \rightsquigarrow$ .....

$$\begin{cases} \mathcal{H}_0 : \mu = \mu_0 \\ \mathcal{H}_1 : \mu < \mu_0 \end{cases} \quad \begin{array}{ll} Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < -u_{1-\alpha} & si\ \sigma^2\ connue \\ Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < -t_{1-\alpha}^{n-1} & si\ \sigma^2\ inconnue \end{array}$$

$$\begin{cases} \mathcal{H}_0 : \mu = \mu_0 \\ \mathcal{H}_1 : \mu > \mu_0 \end{cases} \quad \begin{array}{ll} Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T > u_{1-\alpha} & si\ \sigma^2\ connue \\ Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T > t_{1-\alpha}^{n-1} & si\ \sigma^2\ inconnue \end{array}$$

$$\begin{cases} \mathcal{H}_0 : \mu = \mu_0 \\ \mathcal{H}_1 : \mu \neq \mu_0 \end{cases} \quad \begin{array}{ll} Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < -u_{(1-\alpha/2)}\ ou\ T > u_{(1-\alpha/2)} & si\ \sigma^2\ connue \\ Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < -t_{(1-\alpha/2)}^{n-1}\ ou\ T > t_{(1-\alpha/2)}^{n-1} & si\ \sigma^2\ inconnue \end{array}$$

$$\begin{cases} \mathcal{H}_0 : \sigma^2 = \sigma_0^2 \\ \mathcal{H}_1 : \sigma^2 < \sigma_0^2 \end{cases} \quad Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < .....$$

$$\begin{cases} \mathcal{H}_0 : \sigma^2 = \sigma_0^2 \\ \mathcal{H}_1 : \sigma^2 > \sigma_0^2 \end{cases} \quad Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T > .....$$

$$\begin{cases} \mathcal{H}_0 : \sigma^2 = \sigma_0^2 \\ \mathcal{H}_1 : \sigma^2 \neq \sigma_0^2 \end{cases} \quad Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < z_{\alpha/2}^{n-1} \quad ou \quad T > z_{1-\alpha/2}^{n-1}$$

$$\begin{cases} \mathcal{H}_0 : p = p_0 \\ \mathcal{H}_1 : p < p_0 \end{cases} \quad Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < ..... \quad si\ n\ grand$$

$$\begin{cases} \mathcal{H}_0 : p = p_0 \\ \mathcal{H}_1 : p > p_0 \end{cases} \quad Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T > ..... \quad si\ n\ grand$$

$$\begin{cases} \mathcal{H}_0 : p = p_0 \\ \mathcal{H}_1 : p \neq p_0 \end{cases} \quad Rejet\ de\ \mathcal{H}_0 \Longleftrightarrow T < -u_{(1-\alpha/2)} \quad ou \quad T > u_{(1-\alpha/2)} \quad si\ n\ grand$$

Calcul de la p_{valeur} :

- Test unilatéral inférieur : $p_{valeur} = P_{\mathcal{H}_0}(T < .....)$
- Test unilatéral supérieur : $p_{valeur} = P_{\mathcal{H}_0}(T > .....)$
- Test bilatéral (lois Normale et Student) : $p_{valeur} = 2 * [1 - P_{\mathcal{H}_0}(T < .....)]$
- Test bilatéral (lois Khi-deux et Fisher) : $p_{valeur} = 2 * Min\,\{P_{\mathcal{H}_0}(T < T_{calc})\,;\ .....\}$

Test de comparaison de deux échantillons appariés :

Soient $(x_1, ..., x_n)$ et $(x'_1, ..., x'_n)$ deux séries de données mesurées sur le même échantillon d'individus et sur la même variable X. Soient $D_i = X_i - X'_i$ toutes les différences. On suppose que $D \rightsquigarrow \mathcal{N}(\mu_d, \sigma_d^2)$.

$\begin{cases} \mathcal{H}_0 : \mu_d = 0 \\ \mathcal{H}_1 : \mu_d \neq 0 \end{cases}$ La statistique est $T = \dfrac{\overline{D}\sqrt{n-1}}{S_d} \rightsquigarrow \mathcal{T}_{n-1}$

Rejet de $\mathcal{H}_0 \Longleftrightarrow T <$ ..... *ou* $T >$ .....

Pour un test unilatéral supérieur (si inférieur, échanger X et X') : *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T >$ .....

Test de comparaison de deux échantillons indépendants :

Soient $(X_1, ..., X_{n_1})$ un premier échantillon de loi $\mathcal{N}(\mu_1, \sigma_1^2)$, et $(Y_1, ..., Y_{n_2})$ un second échantillon de loi $\mathcal{N}(\mu_2, \sigma_2^2)$.

On suppose les deux échantillons indépendants.

1. $\begin{cases} \mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 \\ \mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \Longleftrightarrow \begin{cases} \mathcal{H}_0 : \sigma_1^2/\sigma_2^2 = 1 \\ \mathcal{H}_1 : \sigma_1^2/\sigma_2^2 \neq 1 \end{cases}$; $T = \dfrac{S_1'^2}{S_2'^2} \rightsquigarrow$ .....

   Il faut que T>1. Si T<1 échangez les rôles de X et Y (Attention aux degrés de liberté de la loi de Fisher).

   Rejet de $\mathcal{H}_0 \Longleftrightarrow T > f_{(1-\alpha/2)}^{(n_1-1;n_2-1)}$

   Pour un test unilatéral supérieur : *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T >$ .....

2. $\begin{cases} \mathcal{H}_0 : \mu_1 = \mu_2 \\ \mathcal{H}_1 : \mu_1 \neq \mu_2 \end{cases}$ *si* $\sigma_1^2 = \sigma_2^2$ $\quad T = \sqrt{\dfrac{n_1+n_2-2}{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \dfrac{.....}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \rightsquigarrow$ .....

   Rejet de $\mathcal{H}_0 \Longleftrightarrow T < -t_{(1-\alpha/2)}^{(n_1+n_2-2)}$ *ou* $T > t_{(1-\alpha/2)}^{(n_1+n_2-2)}$

   Pour un test unilatéral supérieur : *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T >$ .....

3. Tests des moyennes dans le cas de grands échantillons : $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \rightsquigarrow$ .....

   Test bilatéral : *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T < -u_{(1-\alpha/2)}$ *ou* $T > u_{(1-\alpha/2)}$
   Test unilatéral supérieur : *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T >$.....

Test de comparaison de deux proportions :

Soient p_1 et p_2 les proportions d'un même caractère mesuré sur deux populations de taille n_1 et n_2 (grands).

$\begin{cases} \mathcal{H}_0 : \ p_1 = p_2 \\ \mathcal{H}_1 : \ p_1 \neq p_2 \end{cases}$ La statistique est $T = \dfrac{F_1 - F_2}{\sqrt{F(1-F)(\dfrac{1}{n_1}+\dfrac{1}{n_2})}} \rightsquigarrow$ .....

avec $F =$ .....

Rejet de $\mathcal{H}_0 \Longleftrightarrow T <$ ..... *ou* $T >$ .....

Test unilatéral supérieur (sinon inverser les rôles des échantillons) : *Rejet de* $\mathcal{H}_0 \Longleftrightarrow T >$.....

Test du Khi-deux d'ajustement (adéquation) :

X	$c_1 \cdots\cdots c_r$	total
Données	$n_1 \cdots\cdots n_r$	n
Loi P	$np_1 \cdots\cdots np_r$	n

n est la taille de l'échantillon
n_i est l'effectif observé de la classe c_i
p_i est la probabilité théorique de la loi P pour la classe c_i

$\begin{cases} \mathcal{H}_0 : X \text{ suit la loi } P \\ \mathcal{H}_1 : X \text{ ne suit pas la loi } P \end{cases}$ la statistique $T = \sum_{i=1\ldots r} \frac{(n_i - np_i)^2}{np_i} \rightsquigarrow$ .....

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

Conditions : les effectifs théoriques doivent être supérieurs à 5, et n>50.

S'il est nécessaire d'estimer k paramètres pour déterminer la loi P, alors : $T \rightsquigarrow$ .....

Test du Khi-deux d'indépendance (contingence) :

X Y	y_1 $\cdots\cdots$ y_q	marge ligne
x_1	n_{11} $\cdots$ n_{1q}	$n_{1\bullet}$
$\vdots$	$\vdots$ $\quad$ $\vdots$	$\vdots$
$\vdots$	$\vdots$ $\quad$ $\vdots$	$\vdots$
x_p	n_{p1} $\ldots$ n_{pq}	$n_{p\bullet}$
marge colonne	$n_{\bullet 1}$ $\cdots\cdots$ $n_{\bullet q}$	n

n est la taille de l'échantillon
n_{ij} est l'effectif croisé entre les modalités X_i et Y_j
$n_{i\bullet}=\sum_j n_{ij}$ $\qquad$ $n_{\bullet j}=\sum_i n_{ij}$
$\frac{n_{ij}}{n_{i\bullet}}$ et $\frac{n_{ij}}{n_{\bullet j}}$ sont les profil-lignes (resp. profil-colonnes)

$\begin{cases} \mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes} \\ \mathcal{H}_1 : X \text{ et } Y \text{ non indépendantes} \end{cases}$

La statistique est $T = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{.....} = n \left(\sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - ..... \right) \rightsquigarrow .....$

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

Conditions : les effectifs théoriques doivent être supérieurs à 5, et n>50.