STATISTIQUE MATHÉMATIQUE - TRAVAUX DIRIGÉS : ÉNONCÉS DES EXERCICES.

ENSAE - Mastères spécialisés - 2025/2026

ENSAI École nationale de la statistique et de l'analyse de l'information

1 Modèles statistiques et estimateurs

1.1 Révisions sur les lois et quelques calculs classiques.

1°. Soit X une variable aléatoire de loi exponentielle de paramètre λ. Déterminer la loi de $Y = [X]$, partie entière de X.

2°. Soit $X_1, ..., X_n$, n v.a.i.i.d. de loi exponentielle de paramètre λ et $M_n = \text{Min}(X_1, ..., X_n)$. Déterminer la loi de M_n.

3°. Déterminer la loi du minimum de deux lois uniformes sur $[0,1]$, indépendantes.

4°. Soit X une variable aléatoire suivant une loi de Laplace. Déterminer la loi de $Y = |X|$ et calculer $\mathbb{P}[X \geq 2]$, $\mathbb{P}[X < 0]$, $\mathbb{P}[X = 0]$.

5°. Soit X une variable aléatoire de loi uniforme sur $[-\pi/2, \pi/2]$. Déterminer la loi de $Y = \tan X$.

1.2 Moyenne et variance de la moyenne et de la variance empirique d'un échantillon.

On considère $X_1, ..., X_n$, n v.a.i.i.d. et l'on note $\overline{X}$ leur moyenne empirique et S^2 leur variance empirique :

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

1°. Déterminer la moyenne et la variance de $\overline{X}$.

2°. Déterminer la moyenne et la variance de S^2.

1.3 Calculs sur un échantillon uniforme

Soit X une variable aléatoire de loi uniforme sur $[0, 2a]$ et $X_1 ..., X_n$ un n échantillon de X. Soit $\overline{X}$ la moyenne empirique des X_i et M le maximum.

1°. Calculer $\mathbb{E}[X]$ et $\mathbb{V}(X)$.

2°. Calculer $\mathbb{E}[\overline{X}]$ et $\mathbb{V}(\overline{X})$.

3°. Calculer $\mathbb{E}[M]$ et $\mathbb{V}(M)$.

4°. Comparer les résultats des questions 2° et 3°.

1.4 Moyenne et variance empirique d'un échantillon gaussien.

On considère le modèle d'échantillonnage gaussien $X_1, ..., X_n$ de n v.a.i.i.d. de loi $\mathcal{N}(m, \sigma^2)$ et l'on note $\overline{X}$ la moyenne empirique et S^2 la variance empirique :

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

1°. Parmi les variables aléatoires suivantes, lesquelles sont des statistiques ?

$\overline{X}$? S^2 ? nS^2/σ^2 ? $T = \sqrt{n-1}(\overline{X} - m)/S$?

2°. Déterminer le modèle image par la moyenne empirique.

3°. Déterminer le modèle image par la statistique $(\overline{X}, S^2)$ en opérant de la façon suivante : écrivez S^2 comme fonction de $Y_i = X_i - \overline{X}$ puis en déduire que S^2 est indépendante de $\overline{X}$; déterminer la loi de nS^2/σ^2. Enfin, écrivez le modèle image sous la forme d'un triplet.

4°. Déterminer la loi de T.

1.5 Statistiques d'ordre.

On considère le modèle statistique d'échantillonnage où $\mathscr{P}$ est l'ensemble des lois de probabilités définies sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Soit F la fonction de répartition d'une loi $P \in \mathscr{P}$.

1°. Quel est le modèle image par la statistique d'ordre $X_{(1)} = \min_{i=1..n} X_i$? Par $X_{(n)} = \max_{i=1...n} X_i$?

2°. Soit F_n la fonction de répartition empirique du n-échantillon. Déterminer le modèle image par la statistique $nF_n(x)$.

3°. Déterminer le modèle image par la statistique d'ordre $X_{(i)}$, pour $i = 1, ..., n$.

1.6 Exemples de modèles exponentiels.

Pour les différents modèles proposés, on note X la variable des observations. Exhibez pour chacun une mesure dominante, spécifiez si les modèles sont exponentiels et dans l'affirmative, déterminez une statistique canonique, un paramètre canonique, puis calculer l'espérance et la variance de la statistique canonique.

1°. $(X_1, ..., X_n)$ est un n-échantillon gaussien.

2°. X est issu d'un modèle hypergéométrique.

3°. X est issu d'un modèle binomial négatif.

4°. $X = \epsilon + (1-\epsilon)Y$ où $\epsilon \sim b(\alpha)$ suit une loi de Bernoulli de paramètre α et $Y \sim \mathcal{N}(0,1)$ est indépendante de ϵ.

1.7 Statistiques de rang.

Soit $X = (X_1, ..., X_n)$ un n-échantillon issu d'un modèle $\mathscr{P}$ de loi sans atome sur $\mathbb{R}$. Le rang de X_i est

$$R_i(X) = \text{card}\{j : X_j \leq X_i\} \tag{3}$$

On note $R(X) = (R_1(X), ..., R_n(X))$.

1°. Faites le lien entre le rang et les statistiques d'ordre.

2°. Montrer que pour tout $i = 1, ..., n$, on a

$$\mathrm{R}_i(\mathrm{X}) = \sum_{j=1}^{n} \mathbb{1}_{[\mathrm{X}_i - \mathrm{X}_j \geq 0]} \tag{4}$$

3°. Montrer que pour tout $x \in \mathbb{R}^n$, $\mathrm{R}(x)$ est une permutation de $(1, 2, ..., n)$ $\mathbb{P}$-p.s.

4°. Quel est le modèle image par R ?

1.8 Loi uniforme : différents estimateurs

Soit $(\mathrm{X}_1, ..., \mathrm{X}_n)$ un n-échantillon de loi uniforme sur $[0, \theta]$.

1°. Déterminez un estimateur θ_1 de θ par la méthode des moments. Étudiez son biais, son erreur quadratique, sa convergence.

2°. Déterminez un estimateur θ_2 de θ par la méthode du maximum de vraisemblance. Étudiez son biais, son erreur quadratique, sa convergence. En déduire un estimateur θ_3 de θ, sans biais.

3°. On suppose que $n = 2m + 1$. Déterminez un estimateur plug-in θ_4 de θ. Étudiez son biais, son erreur quadratique, sa convergence.

4°. Quel est le meilleur des estimateurs précédents ?

5°. On muni le modèle d'une loi *a priori* de densité :

$$\pi(\theta) = \frac{1}{\theta^2} \mathbb{1}_{]1,\infty[}(\theta) \tag{5}$$

Déterminez la loi *a posteriori* du modèle et proposer un estimateur θ_5 de θ.

1.9 Loi uniforme : encore d'autres estimateurs.

On considère un échantillon $(\mathrm{X}_1, ..., \mathrm{X}_n)$ de loi uniforme sur l'intervalle $[\theta, \theta + 1]$ où θ est inconnu. On pose :

$$\widehat{\theta}_1 = \overline{\mathrm{X}}_n - \frac{1}{2} \tag{6}$$

$$\widehat{\theta}_2 = \min(\mathrm{X}_1, ..., \mathrm{X}_n) \tag{7}$$

$$\widehat{\theta}_3 = \max(\mathrm{X}_1, ..., \mathrm{X}_n) - 1 \tag{8}$$

1°. Démontrer que $\widehat{\theta}_1$ est l'estimateur des moments de θ. En déduire qu'il est sans biais ; déterminer son erreur quadratique et démontrer qu'il converge au sens des moindres carrés. Démontrer qu'il est fortement consistant, puis qu'il est asymptotiquement normal (A.N.) en précisant la loi limite et la vitesse de convergence.

2°. Démontrer que $\widehat{\theta}_2$ est un estimateur plug-in de θ.

3°. Déterminer

$$\mathbb{P}\left[n(\widehat{\theta}_2 - \theta) \leq x\right] \tag{9}$$

pour x variant dans $\mathbb{R}$ et en déduire la loi limite de $n(\widehat{\theta}_2 - \theta)$ lorsque n tend vers l'infini. Préciser la vitesse de convergence.

4°. Déterminer la fonction de répartition de $\widehat{\theta}_2$ et sa fonction de densité. En déduire son espérance et sa variance. $\widehat{\theta}_2$ converge-t-il au sens des moindres carrés ?

5°. Démontrer que $\widehat{\theta}_3$ est un estimateur plug-in de θ, déterminer

$$\mathbb{P}\left[n(\widehat{\theta}_3 - \theta) \leq x\right] \tag{10}$$

pour x variant dans $\mathbb{R}$ et en déduire la loi limite de $n(\widehat{\theta}_3 - \theta)$ lorsque n tend vers l'infini. Préciser la vitesse de convergence.

6°. Démontrer que la vraisemblance de l'échantillon est maximale sur l'intervalle $[x_{(n)} - 1; x_{(1)}]$. En déduire un estimateur du maximum de vraisemblance. Est-il unique ?

1.10 Loi de Poisson : comparaison de deux estimateurs.

Soit X une va suivant une loi de Poisson de paramètre θ. Soit $(\mathrm{X}_1, ..., \mathrm{X}_n)$ un n-échantillon de X.

1°. Déterminer deux estimateurs de θ à partir de la moyenne et de la variance de l'échantillon.

2°. Comparer ces estimateurs.

1.11 Loi de Pareto : estimation du paramètre de position par trois estimateurs.

Soit $\theta > 0$ et soit X une variable aléatoire de densité

$$f(t) = \frac{3t^2}{\theta^3} \mathbb{1}_{[0,\theta]}(t)$$

Soit $(\mathrm{X}_1, ..., \mathrm{X}_n)$ un n-échantillon de X. On pose :

$$\mathrm{S}_n = \frac{1}{n} \sum_{k=1}^{n} \mathrm{X}_k \ , \ \mathrm{T}_n = \max(\mathrm{X}_1, ..., \mathrm{X}_n)$$

$$\text{et } \mathrm{Z}_n = (\mathrm{X}_1 ... \mathrm{X}_n)^{1/n}$$

1°. Déterminer la fonction de répartition de X, puis calculer l'espérance et la variance de S_n.

2°. Déterminer un estimateur sans biais $\widehat{\mathrm{S}_n}$ de θ de la forme $\alpha \mathrm{S}_n$. Calculer sa variance et montrer qu'il converge en moyenne quadratique vers θ.

3°. Donner une densité de T_n et démonter qu'il converge en moyenne quadratique vers θ.

4°. Calculer $\mathbb{E}\left[\mathrm{X}^{1/n}\right]$, $\mathbb{E}[\mathrm{Z}_n]$ et déterminer a tel que $\lim_{n \to +\infty} \mathbb{E}[\widehat{\mathrm{Z}_n}] = \theta$ où $\widehat{\mathrm{Z}_n} = a\mathrm{Z}_n$. Calculer la variance et en déduire la convergence en moyenne quadratique.

5°. Lequel de ces trois estimateurs est-il préférable de choisir ?

1.12 Loi de Pareto : différents estimateurs.

Soit $n \geq 3$ un entier et α, β deux réels strictement positifs. Soit X la variable aléatoire réelle de densité

$$f(t) = \frac{\alpha \beta^\alpha}{t^{\alpha+1}} \mathbb{1}_{[t \geq \beta]} \tag{11}$$

On dit que X suit une loi de Paréto notée $\mathscr{P}(\alpha,\beta)$.

1°. Déterminer la fonction de répartition F de X, son espérance et sa variance.

2°. Déterminer la loi de la variable $Y = \ln(X/\beta)$.

3°. On suppose $\alpha > 2$ connu et on veut estimater β. On pose $Z_n = \min(X_1,...,X_n)$. Déterminer la loi de Z_n, son espérance, sa variance, puis déterminer un réel c_n tel que $\widehat{Z_n} = c_n Z_n$ soit un estimateur sans biais de β.

4°. Montrer que $\widehat{Z_n}$ converge en moyenne quadratique vers β.

On suppose maintenant β connu et on souhaite estimer α. on pose

$$W_n = \frac{n}{\sum_{k=1}^{n} \ln(X_k/\beta)} \tag{12}$$

5°. Déterminer la densité de $\sum_{k=1}^{n} \ln(X_k/\beta)$, calculer $\mathbb{E}[W_n]$ et en déduire que $\widehat{W}_n = \frac{n-1}{n} W_n$ est un estimateur sans biais de α.

6°. Calculer sa variance.

7°. On suppose dans cette question que $1 < \alpha < 2$. On effectue 402 observations de X et on trouve que $\widehat{W}_{402} = 1,4$. Donner un intervalle de confiance pour α au risque $0,05$.

1.13 Loi uniforme : risques de différents estimateurs.

Soit X une variable aléatoire de loi uniforme sur $[0,\theta]$ et $(X_1,...,X_n)$ un n-échantillon de X.

1°. À l'aide de l'inégalité de Markov, montrer que si T_n est un estimateur de θ asymptotiquement de risque nul, alors la suite $(T_n)_n$ converge en probabilité vers θ. En déduire une condition suffisante sur l'espérance et la variance de T_n pour que cette suite converge en probabilité vers θ.

2°. Soit S_n la moyenne empirique des X_i. Montrer que $W_n = 2S_n$ est un estimateur sans biais de θ. Déterminer sa variance et en déduire le risque quadratique. Montrer qu'il converge en probabilité.

3°. Soit $Z_n = \max(X_1,...,X_n)$. Déterminer sa fonction de répartition, sa densité et son biais en tant qu'estimateur de θ. En déduire un estimateur Y_n de θ sans biais, déterminer son risque quadratique et donner un équivalent de ce risque quand n tend vers l'infini. En déduire que Y_n converge en probabilité vers θ.

4°. On pose $I_n = \min(X_1,...,X_n)$. Déterminer sa fonction de répartition F_n. Montrer que

$$\mathbb{E}[I_n] = \theta - \int_0^{\theta} F_n(t)\,dt \tag{13}$$

Cacluler $\mathbb{E}[I_n]$, $\mathbb{E}[I_n^2]$ et en déduire $\mathbb{V}(I_n)$.

5°. On pose

$$A_n = \frac{nZ_n - I_n}{n-1} \tag{14}$$

et $D_n = Z_n - I_n$. Démontrer qu'une densité de D_n est donnée par

$$h_n(t) = \frac{n(n-1)}{\theta^n}\left[\theta t^{n-2} - t^{n-1}\right]\mathbb{1}_{[0,\theta]}(t)$$

Vérifier que A_n est un estimateur sans biais de θ, calculer $\mathbb{E}[D_n]$, calculer

$$J = \int_0^{+\infty} t^2 h_n(t)\,dt \tag{15}$$

et en déduire $\mathbb{V}(D_n)$. Calculer $\mathrm{cov}(Z_n, I_n)$, $\mathbb{V}(A_n)$ et donner un équivalent du risque quadratique de A_n.

6°. On pose, pour tout entier naturel n non nul,

$$L_n = \frac{1}{n}\sum_{k=1}^{n} \ln X_k \text{ et } T_n = \exp(L_n + 1) \tag{16}$$

On pose également $Y = \ln X$. Montrer que Y possède une espérance et la calculer. Montrer que $L_n + 1$ est un estimateur sans biais de $\ln\theta$.

7°. Montrer que

$$\mathbb{E}[T_n] = e \times \mathbb{E}\left[\prod_{k=1}^{n} X_k^{1/n}\right] \tag{17}$$

Monter que pour tout $k = 1,...,n$,

$$\mathbb{E}\left[X_k^{1/n}\right] = \frac{n}{n+1}\theta^{1/n} \tag{18}$$

et en déduire $\lim_{n\to+\infty} \mathbb{E}[T_n] = \theta$. Montrer que

$$\mathbb{E}[T_n^2] = e^2 \times \mathbb{E}\left[\prod_{k=1}^{n} X_k^{2/n}\right] \tag{19}$$

En déduire que le risque quadratique de T_n vaut

$$r_\theta(T_n) = \theta^2\left[\frac{e^2}{(1+2/n)^n} - \frac{2e}{(1+1/n)^n} + 1\right] \tag{20}$$

puis que

$$r_\theta(T_n) = \frac{\theta^2}{n} + o\left(\frac{1}{n}\right) \tag{21}$$

et déduire de ce qui précède que $(T_n)_n$ converge en probabilité vers θ.

8°. Donner pour chacun des estimateurs un équivalent de leur risque quadratique et comparer les vitesses de convergence, en probabilité, vers θ.

1.14 Loi demi-gaussienne : différents estimateurs.

Soit $X \sim \mathcal{N}(0,\sigma^2)$, où σ est un paramètre inconnu que l'on va chercher à estimer. Soit $(X_1,...,X_n)$ un n-échantillon de X. On note $Y = |X|$ et $Y_i = |X_i|$. On pose également

$$D_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } S_n^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \tag{22}$$

Soit enfin V_n l'estimateur du maximum de vraisemblance.

1°. Calculer $\mathbb{E}[Y]$ et $\mathbb{V}(Y)$.

2°. Déduire de D_n un estimateur T_n, sans biais, de σ, puis montrer que T_n converge en moyenne quadratique vers σ. Déterminer la loi limite de T_n et construire pour σ un intervalle de confiance asymptotique de niveau $1-\alpha$, avec $\alpha \in [0,1[$.

3°. Rappeler la loi de probabilité de S_n^2, calculer son espérance et sa variance. Démonter que

$$\mathbb{E}[S_n] = \sigma\sqrt{1 - \frac{1}{2n} + o\left(\frac{1}{n}\right)} \tag{23}$$

En déduire une valeur approchée de $\mathbb{V}(S_n)$.

4°. Expliciter la vraisemblance de l'échantillon et en déduire l'expression de V_n. Monter qu'il est asymptotiquement sans biais et qu'il converge en probabilité vers σ.

5°. Comparer T_n et V_n.

6°. Construire un intervalle de confiance de niveau $1-\alpha$ pour σ.

1.15 Durée de vie d'un système.

Un système fonctionne en utilisant deux machines de types différents. Les durées de vie X_1 et X_2 des deux machines suivent des lois exponentielles de paramètre respectif λ_1 et λ_2. Les variables aléatoires X_1 et X_2 sont supposées indépendantes.

1°. Soit X une variable aléatoire réelle. Montrer que

$$X \sim \mathscr{E}(\lambda) \Rightarrow \forall x > 0,\ \mathbb{P}(X > x) = \exp(-\lambda x)$$

2°. Soit $t \geq 0$. Calculer la probabilité pour que le système ne tombe pas en panne avant la date t. En déduire la loi de la durée de vie Z du système. Calculer la probabilité pour que la panne du système soit due à une défaillance de la machine 1.

3°. On dispose de n systèmes identiques et fonctionnant indépendamment les uns des autres, et dont on observe les durées de vie $Z_1, \ldots, Z_n$.

a. Écrire le modèle statistique correspondant et la vraisemblance associée. Le paramètre bidimensionnel (λ_1, λ_2) est-il identifiable ?

b. Supposons que l'on observe à la fois les durées de vie des systèmes et la cause de la défaillance (machine 1 ou 2), notée T_i. Écrire la vraisemblance associée au nouveau modèle statistique. Le paramètre bidimensionnel (λ_1, λ_2) est-il identifiable ?

Dans cette question, on considère un seul système utilisant une machine de type 1 et une machine de type 2, mais on suppose que l'on dispose d'un stock n_1 de machines de type 1, de durées de vie $X_1^1, \ldots, X_1^{n_1}$ et d'un stock de n_2 machines de type 2, de durées de vie $X_2^1, \ldots, X_2^{n_2}$. Quand une machine tombe en panne, on la remplace par une machine de même type, tant que le stock correspondant n'est pas épuisé. Quand cela arrive, on dit que le système tombe en panne. On note toujours Z la durée de vie du système. Le cas $n_1 = n_2 = 1$ correspond donc à la première question (pas de stock).

a. Donner la loi de la somme de n variables aléatoires i.i.d. de loi exponentielle de paramètre $\lambda > 0$.

b. Écrire Z en fonction des X_j^i, $j = 1,2$, $i = 1, \ldots, n_j$ et en déduire $\mathbb{P}[Z \geq t]$., en fonction de $t, n_1, n_2, \lambda_1, \lambda_2$.

On note N le nombre de machines (des deux types confondus) sorties du stock quand le système tombe en panne et Z_0 la durée écoulée avant la première panne d'une machine. On note Z_i la durée écoulée entre la i-ème panne et la $(i+1)$-ième panne. La durée de vie totale du système est donc :

$$Z = \sum_{i=0}^{N} Z_i \tag{24}$$

La $(N+1)$-ème panne est donc la panne fatale au système.

c. Montrer que les variables Z_i sont i.i.d. et donner leur loi. On pourra utiliser (après l'avoir démontré et interprété) le résultat suivant : si X est une variable aléatoire de loi exponentielle de paramètre $\lambda > 0$, alors

$$\forall s, t \geq 0,\ \mathbb{P}[X \geq s+t | X \geq s] = \mathbb{P}[X \geq t] = e^{-\lambda t} \tag{25}$$

d. Préciser l'ensemble des valeurs possibles pour la variable N et en donner la loi.

e. Montrer que N et Z_i sont indépendantes. Calculer $\mathbb{E}[Z|N]$ en fonction de N, λ_1, λ_2 et en déduire l'expression de $\mathbb{E}[Z]$ en fonction de $\mathbb{E}[N]$, λ_1 et λ_2.

1.16 Loi exponentielle : différents estimateurs.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi exponentielle de paramètre $1/\theta$.

1°. Expliciter le modèle.

2°. Estimer θ par la méthode du maximum de vraisemblance. Quelles sont les propriétés non asymptotiques et asymptotiques de l'estimateur ?

3°. Proposez un estimateur de θ par la méthode des moments.

4°. Soit Z le nombre d'observations de l'échantillon qui sont supérieures ou égales à 2. Déterminez la loi de Z et déduisez-en un estimateur dont vous étudierez la convergence.

1.17 Loi de Cauchy : différents estimateurs.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi de Cauchy de densité

$$f_\theta(x) = \frac{1}{\pi} \times \frac{1}{1 + (x-\theta)^2} \tag{26}$$

1°. On veut estimer θ par la moyenne empirique. Est-ce un bon estimateur ?

2°. Proposez un estimateur de θ par la méthode des moments.

3°. Proposez un estimateur plug-in de θ à partir de la médiane de l'échantillon.

4°. Étudiez l'estimateur du maximum de vraisemblance de θ. Existe-t-il une solution explicite ? Comment peut-on établir sa consistance ? Sa vitesse de convergence ?

Soit Y_n la v.a. représentant le nombre d'observations de l'échantillon négatives ou nulles. On pose également $p(\theta) = \mathbb{P}_\theta[X_1 \leq 0]$.

5°. Déterminer le modèle image par Y_n et en déduire un estimateur p_n de $p = p(\theta)$.

6°. Proposez ensuite un estimateur de θ à partir de p_n. Étudiez ses propriétés asymptotiques.

1.18 Comportement asymptotique de la variance empirique.

Soient $(X_1, \ldots, X_n)$ un n-échantillon d'une variable aléatoire X, telle que $\mathbb{E}[X^4] < \infty$ (on notera $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}(X)$ et $\mathbb{V} = \mathbb{V}((X-\mu)^2)$). On considère les statistiques suivantes :

$$\begin{aligned} \overline{X}_n &= \frac{1}{n}\sum_{i=1}^n X_i \\ V_n &= \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \\ S_n^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \end{aligned}$$

On notera que V_n est la variance empirique dans le cas où l'espérance est connue mais est incalculable. Le but de l'exercice est de montrer que V_n et S_n^2 ont le même comportement asymptotique.

1°. Montrer que

$$\sqrt{n}\left(V_n - \sigma^2\right) \rightsquigarrow \mathcal{N}(0, \mathbb{V})$$

2°. Montrer que $S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$

3°. Montrer que $S_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2$.

4°. On suppose $\mathbb{E}[X] = 0$, quitte à travailler sur les variables centrées. Montrer, grâce au théorème de Slutsky, que

$$\sqrt{n}\left(S_n^2 - \sigma^2\right) \rightsquigarrow \mathcal{N}(0, \mathbb{V})$$

5°. En travaillant sur le TCL appliqué au couple aléatoire $(\overline{X}_n, \overline{X_n^2})$ et en utilisant la delta-méthode, retrouver le résultat précédent.

6°. Construire un intervalle de confiance à 95% pour σ^2.

1.19 Modèle à variable cachée.

On considère un ensemble de n individus donnés, au sein d'une population. À chacun de ces n individus est envoyé un questionnaire, sur lequel il est demandé d'indiquer depuis combien de temps l'individu n'est pas tombé malade (sans compter une éventuelle maladie actuelle). On modélise cette durée par une variable aléatoire X_i, de loi exponentielle de paramètre inconnu θ. Cependant, tous les individus ne renvoient pas le questionnaire. Pour chaque individu i, on considère une variable aléatoire Z_i, non observée, qui prend la valeur 1 si l'individu i a répondu au questionnaire, et la valeur 0 sinon. On suppose que pour chaque individu i, les variables X_i et Z_i sont indépendantes, et que les couples $(X_1, Z_1), \ldots, (X_n, Z_n)$ sont i.i.d. On note p le paramètre (inconnu) de la loi de Bernoulli des Z_i.

On suppose qu'après un traitement informatique des réponses obtenues, on observe les variables $Y_i = Z_i X_i + (1 - Z_i)a, i = 1, \ldots, n$ où a est un réel positif donné, arbitraire. En particulier, si l'individu i n'a pas répondu au questionnaire, la valeur a est observée. On notera $\mathbb{P}_{\theta,p}$ la mesure de probabilité associée à la loi d'une observation.

1°. Montrer que $\mathbb{P}_{\theta,p}$ est absolument continue par-rapport à $\mu = \delta_a + \lambda$, où δ_a est la mesure de Dirac en a et λ est la mesure de Lebesgue sur $\mathbb{R}$. Calculer la densité de $\mathbb{P}_{\theta,p}$ par-rapport à μ.

2°. Ecrire le modèle statistique associé aux observations $(Y_1, \ldots, Y_n)$ et calculer la vraisemblance du modèle par rapport à la mesure dominante μ.

3°. Le couple (p, θ) est-il identifiable ?

1.20 Processus de Poisson.

Soient $\tau_1, ..., \tau_n$ n v.a.i.i.d. de loi exponentielle de paramètre θ. Soit $T_n = \tau_1 + ... + \tau_n$ et

$$N_t = \sum_{n\geq 1} \mathbb{1}_{[T_n \leq t]} = \inf\{n : T_n > t\} \tag{27}$$

$t \geq 0$, $\theta > 0$, $n \in \mathbb{N}^*$.

1°. Déterminer la loi de T_n et celle de N_t.

2°. Soit $T_0^{(t)} = 0$ et $T_n^{(t)} = T_{N_t+n}$. Monter que les variables aléatoires $\left(T_{n+1}^{(t)} - T_n^{(t)}\right)_{n\geq 1}$ sont des v.a.i.i.d. de loi exponentielle de paramètre θ.

On appelle processus de Poisson un processus de comptage $(X_t)_{t\geq 0}$ stationnaire et à accroissements indépendants, c'est à dire tel que :

- $X_{t+s} - X_s$ a même loi que X_t, $\forall s, t \geq 0$.
- $X_{t+s} - X_t$ indépendant de $(X_u)_{u\leq s}$, $\forall s, t \geq 0$.

3°. Démontrer que $(N_t)_{t\geq 0}$ est un processus de Poisson.

On souhaite estimer l'intensité θ d'un processus de Poisson, de deux manières différentes.

4°. On suppose que l'on a observé le processus jusqu'au temps t. Calculer la vraisemblance de l'observation et la valeur $\widehat{\theta}$ de θ qui maximise cette vraisemblance. Montrer que $\widehat{\theta}$ est un estimateur sans biais de θ. En remarquant que

$$N_t = N_t - N_{[t]} + \sum_{i=1}^{[t]-1} (N_{i+1} - N_i) \tag{28}$$

Montrer que

$$\frac{N_t}{t} \underset{\mathbb{P}-\text{p.s.}}{\overset{t\to+\infty}{\longrightarrow}} \theta \tag{29}$$

puis que

$$\sqrt{\frac{t}{\theta}}\left(\frac{N_t}{t}-\theta\right) \overset{t\to+\infty}{\rightsquigarrow} \mathcal{N}(0,1) \tag{30}$$

5°. On suppose que l'on observe uniquement le n-ième instant. Calculer la vraisemblance de cette observation et en déduire que l'estimateur du maximum de vraisemblance est n/T_n. En remarquant que $T_n = \sum_{i=1}^{n-1}(T_{i+1}-T_i)$, montrer que

$$\frac{n}{T_n} \underset{\mathbb{P}-\text{p.s.}}{\overset{n\to+\infty}{\longrightarrow}} \theta \tag{31}$$

puis que

$$\sqrt{n}\left(\theta\frac{T_n}{n}-1\right) \overset{n\to+\infty}{\rightsquigarrow} \mathcal{N}(0,1) \tag{32}$$

1.21 Taux de hasard d'un processus.

On considère des temps d'arrivées (moments d'occurrence d'un phénomène, par exemple). $0 = T_0 < T_1 < T_2 < \dots$ et $X_i := T_i - T_{i-1}$, $i \geq 1$, les durées inter-arrivées correspondantes. On parle d'un processus de renouvellement quand les X_i sont i.i.d. Dans ce cas, on suppose alors que les X_i admettent une densité $f(x)$ et l'on note $F(x)$ leur fonction de répartition et $R(x) = 1-F(x)$ leur fonction de survie. On s'intéresse au taux de hasard (*hazard rate*) définit par

$$B(x) = \frac{f(x)}{1-F(x)} \tag{33}$$

On souhaite estimer $B(x)$ pour $x \in]0,+\infty[$.

1°. Exprimer $B(x)$ en fonction de $R(x)$ et montrer que

$$\frac{\mathbb{P}[x < X \leq x+h]}{\mathbb{P}[X > x]} = B(x)h + o(h) \tag{34}$$

2°. En déduire une interprétation de $B(x)$.

On suppose à partir de maintenant que les X_i sont i.i.d. de loi exponentielle de paramètre θ et l'on considère les deux schémas d'observation suivants :

Schéma d'observation 1 : on observe les n premières arrivées, $(X_i\,; 1 \leq i \leq n)$.

Schéma d'observation 2 : On observe les arrivées jusqu'à un temps $T > 0$ donné,

$$(N_T\,; X_i\,; 1 \leq i \leq N_T) \text{ avec } N_T := \sum_{i=1}^{\infty} \mathbb{1}_{[T_i \leq T]}$$

3°. Définir les deux modèles statistiques correspondant.

4°. Calculer $B(x)$ et interpréter le résultat.

5°. Pour le schéma 1, construire l'estimateur du maximum de vraisemblance $\widehat{B_1}$ de B et expliciter la vitesse de convergence de son risque quadratique lorsque $n \to +\infty$.

6°. On se place dans le schéma 2. Quelle est la loi de N_T ? Écrire la vraisemblance et trouver un estimateur $\widehat{B_2}$ de B.

7°. On pose

$$\widetilde{B_2} = \left(\frac{1}{N_T}\sum_{i=1}^{N_T} X_i\right)^{-1} \tag{35}$$

Interpréter cet estimateur et expliciter la vitesse de convergence de son risque quadratique quand $T \to +\infty$.

8°. En déduire la vitesse de convergence de $\widehat{B_2}$ quand $T \to +\infty$.

1.22 Sondages et estimateur de Horvitz-Thompson.

On considère une population $U = \{1,\dots,k,\dots,N\}$. Un sondage aléatoire consiste à sélectionnner dans U dans un certain nombre d'individus, avec ou sans remise. Dans toute la suite de cet exercice, nous supposerons que la sélection se fait sans remise. Un échantillon s est donc un sous-ensemble de U. On appelle plan de sondage, une probabilité sur l'ensemble $\mathcal{S}$ de tous les échantillons s possibles, obtenus à partir de la population U.

On notera S un échantillon aléatoire, c'est à dire une variable aléatoire à valeurs dans $\mathcal{S}$ et l'on notera

$$p(s) = \mathbb{P}[S = s] \tag{36}$$

La taille $n(S) = \text{card}(S)$ d'un échantillon aléatoire est une variable aléatoire.

On définit les variables aléatoires de Cornfield (1944) par : quelque soit $k \in U$,

$$\delta_k = \mathbb{1}_{[k\in S]} \tag{37}$$

et l'on définit les probabilités d'inclusion simples et doubles par : quelques soient $k, l \in U$,

$$\pi_k = \mathbb{P}[k \in S] = \sum_{s\in\mathcal{S}:k\in s} p(s) \tag{38}$$

$$\pi_{k,l} = \mathbb{P}[k \in S; l \in S] \tag{39}$$

1°. Démontrer que pour un échantillon sans remise, de taille fixe n (on parle alors de sondage aléatoire simple), on a

$$\sum_{k\in U} \pi_k = n \tag{40}$$

2°. Calculer $\mathbb{E}[\delta_k]$, $\mathbb{E}[\delta_k\delta_l]$, $\mathbb{V}(\delta_k)$ et $\Delta_{kl} = \text{cov}(\delta_k,\delta_l)$.

3°. En déduire que

$$\sum_{k,l\in U:k\neq l} \pi_{k,l} = n(n-1) \text{ et } \sum_{k\in U}\Delta_{kl} = 0 \tag{41}$$

On considère un caractère x que l'on souhaite mesurer dans la population. Ce caractère va être estimé à partir des valeurs qu'il prend dans l'échantillon. On peut donc considérer que x est la réalisation d'une variable aléatoire X définie sur U. On notera T (pour total) la statistique

$$T = \sum_{k\in U} X_k \tag{42}$$

Le paramètre θ que l'on va estimer est donc le nombre réel (déterministe) $\theta = t = \sum_{k\in U} x_k$. En 1952, Horvitz et

Thompson ont proposé l'estimateur suivant pour estimer la somme T

$$\widehat{\mathrm{T}}(\mathrm{X}) = \sum_{k\in\mathrm{S}} \frac{\mathrm{X}_k}{\pi_k} \tag{43}$$

4°. Qu'est-ce qui est aléatoire dans la formule précédente ? Interpréter cet estimateur.

5°. Démontrer que si $\pi_k > 0\ \forall k$, alors $\widehat{\mathrm{T}}(\mathrm{X})$ est sans biais.

6°. Démontrer que

$$\mathbb{V}\left(\widehat{\mathrm{T}}(\mathrm{X})\right) = \sum_{k\in\mathrm{U}}\sum_{l\in\mathrm{U}} \frac{\mathrm{X}_k\mathrm{X}_l}{\pi_k\pi_l}\Delta_{kl} \tag{44}$$

7°. Montrer que si le plan de sondage est de taille fixe, on a également la formule de Yates-Grundy (1953) suivante :

$$\mathbb{V}\left(\widehat{\mathrm{T}}(\mathrm{X})\right) = -\frac{1}{2}\sum_{k,l\in\mathrm{U}:k\neq l}\left(\frac{\mathrm{X}_k}{\pi_k} - \frac{\mathrm{X}_l}{\pi_l}\right)^2 \Delta_{kl} \tag{45}$$

8°. Dans la population $\mathrm{U} = \{1,2,3\}$ dans laquelle on définit le plan de sondage $p(\{1,2\}) = 1/2$, $p(\{1,3\}) = 1/4$ et $p(\{2,3\}) = 1/4$, on considère une variable X définie sur U par $x_1 = x_2 = 3$ et $x_3 = 6$ dont on veut estimer le total T.

Déterminer les probabilités d'inclusion simples et doubles, donner la distribution de l'estimateur de Horvitz-Thompson $\widehat{\mathrm{T}}$, calculer la variance de cet estimateur. Donner la distribution de probabilité d'un estimateur de la variance de $\widehat{\mathrm{T}}$.

1.23 Estimateur du coefficient de corrélation empirique.

Soit $(\mathrm{X}_1,\mathrm{Y}_1),\ldots,(\mathrm{X}_n,\mathrm{Y}_n)$ un n-échantillon d'un vecteur (X,Y) dont le coefficient de corrélation linéaire est noté ρ. Le coefficient de corrélation empirique est par définition égal à

$$\rho_n = \frac{\sum_{i=1}^n \left(\mathrm{X}_i - \overline{\mathrm{X}}\right)\left(\mathrm{Y}_i - \overline{\mathrm{Y}}\right)}{\left[\sum_{i=1}^n \left(\mathrm{X}_i - \overline{\mathrm{X}}\right)^2 \times \sum_{i=1}^n \left(\mathrm{Y}_i - \overline{\mathrm{Y}}\right)^2\right]^{1/2}} \tag{46}$$

où $\overline{\mathrm{X}}$ et $\overline{\mathrm{Y}}$ sont respectivement les moyennes empiriques des X_i et des Y_i.

1°. On suppose que X et Y ont des moments d'ordre 4. Montrer que

$$\sqrt{n}(\rho_n - \rho) \rightsquigarrow \mathcal{N}\left(0, c^2\right) \tag{47}$$

où c est une constante à déterminer en fonction des moments de X et Y.

2°. Si (X,Y) est gaussien, montrer que $c = 1-\rho^2$. Déterminer une fonction différentiable ψ dont la dérivée vaut $(1-\rho^2)^{-1}$ et montrer que

$$\sqrt{n}(\psi(\rho_n) - \psi(\rho)) \rightsquigarrow \mathcal{N}(0,1) \tag{48}$$

3°. Si X et Y sont indépendants et gaussiens, montrer que la densité de ρ_n est donnée par

$$f(t) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}\left(1-t^2\right)^{(n-4)/2}\mathbb{1}_{]-1,1[}(t) \tag{49}$$

Retrouver le résultat de la première question en appliquant le théorème de Scheffé à la densité de $\sqrt{n}\rho_n$.

1.24 Loi exponentielle et variable inobservée

On considère un n-échantillon $\mathrm{X} = (\mathrm{X}_1,\ldots,\mathrm{X}_n)$ d'une variable aléatoire dont la loi a pour densité :

$$f(x) = \theta e^{-\theta x}\mathbb{1}_{[0,+\infty[}(x) \tag{50}$$

avec $\theta > 0$ paramètre inconnu. On suppose que les X_i ne sont pas observés directement. La seule information connue est le fait que X_i soit supérieur à 2 ou non. On pose $\mathrm{Y}_i = \mathbb{1}_{[\mathrm{X}_i>2]}$ pour $i = 1,\ldots,n$ et

$$\overline{\mathrm{Y}}_n = \frac{1}{n}\sum_{i=1}^n \mathrm{Y}_i. \tag{51}$$

1°. Préciser la loi de Y_1 ainsi que celle de $n\overline{\mathrm{Y}}_n$. Calculer $\mathbb{E}[\mathrm{Y}_1]$ en fonction de θ.

On souhaite estimer le paramètre θ. On pose $\lambda = e^{-2\theta}$ et

$$\widehat{\theta}_n = \begin{cases} -\frac{1}{2}\ln\overline{\mathrm{Y}}_n & \text{si } \overline{\mathrm{Y}}_n > 0 \\ 0 & \text{sinon.} \end{cases} \tag{52}$$

2°. Calculer $\mathbb{P}\left[\overline{\mathrm{Y}}_n \neq 0\right]$ et en déduire

$$\lim_{n\to+\infty} \mathbb{P}\left[\overline{\mathrm{Y}}_n \neq 0\right]. \tag{53}$$

3°. Démontrer que $\overline{\mathrm{Y}}_n$ converge presque sûrement vers λ et que $\sqrt{n}(\overline{\mathrm{Y}}_n - \lambda)$ est asymptotiquement normal. Préciser la variance de la loi limite en fonction de θ.

4°. Expliquer pourquoi $\overline{\mathrm{Y}}_n$ est différent de 0, presque sûrement, lorsque n suffisamment grand.

5°. Démontrer que $\widehat{\theta}_n$ est fortement consistant et asymptotiquement normal. Préciser la variance de la loi limite en fonction de θ.

1.25 Loi exponentielle : estimation d'un couple de paramètres

Dans tout cet exercice, n est un entier naturel non nul. Soient $\theta \geq 0$, $\beta > 0$ et f la fonction définie sur $\mathbb{R}$ par

$$f(x) = \begin{cases} \dfrac{1}{\beta}e^{-\dfrac{x-\theta}{\beta}} & \text{si } x \geq \theta, \\ 0 & \text{sinon.} \end{cases} \tag{54}$$

Soit $(\mathrm{X}_1,\ldots,\mathrm{X}_n)$ un n-uplet de variables aléatoires réelles mutuellement indépendantes et identiquement distribuées, dont la loi a pour densité f.

1°. Vérifier que f est une densité d'une loi de probabilité.

2°. Calculer $\mathbb{E}[\mathrm{X}_1]$ et $\mathbb{V}(\mathrm{X}_1)$ en justifiant leur existence.

3°. On pose $\mathrm{Y}_n = \min(\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_n)$. Déterminer une densité de Y_n. Calculer son espérance et sa variance.

4°. Y_n est-il un estimateur sans biais de θ ? Asymptotiquement sans biais ?

5°. Déduire des questions précédentes l'erreur quadratique moyenne $\mathbb{E}\left[(\mathrm{Y}_n - \theta)^2\right]$.

La suite $(\mathrm{Y}_n)_n$ converge-t-elle dans L^2 ? En probabilité ?.

Soient $S_n = \sum_{i=1}^{n} X_i$ et $Z_n = \frac{1}{n} S_n - Y_n$.

6°. Calculer $\mathbb{E}[Z_n]$. Z_n est-il un estimateur sans biais de β ? Asymptotiquement sans biais ?

7°. Calculer $\mathbb{V}(Z_n)$ en fonction de $\mathrm{cov}(S_n, Y_n)$ et montrer que $\mathbb{V}(Z_n)$ tend vers zéro quand n tend vers l'infini. La suite $(Z_n)_n$ converge-t-elle dans L^2 ? En probabilité ?

8°. Démontrer que le couple $(\widehat{\theta}_n, \widehat{\beta}_n)$ donné par

$$\begin{cases} \widehat{\theta}_n = \dfrac{1}{n-1}\left(nY_n - \dfrac{S_n}{n}\right) \\ \widehat{\beta}_n = \dfrac{1}{n-1}\left(S_n - nY_n\right) \end{cases} \tag{55}$$

est un estimateur sans biais du couple (θ, β). Calculer la variance de $\widehat{\theta}_n$ et celle de $\widehat{\beta}_n$.

9°. Démontrer que

$$\frac{\sqrt{n}}{\beta}\left(\frac{S_n}{n} - (\theta + \beta)\right) \tag{56}$$

converge en loi vers une variable aléatoire dont on précisera la loi.

10°. Soit

$$T_n = \frac{\sqrt{n}}{\beta}\left(Y_n - \theta\right). \tag{57}$$

Déterminer une fonction de densité de T_n et étudier la convergence en probabilité de T_n.

11°. Démontrer que

$$\frac{\sqrt{n}}{\beta}\left(Z_n - \beta\right) \tag{58}$$

converge en loi vers une variable aléatoire dont on précisera la loi.

On suppose θ connu et l'on souhaite à nouveau estimer β. Soit V_i la variable aléatoire définie par $V_i = 1$ si $X_i \geq \theta + 1$ et 0 sinon. On note $\overline{V}_n$ la moyenne empirique des V_i.

12°. Déterminer la loi de $\sum_{i=1}^{n} V_i$ en précisant ses paramètres. Démontrer que $\overline{V}_n$ est un estimateur de $\psi(\beta)$, où ψ est une fonction à préciser. En déduire un estimateur de β, fonction de $\overline{V}_n$.

1.26 Loi puissance

On considère un n-échantillon $X_1, ..., X_n$ d'une v.a. X dont la loi a pour densité

$$f(x) = \frac{1}{2\sqrt{x\theta}} \mathbb{1}_{]0,\theta]}(x) \tag{59}$$

où $\theta > 0$ est un paramètre que l'on souhaite estimer. On note $\overline{X}$ la moyenne empirique de l'échantillon.

1°. Montrer que la fonction de répartition de X est, quelque soit $x \in]0, \theta]$,

$$F(x) = \sqrt{\frac{x}{\theta}} \tag{60}$$

2°. Calculer la vraisemblance $L(x_1, ..., x_n, \theta)$ du n-échantillon.

3°. Déterminer l'estimateur du maximum de vraisemblance de θ. Le modèle est-il régulier ?

4°. On pose

$$\widehat{\theta}_1 = X_{(n)} = \max_{i=1}^{n} X_i \tag{61}$$

Déterminer la fonction de répartition F_n de $\widehat{\theta}_1$ et montrer que sa densité est donnée par

$$f_n(x) = \frac{n}{2\theta^{n/2}} x^{(n-2)/2} \mathbb{1}_{]0,\theta]}(x) \tag{62}$$

5°. Démonter que $\widehat{\theta}_1$ est biaisé, mais asymptotiquement sans biais. Calculer son risque quadratique et étudier la consistance de cet estimateur.

6°. Démontrer que $n\left(\widehat{\theta}_1 - \theta\right)$ converge en loi, quand n tend vers l'infini, vers une loi de densité

$$h(x) = \frac{1}{2\theta} \exp\left(\frac{x}{2\theta}\right) \mathbb{1}_{]-\infty,0]}(x) \tag{63}$$

Expliquer pourquoi $\widehat{\theta}_1$ n'est pas asymptotiquement normal.

7°. Calculer $\mathbb{E}_\theta[X]$ et démontrer que l'estimateur des moments de θ est $\widehat{\theta}_2 = 3\overline{X}$.

8°. Démontrer que $\widehat{\theta}_2$ est un estimateur sans biais de θ et que

$$\mathbb{V}_\theta(\widehat{\theta}_2) = \frac{4\theta^2}{5n} \tag{64}$$

En déduire que cet estimateur est consistant et que

$$\sqrt{n}\left(\widehat{\theta}_2 - \theta\right) \rightsquigarrow \mathcal{N}\left(0, \frac{4\theta^2}{5}\right) \tag{65}$$

9°. Déterminer la limite en loi de la suite de v.a.

$$\sqrt{n}\left(\frac{\sqrt{5}}{2}\ln(3\overline{X}) - \frac{\sqrt{5}}{2}\ln\theta\right) \tag{66}$$

1.27 Loi binomiale négative et loi de Pascal : estimations de paramètres

Dans tout cet exercice, n est un entier naturel non nul. On considère une suite de v.a.i.i.d. $(X_n)_{n\geq 1}$ de loi de Bernoulli de paramètre $p \in]0,1[$. On notera $q = 1 - p$. L'événement $[X_n = 1]$ représente un succès au n-ième tirage.

1°. Soit Y la variable aléatoire égale au nombre de tirages nécessaires avant d'obtenir un premier succès. Démontrer que

$$\forall k \geq 1,\ \mathbb{P}[Y = k] = q^{k-1} p.$$

Déterminer l'expression de l'espérance $\mathbb{E}[Y]$ et de la variance $\mathbb{V}(Y)$ de Y, en fonction de p et q.

2°. On considère deux v.a. indépendantes Y_1 et Y_2 de même loi géométrique de paramètre p. On note $S = Y_1 + Y_2$. Démontrer que

$$\forall n \geq 2,\ \mathbb{P}[S = n] = (n-1)p^2 q^{n-2}.$$

3°. Déterminer la probabilité conditionnelle $\mathbb{P}[Y_1 = k|S = n]$. Interpréter le résultat.

On note $S_n = X_1 + ... + X_n$ la somme des n premiers tirages et T_n le nombre de tirages nécessaires pour obtenir n succès. On dit que T_n suit une loi de Pascal de paramètres n et p.

4°. Reconnaître la loi suivie par S_n, puis démontrer que pour tout $k \geq n$,

$$[T_n = k] = [S_{k-1} = n-1] \cap [X_n = 1].$$

5°. En déduire que

$$\forall k \geq n,\ \mathbb{P}[T_n = k] = \begin{pmatrix} k-1 \\ n-1 \end{pmatrix} p^n q^{k-n}.$$

6°. On pose $Y_1 = T_1$ et pour tout $i \geq 2$, $Y_i = T_i - T_{i-1}$. Montrer que les $(Y_i)_i$ forment une suite de variables de loi géométrique de paramètre p. Montrer qu'elles sont mutuellement indépendantes et que

$$T_n = Y_1 + ... + Y_n.$$

7°. Déterminer $\mathbb{E}[T_n]$ et $\mathbb{V}(T_n)$.

8°. On note V_n le nombre d'échecs dans la séquence $(X_i)_i$ nécessaires avant d'obtenir n succès. On dit que V_n suit une loi binomiale négative de paramètres n et p. Démontrer que pour tout $n \geq 1$, V_n et T_n sont liées par la relation suivante :

$$T_n = V_n + n.$$

En déduire que

$$\forall k \geq 0,\ \mathbb{P}[V_n = k] = \begin{pmatrix} k+n-1 \\ k \end{pmatrix} p^n q^k.$$

On considère un couple de v.a. (X, Λ) dont la loi conditionnelle de X sachant $[\Lambda = \lambda]$ suit une loi de Poisson de paramètre $\lambda > 0$. On suppose que Λ suit une loi gamma $\gamma(n, \theta)$ de paramètres $n \in \mathbb{N}^*$ et $\theta > 0$, dont la densité g est définie par

$$\forall t > 0,\ g(t) = \frac{t^{n-1}}{\Gamma(n)\theta^n} e^{-t/\theta}, \tag{67}$$

et Γ est la fonction Gamma d'Euler, définie pour $z > 0$ par

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t}\, dt. \tag{68}$$

On dit que la loi de X est un mélange Poisson-Gamma.

9°. Démontrer que la densité jointe du couple (X, Λ) s'écrit :

$$\forall \lambda > 0,\ \forall k \geq 0,\ f(k, \lambda) = \frac{\lambda^{n+k-1}}{k!\,\Gamma(n)\,\theta^n} e^{-\lambda(1+1/\theta)}. \tag{69}$$

Déterminer et reconnaître la loi marginale de X en précisant ses paramètres.

10°. Démontrer que

$$\mathbb{E}\left[\frac{n-1}{T_n - 1}\right] = p \text{ et } \mathbb{E}\left[\frac{n}{T_n}\right] > p,$$

après avoir justifié de l'existence de ces espérances.

11°. On suppose p inconnu et on souhaite l'estimer à partir des observations. Démontrer que la suite de v.a. $(n/T_n)_n$ converge en probabilité et préciser sa limite. En déduire un estimateur $\hat{p}$ de p. Est-il biaisé ? Proposer un estimateur $\tilde{p}$ non biaisé.

1.28 Convergences en loi et en probabilité de suites de v.a.

Soit θ un réel strictement positif. Toutes les variables aléatoires sont définies sur un même espace probabilisé $(\Omega, \mathscr{F}, \mathbb{P})$. Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles, mutuellement indépendantes, de même loi uniforme sur le segment $[-\theta\sqrt{3}, \theta\sqrt{3}]$. Pour $n \geq 1$, on pose :

$$S_n = X_1^2 + ... + X_n^2, \qquad T_n = \sqrt{\frac{S_n}{n}},$$

$$U_n = \sqrt{n}(T_n - \theta), \qquad V_n = \frac{\sqrt{n}}{2\theta}(T_n^2 - \theta^2).$$

1°. Justifier le fait que X_1 possède des moments d'ordre 4, puis calculer $\mathbb{E}[X_1^4]$ et $\mathbb{V}(X_1^2)$.

2°. Démontrer que $(S_n/n)_n$ converge en probabilité vers θ^2 et en déduire que $(T_n)_n$ converge en probabilité. Préciser sa limite.

3°. Montrer que $(V_n)_n$ converge en loi vers une variable aléatoire Z suivant une loi normale centrée et de variance $\theta^2/5$.

4°. Démontrer que pour tout a réel non nul fixé,

$$\forall x \in \mathbb{R} \text{ tel que } x \neq -a,\ x - a = \frac{x^2 - a^2}{2a} - \frac{(x^2 - a^2)^2}{2a(x+a)^2}.$$

5°. Montrer que $U_n = V_n - W_n$ où W_n est une variable aléatoire vérifiant, pour tout $n \geq 1$,

$$0 \leq W_n \leq \frac{\sqrt{n}}{2\theta^3}\left(T_n^2 - \theta^2\right)^2.$$

6°. Montrer que

$$\lim_{n\to+\infty} \mathbb{E}[W_n] = 0,$$

puis que $(W_n)_n$ converge vers 0 en probabilité.

Une fonction $f : \mathbb{R} \longrightarrow \mathbb{R}$ est à support compact s'il existe un intervalle $K = [\alpha, \beta] \subset \mathbb{R}$ tel que pour tout $x \notin K$, $f(x) = 0$.

On rappelle qu'une suite de variables aléatoires réelles $(U_n)_n$ converge en loi vers une variable aléatoire réelle U si, et seulement si, pour toute fonction f continue sur $\mathbb{R}$ et à support compact, on a

$$\lim_{n\to+\infty} \mathbb{E}\left[|f(U_n) - f(U)|\right] = 0.$$

7°. Soit f une fonction continue sur $\mathbb{R}$, à support compact K. Démontrer l'existence de

$$M = \sup_{x\in\mathbb{R}} |f(x)|.$$

8°. Soit $\epsilon > 0$. Démontrer qu'il existe $\delta > 0$ tel que

$$\mathbb{E}\left[|f(U_n) - f(V_n)|\right] \leq \frac{\epsilon}{2} + 2M \times \mathbb{P}[|W_n| \geq \delta].$$

9°. En déduire que

$$\lim_{n\to+\infty} \mathbb{E}\left[|f(U_n) - f(V_n)|\right] = 0.$$

10°. En déduire que $(U_n)_n$ converge en loi vers une variable aléatoire U dont on donnera la loi.

1.29 Analyse statistique d'un réseau social : estimation du nombre d'amis dans facebook.

2 Information, exhaustivité, optimalité

2.1 Exemples de statistiques exhaustives.

Dans chaque question, $(X_1, ..., X_n)$ un n-échantillon d'une va X dont la loi est donnée. On cherche à déterminer une statistique exhaustive.

1°. X est une va de loi uniforme sur $[0, \theta]$. Déterminer une statistique exhaustive pour θ. Montrer qu'elle est complète. Est-elle admissible ? Efficace ?

2°. X est une va de loi exponentielle de paramètre λ. Démontrer que $\sum X_i$ est une statistique exhaustive pour λ.

3°. X est une va de loi de Poisson de paramètre θ. Démontrer, de trois façons différentes, que $S = \sum_{i=1}^n X_i$ est une statistique exhaustive minimale pour θ. Démontrer que deux façons différentes qu'elle est complète.

4°. $X \sim \mathcal{N}(m, \sigma^2)$. Soient $\overline{X}$ la moyenne empirique et $S = \sum_{i=1}^n X_i^2$ la somme des carrés. Soit $T = (\overline{X}, S)$. Montrer que T est une statistique exhaustive pour le couple (m, σ^2).

5°. Soit $\theta > 0$ et X une variable aléatoire de densité

$$f(x, \theta) = \frac{\theta e^{\theta x}}{e^{\theta^2} - 1} \mathbb{1}_{[0,\theta]}(x) \tag{70}$$

La statistique $S = \sum_{i=1}^n X_i$ est-elle exhaustive ?

2.2 Variables inobservées.

On dispose d'observations binaires $(Y_1, \ldots, Y_n)$ i.i.d. qui dépendent d'un phénomène sous-jacent. On modélise cela par des variables aléatoires $(Y_1^*, \ldots, Y_n^*)$ inobservées qui sont des tirages i.i.d. suivant une loi normale (m, σ^2). On a donc :

$$\forall i, \qquad Y_i = \begin{cases} 1 & \text{si} \quad Y_i^* > 0 \\ 0 & \text{si} \quad Y_i^* \leq 0 \end{cases}$$
$$Y_i^* \sim \quad \mathcal{N}(m, \sigma^2)$$

1°. On note Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$. Exprimer la loi de Y_i en fonction de Φ.

2°. Ecrire le modèle statistique associé aux observations $(Y_1, \ldots, Y_n)$.

3°. Montrer que le couple (m, σ^2) n'est pas identifiable. Quel paramètre peut-on identifier ? Ce paramètre est noté θ. Réécrire le modèle statistique, avec θ comme nouveau paramètre d'intérêt.

4°. Trouver une statistique exhaustive et complète.

2.3 Procédé de capture / recapture.

On veut compter le nombre θ de poissons dans un lac fermé. Pour cela, on tire un poisson au hasard, on le marque et on le remet dans le lac. On tire un second poisson. S'il est déjà marqué, on en prend note et on le remet dans le lac. Sinon, on le marque à son tour et on le remet dans le lac. Et ainsi de suite.

On tire n poissons selon la procédure ci-dessus. Au n-ième tirage, l'observation consiste en une variable aléatoire Y_n qui vaut 1 si le poisson est déjà marqué, 0 sinon. Par définition, on a $Y_1 = 0$. Le but de l'exercice est de montrer que :

$$R_n = \sum_{i=1}^n Y_i$$

est une statistique exhaustive pour θ.

1°. Montrer que :

$$\mathbb{P}\left[Y_n = y_n, ..., Y_1 = y_1\right] =$$
$$\begin{cases} & \mathbb{P}\left[Y_n = y_n | Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1\right] \\ \times & \mathbb{P}\left[Y_{n-1} = y_{n-1} | Y_{n-2} = y_{n-2}, ..., Y_1 = y_1\right] \\ \vdots & \\ \times & \mathbb{P}\left[Y_1 = y_1\right] \end{cases}$$

2°. Montrer que la loi conditionnelle de Y_n sachant $Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1$ est une loi de Bernoulli de paramètre :

$$\frac{n - r_{n-1} - 1}{\theta}$$

et en déduire que la vraisemblance est proportionnelle à :

$$\prod_{i=1}^n \frac{(\theta - i + 1 + r_{i-1})^{1-y_i}}{\theta} \tag{71}$$

3°. Montrer que l'expression précédente se reécrit :

$$\frac{1}{\theta^{n-1}} \frac{(\theta - 1)!}{(\theta - n + r_n)!}$$

4°. En déduire que R_n est une statistique exhaustive pour θ.

2.4 Information de Fisher.

Calculer, lorsqu'elle existe, l'information de Fisher dans les modèles statistiques associés aux échantillons suivants :

1°. Un échantillon de n v.a.i.i.d. de loi de Poisson de paramètre λ :

$$\mathbb{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!} \text{ pour } k \in \mathbb{N}$$

2°. Un échantillon de n v.a.i.i.d. de loi de Pareto de paramètres α et θ avec $\alpha > 1$ et $\theta > 0$, de densité :

$$f(x) = \frac{\alpha - 1}{\theta}\left(\frac{\theta}{x}\right)^{\alpha}\mathbb{1}_{[x\geq\theta]}$$

3°. Un échantillon de n v.a.i.i.d. de loi de Weibull de paramètres α et θ avec $\alpha > 0$ et $\theta > 0$ de densité :

$$f(x) = \alpha\theta x^{\alpha-1}e^{-\theta x^{\alpha}}$$

4°. Un échantillon de n v.a.i.i.d. de loi uniforme sur $[0,\theta]$ avec $\theta > 0$ inconnu.

2.5 Score.

On étudie une variable aléatoire réelle X, de densité $f(\cdot,\theta)$ où $\theta \in \mathbb{R}^d$ est un paramètre vectoriel inconnu, et f est supposée connue, de classe C^1 sur $\mathbb{R}\times\mathbb{R}^d$.

1°. Quelle est le score du modèle, noté $S_X(x,\theta)$? Donner l'expression de l'information de Fisher $I_X(\theta)$.

2°. Supposons que l'on ne parvient pas à observer X, mais que seule est disponible la variable Y, définie par :

$$Y = \mathbb{1}_{[X\geq s]}$$

où s est un seuil connu. En supposant que l'on peut intervertir $\int_{\mathscr{X}}$ et $\frac{\partial}{\partial\theta}$, donner le score du nouveau modèle, noté $S_Y(y;\theta)$. En déduire que

$$S_Y(y;\theta) = \mathbb{E}[S_X(X;\theta)|Y = y],\ y \in \{0,1\}.$$

3°. En déduire alors que $I_X(\theta) >> I_Y(\theta)$, où $I_Y(\theta)$ est l'information de Fisher associée à Y. Cette inégalité s'entend au sens des matrices symétriques réelles. Quelle interprétation pourriez-vous donner à l'inégalité ci-dessus, dans le cas où $d = 1$?

2.6 Estimation d'une fonction de survie.

On dispose de n observations indépendantes des durées de vie de certains composants industriels. On suppose que les variables aléatoires $Y_1,\dots,Y_n$ associées sont i.i.d. de densité

$$f(t) = \theta e^{-\theta t}\mathbb{1}_{[t\geq 0]} \tag{72}$$

où $\theta > 0$ est un paramètre inconnu. Soit F la fonction de répartition de Y_1. On cherche à estimer la fonction de survie de Y_1, c'est-à-dire $\overline{F}(t) = 1 - F(t)$, à un instant t donné et connu.

1°. Proposer un estimateur $\widehat{F}(t)$ qui soit sans biais et convergent quelle que soit la loi des $(Y_i)_i$. Intuitivement, cet estimateur est il optimal (parmi les estimateurs sans biais) ?

2°. Calculer $\overline{F}(t)$ en fonction de t et θ.

3°. Calculer l'estimateur du maximum de vraisemblance de θ et en déduire un estimateur convergent $\widehat{F}(t)$ de $\overline{F}(t)$.

On admettra par la suite que $\widehat{F}(t)$ est biaisé.

4°. Calculer la loi limite de $\sqrt{n}\left(\widehat{F}(t) - \overline{F}(t)\right)$.

5°. Soit T la variable aléatoire définie par :

$$T = \mathbb{1}_{[Y_1\geq t]} \tag{73}$$

On note par ailleurs $S = Y_1 + \dots + Y_n$.

Déterminer la loi de Y_1 conditionnellement à S. Calculer $T^* = \mathbb{E}[T|S]$. Comment s'appelle cet estimateur ? Montrer que T^* est l'estimateur sans biais de $\overline{F}(t)$ optimal (parmi les estimateurs sans biais). T^* est-il efficace ?

2.7 Loi de Poisson : estimateur de la probabilité que $X = 0$.

On s'intéresse à l'estimation de $\theta = e^{-\lambda} = \mathbb{P}_\theta[X = 0]$ basée sur un échantillon $(X_1,\dots,X_n)$ de variables aléatoires i.i.d. de loi de Poisson $X \sim \mathscr{P}(\lambda)$. On considère les trois estimateurs suivants :

$$\begin{cases} \widehat{\theta}_1 = e^{-\overline{X}} \\ \widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[X_i=0]} \\ \widehat{\theta}_3 = \left(1 - \frac{1}{n}\right)^S \end{cases} \tag{74}$$

avec $S = \sum_{i=1}^n X_i$.

1°. Montrer que ces trois estimateurs sont convergents.

2°. Montrer que

$$\mathbb{E}\left[\widehat{\theta}_2 \middle| \sum_{i=1}^n X_i\right] = \widehat{\theta}_3 \tag{75}$$

2.8 Estimation par maximum de vraisemblance

1°. Calculer l'estimateur du maximum de vraisemblance (e.m.v.) $\hat{p}$ de p dans le modèle $X_i \sim B(p)$ et calculer la loi limite de $\sqrt{n}(\hat{p} - p)$.

2°. Calculer l'e.m.v. $(\widehat{m},\widehat{\sigma^2})$ de (m,σ^2) dans le modèle $X_i \sim \mathscr{N}(m,\sigma^2)$ et donner la loi limite du vecteur

$$\sqrt{n}\begin{pmatrix} \widehat{m} - m \\ \widehat{\sigma^2} - \sigma^2 \end{pmatrix} \tag{76}$$

3°. Calculer l'e.m.v. $(\widehat{a},\widehat{b})$ de (a,b) dans le modèle $X_i \sim \mathscr{U}[a,b]$ d'une loi uniforme sur $[a,b]$. Donner la loi limite du vecteur

$$n\begin{pmatrix} \widehat{a} - a \\ \widehat{b} - b \end{pmatrix} \tag{77}$$

2.9 Coordonnées polaires d'un vecteur gaussien.

Soit (X_1, X_2) un 2-échantillon de loi $\mathcal{N}(0, \sigma^2)$ où σ^2 est un paramètre inconnu. Soit (R, θ) le vecteur des coordonnées polaires de (X_1, X_2) ; c'est à dire que $X_1 = R\cos\theta$ et $X_2 = R\sin\theta$.

1°. Préciser le modèle statistique, le modèle image par T.

2°. T(X) est-elle une statistique exhaustive ? Complète ?

3°. Montrer que R(X) est une statistive exhaustive par deux méthodes différentes. Montrer qu'elle est complète. Montrer que $\theta(X)$ est une statistique libre.

4°. Montrer que deux façons différentes que R et θ sont indépendantes.

5°. Calculer l'information de Fisher du modèle.

2.10 Estimation du paramètre d'une loi de Poisson.

On considère un n-échantillon $X = (X_1, ..., X_n)$ suivant une loi de Poisson $\mathcal{P}(\lambda)$ de paramètre λ. On note $S = \sum_{i=1}^n X_i$ et $s = \sum_{i=1}^n x_i$.

1°. Préciser le modèle statistique et calculer la vraisemblance de l'échantillon.

2°. Montrer que S est une statistique exhaustive de l'échantillon pour le paramètre λ. Montrer qu'elle est complète pour λ.

3°. Déduire des questions précédentes un estimateur sans biais de variance minimale (VUMSB) du paramètre λ.

4°. Le modèle est-il régulier ? Si oui, calculer l'information $\mathbb{I}_X(\lambda)$ au sens de Fisher et en déduire un estimateur efficace.

5°. On s'intéresse maintenant au paramètre $\theta = e^{-\lambda}$. Quelle est la signification de θ ? Démontrer que $\widehat{\theta}_1 = \exp(\overline{X})$ est l'estimateur du maximum de vraisemblance de θ et qu'il est biaisé.

6°. Soient $Y_i = \mathbb{1}_{[X_i=0]}$. Montrer que Y_1 est un estimateur des moments de θ et qu'il est non biaisé.

7°. Déterminer la loi conditionnelle de Y_1 sachant S. En déduire l'estimateur VUSMB $\widehat{\theta}_2$ de θ.

8°. L'estimateur $\widehat{\theta}_2$ est-il efficace ?

9°. On considère maintenant l'estimateur $\widehat{\theta}_3 = \overline{Y}$, avec $\overline{Y}$ moyenne arithmétique des Y_i. Démontrer qu'il est VUMSB et efficace pour θ.

2.11 Estimation du paramètre d'une loi uniforme.

Soit X une variable aléatoire de loi uniforme sur $[0, \theta]$ et $(X_1, ..., X_n)$ un n-échantillon de X.

1°. Déterminer une statistique exhaustive du modèle et montrer qu'elle est complète.

2°. Expliquer pourquoi l'estimateur du maximum de vraisemblance est fonction d'une statistique exhaustive du modèle.

3°. Déduisez-en un estimateur $\widehat{\theta}$ qui soit admissible pour le risque quadratique parmi les estimateurs sans biais.

2.12 Maximumu de vraisemblance et reparamétrisation.

Soit $(X_1, ..., X_n)$ un n-échantillon d'une variable aléatoire X de densite

$$f(x, \theta) = \theta(1-x)^{\alpha} \mathbb{1}_{[0,1]}(x) \tag{78}$$

où $\theta > 0$ est le paramètre inconnu.

1°. Déterminer α en fonction de θ et calculer la vraisemblance du modèle.

2°. Déduisez-en un estimateur du maximum de vraisemblance de θ et de $1/\theta$.

3°. Soit $Z_i = -\ln(1 - X_i)$. Déterminer la loi des Z_i. Précisez si les estimateurs obtenus précédemment sont biaisés ou non. Sont-ils UMVUE ?

4°. Calculer les bornes de Cramer-Rao associées à θ et $1/\theta$. Discuter l'efficacité des estimateurs.

2.13 Paradoxe de Basu

Exercice inspiré de *Basu D.(1988) Stastical Information and Likelihood, Springer-Verlag, N.Y.*

Dans une urne contenant 1000 tickets, 20 sont marqués θ et 980 sont marqués 10θ, où θ est un nombre rationnel strictement positif.

1°. Donner l'estimateur du maximum de vraisemblance $\widehat{\theta}$ de θ lorsque l'on tire un unique ticket de valeur X, et montrer que $\mathbb{P}[\widehat{\theta} = \theta) = 0.98$. Expliquer pourquoi on ne pouvait supposer que θ était un réel quelconque pour calculer un estimateur du maximum de vraisemblance.

2°. On renumérote les tickets marqués 10θ par $a_i\theta$ $(1 \leq i \leq 980)$ où les a_i sont des nombres rationnels connus, deux à deux distincts, et compris dans l'intervalle $[10, 10.1]$. Donner le nouvel estimateur du maximum de vraisemblance $\widetilde{\theta}$ et montrer que $\mathbb{P}[\widetilde{\theta} < 10\theta] = 0.02$. Ce résultat vous semble-t-il paradoxal ?

2.14 Paramètres de position et d'échelle d'une loi exponentielle

Soit f la densité de la loi exponentielle de paramètre $\theta > 0$, translatée de $\alpha \in \mathbb{R}$,

$$f(x, \alpha, \theta) = \frac{1}{\theta} \exp\left[-\frac{x-\alpha}{\theta}\right] \mathbb{1}_{[\alpha, +\infty[}(x) \tag{79}$$

On considère un échantillon de n variables aléatoires i.i.d. de densité $f(x, \alpha, \theta)$, où $\theta > 0$ et $\alpha \in \mathbb{R}$ sont des paramètres inconnus.

1°. Donner l'e.m.v. $(\widehat{\alpha}_n, \widehat{\theta}_n)$ de (α, θ).

2°. Calculer la loi de $n(\widehat{\alpha}_n - \alpha)$, pour $n \in \mathbb{N}$.

3°. Déterminer la loi limite de $\sqrt{n}(\hat{\theta}_n - \theta)$.

4°. Rappeler l'expression de la loi de la statistique d'ordre $X_\bullet = (X_{(1)}, \dots, X_{(n)})$ en fonction de f. En déduire la loi du n-uplet

$$(X_{(1)}, X_{(2)} - X_{(1)}, \dots, X_{(n)} - X_{(n-1)}) \tag{80}$$

et en déduire que $\widehat{\theta}_n$ et $\widehat{\alpha}_n$ sont indépendants, pour $n \in \mathbb{N}$.

2.15 Loi de Weibull et modèle du taux de chômage

On souhaite évaluer et analyser le phénomène du chômage. Pour cela, on dispose de n observations sur les durées $y_i, 1 \leq i \leq n$, pendant lesquelles des individus sont restés sans emploi.

On suppose dans la suite que les variables aléatoires correspondantes $(Y_i)_{i \in [\![1,n]\!]}$ sont i.i.d. et suivent une loi de Weibull de paramètres a et b. On rappelle que cette loi est continue sur $\mathbb{R}_+$ et admet la fonction de répartition pour $y > 0$

$$F(y; a, b) = 1 - \exp\left(-a y^b\right) \tag{81}$$

On définit la fonction de survie par

$$S(y) = 1 - F(y) \tag{82}$$

1°. Donner l'expression de la fonction de hasard du modèle.

2°. Quelle est en terme de chômage l'interprétation de la fonction de hasard ? Expliquer alors pourquoi il est important de considérer le cas particulier où cette fonction est constante. Pour quelles valeurs des paramètres, la fonction de hasard est-elle constante ? Quelles sont alors les lois des durées de chômage ?

3°. Étudier l'évolution de la fonction de hasard en fonction de a, puis en fonction de b.

On suppose que $b = 1$. Le modèle est alors uniquement paramétré par a.

3°. Le modèle est-il exponentiel ? Si oui, expliciter une statistique exhaustive.

4°. Déterminer le vecteur du score et vérifier directement qu'il est centré.

5°. Quel est l'estimateur du maximum de vraisemblance $\widehat{a_0}$ de a ? Est-il sans biais, y a-t-il surestimation ou sous-estimation systématique ?

6°. Déterminer la variance asymptotique de cet estimateur $\widehat{a_0}$.

On considère maintenant le cas où a et b sont quelconques (positives).

7°. Le modèle est-il exponentiel avec une statistique exhaustive dont la taille est indépendante du nombre n d'observations ? Si oui, expliciter une telle statistique.

8°. Ecrire les équations de vraisemblance. Sont-elles résolubles sous forme analytique ?

9°. Donner la forme de la variance asymptotique de l'estimateur du maximum vraisemblance $(\widehat{a}, \widehat{b})'$ du paramètre $(a, b)'$.

10°. Comparer les estimateurs $\widehat{a}$ et $\widehat{a_0}$ lorsque $b = 1$. Quelle conclusion en tirer ?

On considère maintenant le cas de T observations $Y_1, \dots, Y_T$ indépendantes, de lois respectives :

$$F(y; e^{\alpha t}, 1)\ ,\ t \in [\![1, T]\!]\ ,\ \alpha \in \mathbb{R} \tag{83}$$

11°. Déterminer la vraisemblance du modèle, et vérifier qu'elle est concave en α à $(y_1, \dots, y_T)$ fixé. En déduire l'équation caractérisant l'estimateur du maximum de vraisemblance $\widehat{\alpha_T}$ de α.

12°. On note $u_t = y_t - e^{-\widehat{\alpha_T} t}$. Donner l'interprétation de u_t.

13°. Montrer que l'équation de la vraisemblance correspond à la condition d'orthogonalité de $(u_1, \dots, u_T)$ et de $1, \dots, T$ pour un certain produit scalaire que l'on précisera.

2.16 Virus et variables inobservées

On considère une population de n individus infectés par un virus ; on étudie leurs durées d'incubation $(T_i)_{i=1..n}$, dont on suppose qu'elle est observable. Pour modéliser l'hétérogénéité de la population, on suppose qu'on peut caractériser chaque individu i par un « facteur de risque » inobservable, réalisation de la variable aléatoire Λ_i, de telle sorte que :

- La loi de T_i, conditionnellement à Λ_i, est la loi exponentielle de paramètre Λ_i.
- Les variables $(\Lambda_i)_{i=1..n}$ sont identiquement distribués de loi $\Gamma(r, \alpha)$, avec $r > 2$.
- les couples (T_i, Λ_i) sont indépendants entre eux.

1°. Donner la vraisemblance de $(T_1, \dots, T_n)$.

2°. Calculer, lorsqu'il existe, le moment d'ordre k $\mathbb{E}[T_1^k]$.

3°. Calculer l'information de Fisher du modèle. Dans le cas où α est connu, calculer l'estimateur du maximum de vraisemblance de r. Que se passe-t-il si α et r sont tous deux inconnus ?

4°. On suppose α connu. Déterminer au moyen de la méthode des moments un estimateur convergent de r. Cet estimateur est-il sans biais ? Est-il asymptotiquement efficace ?

5°. On suppose α et r inconnus. En utilisant les deux premiers moments de T_i, trouver des estimateurs convergents $\tilde{\alpha}$ et $\tilde{r}$ de α et r. Donner la loi limite du vecteur

$$\sqrt{n}\begin{pmatrix} \overline{T} - \mathbb{E}[T] \\ \overline{T^2} - \mathbb{E}[T^2] \end{pmatrix} \tag{84}$$

En déduire la loi asymptotique du vecteur $(\tilde{\alpha}, \tilde{r})$.

2.17 Paramètres d'une loi de Laplace

Soit X une v.a. de densité

$$h(x) = k\exp\left(-\frac{|x-\lambda|}{\mu}\right) \tag{85}$$

où $\mu > 0$, $k > 0$, $\lambda \in \mathbb{R}$.

1°. Déterminer k et donner la fonction de répartition de X.

2°. On pose $Y = (X-\lambda)/\mu$. Déterminer la densité de Y, calculer $\mathbb{E}[X]$ et $\mathbb{V}(X)$.

3°. Soit $(X_1,...,X_n)$ un n-échantillon de X. Déterminer des estimateurs de λ et μ par la méthode du maximum de vraisemblance, puis par la méthode des moments. Étudier les propriétés de ces estimateurs.

4°. On suppose que $\lambda = 0$ et l'on pose $\sigma = 1/\mu$. σ est supposé aléatoire, de loi *a priori* $\gamma(1,\alpha)$. Si α est connu, déterminer un estimateur bayésien de σ. Dans le cas contraire, proposé un estimateur de α basé sur l'échantillon $(X_1,...,X_n)$.

2.18 Introduction à l'apprentissage supervisé.

On considère un n-échantillon de v.a.i.i.d. $\mathscr{D}_n = \{Z_1,...,Z_n\}$ avec $Z_i = (X_i, Y_i)$. Les X_i sont des observations issues d'une v.a. X, ce sont les données que l'on souhaite classer. Les Y_i sont issues d'une v.a. Y et sont les catégories auxquelles appartiennent les X_i (on dit également étiquettes ou labels). L'objectif de l'apprentissage supervisé est de déterminer au mieux la catégorie Y à laquelle appartient la donnée X correspondante, à partir des seules observations de l'échantillon $Z_1,...Z_n$.

On suppose que les v.a. X sont issues d'un espace $\mathbb{X}$, que les v.a. Y sont issues d'un espace $\mathbb{Y}$ et l'on se donne une loi de probabilité (inconnue) $\mathbb{P}$ sur l'espace $\mathscr{E} = \mathbb{X}\times\mathbb{Y}$.

Une fonction de prédiction est un élément $g \in \mathscr{F} = \mathscr{F}(\mathbb{X},\mathbb{Y})$ qui associe une étiquette à une observation. Pour mesurer la qualité de g, on définit différentes fonctions de perte $l : \mathbb{Y}^2 \longrightarrow \mathbb{R}_+$ telles que $l(Y, g(X))$ mesure l'écart entre la vraie valeur Y correspondant à X et la valeur $g(X)$ prédite à partir de la fonction g. Le risque de g est la valeur moyenne des réalisations de toutes les pertes possibles. Autrement dit,

$$R(g) = R_{\mathbb{P}}(g) = \mathbb{E}\left[l(Y, g(X)\right] \tag{86}$$

Le prédicteur de Bayes est l'élément $g^\star$ de $\mathscr{F}$ qui minimise la perte $R(g)$.

Dans cet exercice, nous nous limitons au problème de classification binaire, c'est à dire que Y ne peut prendre que deux valeurs : 0 ou 1. La fonction de perte naturelle associée est alors la fonction

$$l(Y,Y') = \mathbb{1}_{[Y\neq Y']} \tag{87}$$

On note enfin

$$\eta(x) = \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x] \tag{88}$$

Nous allons démontrer que

$$g^\star(x) = \mathbb{1}_{[\eta(x)>1/2]} \tag{89}$$

1°. Montrer que

$$\mathbb{P}[Y = g(X)|X = x] = \tag{90}$$
$$\eta(x)\mathbb{1}_{[g(x)=1]} + (1-\eta(x))\mathbb{1}_{[g(x)=0]} \tag{91}$$

2°. En déduire que

$$\mathbb{P}[Y \neq g^\star(X)|X = x] \leq \mathbb{P}[Y \neq g(X)|X = x] \tag{92}$$

et conclure.

3°. Montrer que le risque de Bayes $R^\star = R(g^\star)$ vérifie

$$R^\star = \mathbb{E}[\eta(X) \wedge (1-\eta(X))] \tag{93}$$
$$= \frac{1}{2}\left(1 - \mathbb{E}[|2\eta(X)-1|]\right) \tag{94}$$

avec $x \wedge y = \inf(x,y)$.

4°. Montrer de façon plus générale que quelque soit la fonction f de $\mathbb{X}$ dans $\mathbb{R}$, $\eta(X)$ minimise l'erreur quadratique lorsque $f(X)$ prédit Y. C'est à dire, montrer que

$$\mathbb{E}\left[(\eta(X)-Y)^2\right] \leq \mathbb{E}\left[(f(X)-Y)^2\right] \tag{95}$$

5°. On prédit la réussite d'un étudiant à un examen en fonction du nombre d'heures X passées à travailler. $Y = 1$ signifie que l'étudiant réussit son examen. On suppose que

$$\eta(x) = \frac{x}{x+c} \tag{96}$$

où $c > 0$. Si X suit une loi uniforme sur $[0, 4c]$, calculer $R^\star$.

2.19 Famille exponentielle sous forme naturelle

On considère un n-échantillon $X_1,...,X_n$ de X, v.a. de carré intégrable et de densité

$$f(x,\theta) = h(x)\exp(\theta x - \psi(\theta)) \tag{97}$$

où $\theta \in \Theta \subset \mathbb{R}$, $x \in \mathbb{X}$, $h(x) > 0$ et ψ de classe C^∞.

1°. Montrer que ψ vérifie

$$\psi(\theta) = \ln\left(\int_{\mathbb{X}} h(x)e^{\theta x}d\mu(x)\right) \tag{98}$$

où μ est une mesure dominante.

2°. Montrer que

$$\psi'(\theta) = \mathbb{E}_\theta[X] \text{ et } \psi''(\theta) = \mathbb{V}_\theta(X) \tag{99}$$

3°. Déduire de la question précédente que ψ' est strictement croissante, puis qu'elle est inversible.

4°. Démontrer que l'e.m.v. $\widehat{\theta}_n$ est un estimateur des moments, dont on donnera l'expression.

5°. En déduire une preuve directe de la convergence et de la normalité asymptotique de $\widehat{\theta}_n$.

6°. Retrouver le fait que l'e.m.v. est asymptotiquement efficace.

7°. Montrer qu'une loi *a priori* de la forme

$$\pi(\theta) = C(a,\lambda)\exp\big(a\theta - \lambda\psi(\theta)\big) \tag{100}$$

pour des hyperparamètres $a>0, \lambda>0$, est conjuguée pour le modèle.

8°. On suppose que X suit une loi de Poisson de paramètre ν. Montrer que le n-échantillon correspondant est bien un modèle exponentiel, déterminer l'e.m.v. de θ et en déduire l'e.m.v. de ν. Reconnaître la famille de loi *a priori*.

9°. On suppose que X suit une loi Gamma de paramètres α (connu) et β (inconnu) :

$$f(x,\beta) = \frac{\Gamma(\alpha)}{\beta^\alpha} x^\alpha e^{-\beta x} \mathbb{1}_{[0,\infty[}(x) \tag{101}$$

Calculer l'e.m.v. de β.

2.20 Loi binomiale négative

On considère un n-échantillon $(X_1,...,X_n)$ de N, variable aléatoire à valeurs entières de loi binomiale négative de paramètres $p\in[0,1]$ et $r\in\mathbb{N}^*$:

$$\mathbb{P}[N=k] = \binom{k+r-1}{k}(1-p)^r p^k \tag{102}$$

$k\in\mathbb{N}$. On rappelle que N mesure le nombre de lancers à pile ou face avec probabilité de faire pile égale à p, avant d'obtenir exactement r piles.

1°. Démontrer que

$$\mathbb{E}[N] = \frac{pr}{1-p} \text{ et } \mathbb{V}(N) = \frac{pr}{(1-p)^2} \tag{103}$$

On suppose r fixé et connu et l'on cherche à estimer p.

2°. Proposer un estimateur des moments de p et donner son comportement asymptotif.

3°. Montrer que le modèle est exponentiel.

4°. Calculer l'e.m.v. de p. Déduire des questions précédentes l'information de Fisher du modèle (sans la calculer).

5°. Montrer que l'e.m.v. est biaisé. Est-il néanmoins possible d'en déduire un estimateur sans biais ?

Montrer que la famille des lois Béta dont la densité est donnée par

$$f(t) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}\mathbb{1}_{]0,1[}(t) \tag{104}$$

pour $a>0, b>0$ est conjuguée pour ce modèle. Donner la loi *a posteriori* correspondant à un *a priori* de type Béta.

On cherche maintenant à estimer p et r.

6°. Peut-on estimer ce vecteur par maximum de vraisemblance ? Pourquoi ?

7°. On suppose maintenant que r est un réel positif et l'on rappelle que les factorielles peuvent s'étendre aux réels via la loi Gamma d'Euler en posant

$$\binom{k+r-1}{k} = \frac{\Gamma(k+r)}{k!\Gamma(r)} \tag{105}$$

Donner les équations de vraisemblance vérifiées par l'e.m.v. de (p,r). Admet-il une expression explicite ?

2.21 Échantillon de Bernoulli : estimateurs du carré du paramètre

On considère un n-échantillon $X=(X_1,...,X_n)$ d'une v.a. de loi de Bernoulli de paramètre inconnu $\theta\in]0,1[$.

On note $S_n = \sum_{i=1}^n X_i$.

1°. Expliquer succintement pourquoi le modèle est régulier.

2°. Calculer la log-vraisemblance $l(x,\theta)$ de l'échantillon et en déduire l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ de θ.

3°. Montrer que cet estimateur est fortement consistant et converge en moyenne quadratique. Démontrer que $\widehat{\theta}_n$ est asymptotiquement normal et que

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \rightsquigarrow \mathcal{N}(0,\theta(1-\theta)) \tag{106}$$

4°. On s'intéresse maintenant au paramètre $\lambda=\theta^2$. Démontrer que l'estimateur du maximum de vraisemblance de λ est $\widehat{\lambda}_n = (S_n/n)^2$.

5°. Démontrer que $\widehat{\lambda}_n$ est biaisé et calculer son biais.

6°. Démontrer que $\widehat{\lambda}_n$ est fortement consistant, asymptotiquement normal et préciser la loi limite de $\sqrt{n}(\widehat{\lambda}_n - \lambda)$.

7°. Calculer l'information de Fisher associée à λ, puis démontrer que $\widehat{\lambda}_n$ est asymptotiquement efficace.

8°. On cherche à construire un estimateur sans biais de λ. Pour tout $i=1,...,n$, on note $X_{\bullet i} = (X_1,...,X_{i-1},X_{i+1},...,X_n)$ le vecteur formé des $(n-1)$ observations différentes de la ième. On note $\widehat{\lambda}_n(i)$ l'estimateur du maximum de vraisemblance du modèle associé à $X_{\bullet i}$. Montrer que

$$\widehat{\lambda}_n(i) = \left(\frac{S_n - X_i}{n-1}\right)^2 \tag{107}$$

9°. Calculer $\mathbb{V}(S_n - X_i)$, en déduire $\mathbb{E}[(S_n-X_i)^2]$, puis démontrer que l'estimateur

$$T_n = n\widehat{\lambda}_n - \frac{n-1}{n}\sum_{i=1}^n \widehat{\lambda}_n(i) \tag{108}$$

est sans biais pour λ.

2.22 Échantillon de Bernoulli : estimateurs bayésiens du paramètre et de son carré

On considère un n-échantillon $X=(X_1,...,X_n)$ d'une variable aléatoire de loi de Bernoulli de paramètre

$\theta \in]0,1[$. On note $S_n = \sum_{i=1}^n X_i$ et $\overline{X} = S_n/n$ la moyenne empirique de l'échantillon.

1°. Déterminer la vraisemblance du n-échantillon $X_1,...,X_n$. Le modèle est-il exponentiel ?

2°. Déterminer un estimateur $\widehat{\theta}_n$ de θ par la méthode du maximum de vraisemblance, puis par la méthode des moments. Que constatez-vous ?

3°. Montrer que cet estimateur est fortement consistant et converge en moyenne quadratique. Démontrer que $\widehat{\theta}_n$ est asymptotiquement normal et déterminer la variance de la loi limite.

4°. Déterminer un intervalle de confiance asymptotique pour θ, de niveau $1-\alpha$.

5°. $\widehat{\theta}_n$ est-il exhaustif ? Minimal ? Complet ?

6°. Le modèle est-il régulier ?

7°. Démontrer que $\widehat{\theta}_n$ est un estimateur VUMSB (sans biais de variance minimale) de θ.

8°. Calculer l'information de Fisher relative à θ et montrer que $\widehat{\theta}_n$ est un estimateur efficace de θ.

On considère que la loi *a priori* de θ est une loi bêta $\mathscr{B}(a,b)$, avec $a,b \in [0,1]$, dont la densité est donnée par

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \mathbb{1}_{[0,1]}(\theta) \tag{109}$$

où Γ est la fonction Gamma d'Euler donnée par

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx \tag{110}$$

On rappelle que l'espérance d'une variable aléatoire H de loi bêta est donnée par

$$\mathbb{E}[H] = \frac{a}{a+b} \tag{111}$$

9°. Déterminer la loi *a posteriori* de θ sachant $[X = x]$. En déduire que l'estimateur de la moyenne *a posteriori* peut s'exprimer au travers de la statistique S_n. Montrer que cet estimateur est biaisé, mais asymptotiquement sans biais.

10°. On suppose $n > 2$. On souhaite estimer la valeur de θ^2. Démontrer que $Y = X_1X_2$ est un estimateur sans biais de θ^2. Démontrer que $T_n = \mathbb{E}[Y|S_n]$ est un estimateur sans biais de variance minimale pour θ^2.

11°. On se propose de calculer explicitement l'expression de T_n. Calculer $\mathbb{P}[Y = 1|S_n = s]$ (on distinguera les cas $s < 2$ et $s \geq 2$). En déduire que

$$T_n = \frac{S_n(S_n-1)}{n(n-1)} \mathbb{1}_{[S_n \geq 2]}. \tag{112}$$

2.23 Loi expontielle anglo-saxonne : estimation non biaisée du paramètre d'échelle

On considère un n-échantillon $X = (X_1,...,X_n)$ d'une v.a. de loi exponentielle de paramètre inconnu $\theta > 0$.

On note $S = \sum_{i=1}^n X_i$ et $\overline{X} = S/n$ la moyenne empirique de l'échantillon.

1°. Déterminer l'estimateur des moments $\widehat{\theta}$ de θ.

2°. Calculer la log-vraisemblance $l(x,\theta)$ de l'échantillon et en déduire l'estimateur du maximum de vraisemblance $\widehat{\theta}_{MV}$ de θ, en fonction de $\overline{X}$.

3°. Montrer que cet estimateur est fortement consistant.

4°. À l'aide des rappels, démontrer que

$$\mathbb{E}\left[\frac{1}{S}\right] = \frac{\theta}{n-1} \tag{113}$$

et en déduire $\mathbb{E}[\widehat{\theta}_{MV}]$. Déduire que l'estimateur $T_n = (n-1)/S$ est non biasé.

5°. En calculant $\mathbb{E}[1/S^2]$ comme précédemment, démontrer que

$$\mathbb{V}(T_n) = \frac{\theta^2}{n-2} \tag{114}$$

En déduire que T_n converge en moyenne quadratique.

6°. Démontrer que $\widehat{\theta}_{MV}$ est asymptotiquement normal et que

$$\sqrt{n}\left(\widehat{\theta}_{MV} - \theta\right) \rightsquigarrow \mathscr{N}(0,\theta^2) \tag{115}$$

7°. Déterminer une statistique exhaustive et complète pour θ puis en en déduire que T_n est VUMSB pour θ.

8°. Expliquez pourquoi le modèle est régulier (on ne justifiera que les propriétés évidentes sans calcul) et calculer l'information de Fisher $\mathbb{I}_X(\theta)$ associée à θ pour l'échantillon $X = (X_1,...,X_n)$.

9°. T_n est-il un estimateur efficace ? Asymptotiquement efficace ?

10°. On suppose maintenant que θ suit une loi *a priori* $\Gamma(\alpha,\beta)$. Calculer la vraisemblance *a posteriori* du modèle et reconnaître la loi de θ sachant les observations. À partir de la moyenne *a posteriori* $\mathbb{E}[\Pi|X]$ déterminer un estimateur bayésien de θ. Quelle est sa limite lorsque n tend vers l'infini ?

2.24 Échantillon de Bernoulli : estimation bayésienne des puissances du paramètre

On considère un n-échantillon $X = (X_1,...,X_n)$ d'une v.a. de loi de Bernoulli de paramètre $\theta \in]0,1[$. On note $\overline{X}$ la moyenne empirique de l'échantillon.

1°. Déterminer la vraisemblance du n-échantillon $X_1,...,X_n$. Le modèle est-il exponentiel ? (2 points).

2°. Déterminer un estimateur $\widehat{\theta}$ de θ par la méthode du maximum de vraisemblance, puis par la méthode des moments. Que constatez-vous ? (2 points).

3°. Montrer que $\widehat{\theta}$ est fortement convergent, sans biais et asymptotiquement normal. (3 points).

4°. Est-il exhaustif ? Minimal ? Complet ? (1 point).

5°. Le modèle est-il régulier ? (1 point).

6°. Démontrer que $\overline{X}$ est un estimateur VUMSB (sans biais de variance minimale) de θ. (2 points).

7°. Calculer l'information de Fisher relative à θ (1 point).

8°. Montrer que $\overline{X}$ est un estimateur efficace de θ. (1 point).

9°. On considère que la loi *a priori* de θ est une loi bêta $\mathscr{B}(a,b)$, avec $a,b \in [0,1]$, dont la densité est donnée par

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{[0,1]}(\theta) \tag{116}$$

où Γ est la fonction Gamma d'Euler donnée par

$$\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx \tag{117}$$

On rappelle que l'espérance d'une v.a. H de loi bêta est donnée par

$$\mathbb{E}[H] = \frac{a}{a+b} \tag{118}$$

Déterminer la loi *a posteriori* de θ sachant $[X = x]$. En déduire que l'estimateur de la moyenne *a posteriori* peut s'exprimer au travers de la statistique $S = \sum_{i=1}^{n} X_i$. Montrer que cet estimateur est biaisé, mais asymptotiquement sans biais. (2 points).

10°. On souhaite estimer, pour un entier k tel que $1 < k \leq n$, la valeur de θ^k. Démontrer que

$$Y_k = \prod_{i=1}^{k} X_i \tag{119}$$

est un estimateur sans biais de θ^k. Exprimer, sous la forme d'une espérance conditionnelle, un estimateur $T = T_{k,n}(X)$ sans biais de variance minimale pour θ^k. (3 points).

11°. Démontrer que

$$T_{k,n}(X) = \frac{\begin{pmatrix} n-k \\ S-k \end{pmatrix}}{\begin{pmatrix} n \\ S \end{pmatrix}}\mathbb{1}_{[S\geq k]} \tag{120}$$

$\begin{pmatrix} n \\ k \end{pmatrix}$ représente le nombre de combinaisons de k éléments parmi n. (2 points).

2.25 Modèle de Hardy-Weinberg

Un modèle génétique des états attribue aux trois génotypes *aa* (état 1), A*a* (état 2) et AA (état 3) les probabilités suivantes d'apparaître :

$$\begin{cases} p_1 = (1-\theta)^2 \\ p_2 = 2\theta(1-\theta) \\ p_3 = \theta^2 \end{cases} \tag{121}$$

où $\theta \in [0,1]$ est la proportion de l'allèle A dans la population. On considère un échantillon de n d'individus dans la population. On dénombre les effectifs X_1, X_2 et X_3 des génotypes dans l'échantillon.

1°. Montrer que (p_1, p_2, p_3) définit une mesure de probabilité sur $\mathbb{X} = \{1,2,3\}$.

2°. On suppose que $X = (X_1, X_2, X_3)$ suit une loi multinomiale $\mathscr{M}(n, p_1, p_2, p_3)$. Décrire le modèle statistique considéré.

3°. Quel est le modèle image par X_1 ? En déduire le modèle image par X_2, puis X_3. Montrer que X_2 et X_3 ne sont pas indépendantes. Le modèle est-il dominé ?

4°. Montrer que l'estimateur du maximum de vraisemblance de θ est

$$\widehat{\theta} = \frac{X_2 + 2X_3}{2n} \tag{122}$$

5°. Cet estimateur est-il de variance uniformément minimale parmi les estimateurs sans biais ? Est-il efficace ?

6°. Déterminer l'estimateur du maximum de vraisemblance de $p = (p_1, p_2, p_3)$. Est-il sans biais ?

2.26 Loi de Paréto et estimateur de Hill

Économiste et sociologiste italien né à Paris en 1848, Vilfredo Pareto est à l'origine de la loi de probabilité que nous allons présenter dans cet exercice. Alors titulaire de la chaire d'économie politique de l'université de Lausanne (il succède à Léon Walras), Pareto s'intéresse à la distribution et à la répartition des revenus dans les différents pays d'Europe.

Disposant des données fiscales pour la France, l'Angleterre, la Suisse, l'Italie, la Russie et la Prusse, il remarque que les inégalités de revenus varient fortement d'un pays à l'autre, mais il met également en lumière une régularité statistique remarquable, vérifiée dans tous les pays pour lesquels il dispose de données. Dans son « essai sur la courbe de la répartition de la richesse » publié en 1896, il écrit : « nous indiquerons par x un certain revenu, et par N le nombre de contribuables ayant un revenu supérieur à x (...). Traçons deux axes (AB) et (AC). Sur (AB) portons les logarithmes de x, sur (AC) les logarithmes de N. Il ressort une relation tout à fait linéaire. » De ce constat empirique, l'auteur en déduit la relation mathématique suivante :

$$\log(N) = B - \alpha \times \log(x) \Leftrightarrow N = \frac{A}{x^\alpha} \tag{123}$$

Avec $B = \log(A)$. Finalement, selon Pareto, le pourcentage de la population dont la richesse est supérieure à une valeur x est toujours proportionnelle à $A \div x^\alpha$. C'est le paramètre α qui varie entre les différents pays et explique des différences dans la distribution des revenus.

Aujourd'hui, la loi de Pareto est encore couramment utilisée en économie ou en sociologie pour étudier les inégalités de revenus dans nos sociétés. Elle a également fait l'objet de multiples applications en gestion des risques, actuariat, dans le domaine du management des entreprises ou dans la gestion des flux de données sur internet.

La densité d'une loi de Pareto $\mathscr{P}(\alpha, c)$ est donnée par

$$f(x) = \frac{\alpha c^\alpha}{x^{\alpha+1}}\mathbb{1}_{[c,+\infty[}(x) \tag{124}$$

1°. Déterminer sa fonction de répartition.

2°. Calculer $\mathbb{E}[X]$ et $\mathbb{V}(X)$.

3°. Soit $X = (X_1, ..., X_n)$ un n-échantillon de Pareto. La notation $\rightsquigarrow$ signifie « converge en loi lorsque n tend vers l'infini ». On pose

$$F(x) = \mathbb{P}[X_i \leq x] \tag{125}$$

$\forall\ i = 1, ..., n$. Si $M_n = \max(X_1, ..., X_n)$ et si F_n est la fonction de répartition de M_n, déterminer le lien entre F_n et F.

4°. Le théorème de Fisher-Tippett assure qu'il existe deux suites m_n et $\sigma_n > 0$ telles que

$$\frac{M_n - m_n}{\sigma_n} \rightsquigarrow H \tag{126}$$

Où H suit une distribution de Gumbel, de Weibull ou de Fréchet.

Démontrer que la loi de Pareto est dans le domaine d'attraction de la loi de Fréchet, qui caractérise les distributions à queues épaisses. Précisément, montrer, en posant $m_n = 0$ et $\sigma_n = (nc)^{1/\alpha}$ que

$$F_n(x\sigma_n + m_n) \rightsquigarrow \phi \tag{127}$$

avec

$$\phi(x) = \exp(-x^{-\alpha})\mathbb{1}_{]0,+\infty[}(x) \tag{128}$$

ϕ est la fonction de répartition d'une loi de Fréchet.

On peut paramétrer cette loi limite à l'aide d'un paramètre γ qui est l'indice de la loi et caractérise l'épaisseur de sa queue :

$$H(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right) \tag{129}$$

$\gamma = 1/\alpha > 0$ et $1 + \gamma x > 0$.

Finalement, l'indice γ de la loi extrême relative à une loi de Pareto est exactement le paramètre $1/\alpha$.

5°. Démontrer que l'estimateur de c par la méthode du maximum de vraisemblance est

$$\widehat{c} = \min_{i=1}^{n} X_i \tag{130}$$

et déterminer sa loi.

6°. On suppose maintenant c connu. Démontrer que l'estimateur du maximum de vraisemblance de α est

$$\widehat{\alpha} = \widehat{\alpha}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \ln \frac{X_i}{c}\right)^{-1} \tag{131}$$

7°. Montrer que $Y_i = \ln(X_i/c)$ suit une loi exponentielle de paramètre $1/\alpha$.

8°. Déterminer la loi de

$$T = \sum_{i=1}^{n} \ln \frac{X_i}{c} \tag{132}$$

9°. Déterminer l'espérance et la variance de $\widehat{\alpha}$ et en déduire un estimateur sans biais $\alpha^\star$ de α. Calculer sa variance.

10°. Montrer que T est une statistique exhaustive et que cette statistique est complète. En déduire que $\alpha^\star$ est l'estimateur VUMSB de α.

11°. Montrer que $\widehat{\alpha}$ et $\alpha^\star$ sont des estimateurs consistants de α. Déterminer la loi limite de $\sqrt{n}(\widehat{\alpha}_n - \alpha)$ et $\sqrt{n}(\alpha_n^\star - \alpha)$.

12°. Montrer que l'estimateur des moments de α (lorsque $\alpha > 2$) est

$$\overline{\alpha} = \overline{\alpha}_n = \frac{\overline{X}}{\overline{X} - c} \tag{133}$$

où $\overline{X}$ est la moyenne empirique de l'échantillon. Calculer la loi limite de $\sqrt{n}(\overline{\alpha}_n - \alpha)$.

On va donc supposer à partir de maintenant et sans perte de généralité, que $c = 1$. Montrer que la fonction de survie de X est donnée, pour $x \geq 1$, par :

$$\overline{F}(x) = x^{-\alpha} \tag{134}$$

13°. $X = (X_1, ..., X_n)$ étant toujours un échantillon de loi de Pareto, notons $X_\bullet = (X_{(1)}, ..., X_{(n)})$ sa statistique d'ordre, avec $\max X_i = X_{(n)} > ... > X_{(1)} = \min X_i$.

Pour un entier $k \in [\![1..n-1]\!]$, on définit l'estimateur de Hill par

$$H_k = \frac{1}{k}\sum_{i=1}^{k} \ln \frac{X_{(n-i+1)}}{X_{(n-k)}} \tag{135}$$

Il est défini à partir des $k+1$ valeurs les plus élevées de l'échantillon et dépend également de n via la valeur de $k = k_n$ qui doit être estimée. Formellement, on se donne un seuil élevé $s = X_{(n-k)}$ et on ne garde que les k plus grandes valeurs qui excèdent ce seuil. Lorsque $k = n$, que s est fixe, déterministe et égal à c, on retrouve l'expression de l'estimateur du maximum de vraisemblance donnée à la section précédente.

L'estimation de l'indice de la queue d'une distribution dans le domaine d'attraction de la loi de Fréchet est à la source de nombreuses publications. Le problème est alors l'estimation de $\gamma = 1/\alpha$ pour une loi vérifiant

$$\overline{F}(x) = x^{-1/\gamma} L(x) \tag{136}$$

où $L(x)$ est une fonction à variation lente vérifiant, pour tout $x > 0$,

$$\lim_{t\to+\infty} \frac{L(tx)}{L(x)} = 1 \tag{137}$$

Le cas qui nous intéresse est celui d'une loi de Pareto stricte. On a alors $L(x) = 1$ qui correspond au cas le plus simple de fonction à variation (très) lente. De façon générale, on peut considérer des distributions pour lesquelles, par exemple, $L(x) = \ln x$. Dès que la distribution s'éloigne de la loi de Pareto stricte, les propriétés de l'estimateur de Hill ne sont plus les mêmes et des biais apparaissent. Dans la suite, nous présenterons à cet effet le « Hill horror plot » qui visualise l'augmentation du biais au fur et à mesure que le nombre d'observations retenues augmente.

Lorsque la suite d'entiers $k_n \to +\infty$ et $k_n = o(n)$ alors H_k est un estimateur faiblement consistant de $\gamma = \alpha^{-1}$. Si la convergence de k_n vers l'infini est suffisamment lente (par exemple si $k_n = \lfloor n^\nu \rfloor$ avec $0 < \nu < 1$ ou si $k_n / \ln\ln n \to +\infty$) alors la consistance forte est également acquise.

Par ailleurs, sous certaines conditions supplémentaires appelées « hypothèses de variations régulières au second ordre », l'estimateur est asymptotiquement normal :

$$\sqrt{k_n}\left(H_{k_n} - \gamma\right) \rightsquigarrow \mathcal{N}(0, \gamma^2) \tag{138}$$

Ces conditions suffisantes portent à la fois sur la forme de la fonction à variation lente $L(x)$ et sur la suite k_n. L'une des conditions, appelée condition de Von Mises impose à L de vérifier

$$\lim_{x\to+\infty} \frac{xF'(x)}{1-F(x)} = \alpha \tag{139}$$

Cette condition est vérifiée, par exemple, lorsque F appartient à la classe de Hall, c'est à dire lorsque

$$\overline{F}(x) = cx^{-\alpha}\left(1 + bx^{-\beta} + o\left(x^{-\beta}\right)\right) \tag{140}$$

avec $c, \alpha, \beta > 0$.

Pour une loi de Pareto stricte, la condition de Von Mises, est vérifiée pour tout x, et pas seulement à la limite (b et le reste sont exactement nuls) et la loi de Pareto appartient à la classe de Hall. Mais surtout, l'estimateur de Hill associé est sans biais pour toute valeur de k (et donc également asymptotiquement). Pour le démontrer, on note

$$\ln X_{(i)} = \gamma Y_{(i)} \tag{141}$$

En considérant le vecteur $Y = (Y_{(1)}, ..., Y_{(k+1)})$, montrer que le vecteur des écarts entre deux coordonnées successives $Y_{(i)} - Y_{(i-1)}$ possède des coordonnées indépendantes qui suivent une loi exponentielle. En utilisant la décomposition de Rényi des statistiques d'ordre exponentielles, montrer que

$$Y_{(i)} = \sum_{j=1}^{i} \frac{Y_j}{n-j+1} \tag{142}$$

En déduire que

$$H_k = \frac{\gamma}{k}\sum_{i=1}^{k}\left(Y_{(n-i+1)} - Y_{(n-k)}\right) = \frac{\gamma}{k}\sum_{i=1}^{k}\sum_{j=i}^{k}\frac{Y_j}{j} \tag{143}$$

Puis démontrer que

$$\Rightarrow H_k = \frac{\gamma}{k}\sum_{j=1}^{k} Y_j \sim \frac{\gamma}{k}\Gamma(k,1) \tag{144}$$

Conclure que

$$\mathbb{E}[H_k] = \gamma \tag{145}$$

Montrer que H_k converge presque sûrement vers γ et déterminer la loi limite de $\sqrt{k}(H_k - \gamma)$.

3 Estimation par régions de confiance

3.1

On souhaite estimer la durée de vie d'un composant électronique. On dispose d'un échantillon des durées de vie de 10 composants. La moyenne empirique de cet échantillon est de $\overline{x} = 1.3$ années, et la variance empirique non biaisée est de $s^2 = 0.0796$.

1°. En considérant que la durée de vie d'un composant suit une loi normale $\mathcal{N}(m, \sigma^2)$, déterminer un intervalle de confiance de niveau $1-\alpha = 0.99$ de la valeur moyenne m de la durée de vie de ce composant.

2°. Déterminez un intervalle de confiance de σ^2 au risque 0.9.

3°. On considère le modèle bayésien $X_1, ..., X_n$ de v.a.i.i.d. $\sim \mathcal{N}(m, s^2)$ où m est un paramètre dont la loi *a priori* est $\mathcal{N}(1,1)$. Construisez un intervalle de crédibilité de niveau 0.99 pour m.

3.2

Deux candidats A et B se présentent à un scrutin. Tous les inscrits votent et le bulletin blanc est proscrit. Un sondage sur $n = 100$ personnes est alors effectué afin d'anticiper l'issue du vote. Sur les 100 personnes, $x = 60$ pensent (et disent qu'ils vont) voter pour A. On veut inférer le pourcentage de votants pour A.

1°. Proposer un modèle statistique adapté.

2°. Donner un intervalle de confiance de niveau 0.99 sur le pourcentage de votant pour le candidat A. Cet intervalle est-il informatif pour vonnaître l'issue du vote ?

3°. Quelle serait la taille minimale de l'échantillon pour qu'un intervalle de confiance de niveau 0.95 soit de longueur inférieure à 0.6 ?

3.3

On considère le modèle d'échantillonnage $X_1, ..., X_n$, v.a.i.i.d. $\sim \mathcal{P}(\lambda)$, $\lambda > 0$.

1°. Étudiez la vitesse de convergence de l'e.m.v. Déduisez-en un intervalle de confiance asymptotique de niveau $1-\alpha$ pour λ.

2°. On considère maintenant le modèle bayésien d'un n-échantillon $X_1, ..., X_n$ de loi de Poisson $\mathcal{P}(\lambda)$ où λ suit une loi *a priori* $\gamma(a, b)$. Construire un intervalle de crédibilité pour λ.

3°. Application numérique : $n = 10$, $y = 6$, $\alpha = 0.1$, $a = b = 1$. Comparer les deux procédures.

3.4

On considère le modèle d'échantillonnage $X_1, ..., X_n$ de v.a.i.i.d. dont la densité est

$$f(x, \theta) = e^{-x+\theta}\mathbb{1}_{]0,\infty[}(x) \tag{146}$$

$\theta \in \mathbb{R}$.

1°. Montrer que θ est un paramètre de position pour $\sum X_i$ et en déduire un pivot dont on précisera la loi. À partir de ce pivot, construisez un intervalle de confiance de niveau α pour θ.

2°. Même question pour $T = X_{(1)}$, minimum de l'échantillon.

3°. Quel intervalle de confiance est meilleur ?

3.5

Au sein d'une population donnée, on s'intéresse à la probabilité p d'être contaminé par une personne contagieuse. On note $q = 1 - p$ et $\epsilon > 0$. On considère un n-échantillon $(Y_1, ..., Y_n)$ d'une v.a. de Bernoulli Y de paramètre p. On note $\overline{Y}_n$ la moyenne empirique de l'échantillon.

1°. Démonter l'inégalité de Tchebychev.

2°. Monter que $\overline{Y}_n$ est un estimateur sans biais de p et déterminer son risque quadratique.

3°. Monter que $\left[\overline{Y}_n - \sqrt{5/n}, \overline{Y}_n + \sqrt{5/n}\right]$ est un intervalle de confiance de p de niveau $0,95$.

4°. Soit $\theta > 0$. Établir que :

$$\mathbb{P}[\overline{Y}_n - p \geq \epsilon] = \mathbb{P}[e^{n\theta\overline{Y}_n} \geq e^{n\theta(p+\epsilon)}] \tag{147}$$

5°. Soit g la fonction définie sur $[0, \infty[$ par

$$g(x) = \ln\left(pe^x + q\right) \tag{148}$$

Démontrer que

$$\mathbb{P}[\overline{Y}_n - p \geq \epsilon] \leq e^{ng(\theta) - \theta(p+\epsilon)} \tag{149}$$

6°. Montrer que g est de classe C^2 et vérifier pour tout $x \geq 0$, l'inégalité $|g''(x)| \leq 1/4$. En déduire que

$$g(\theta) \leq \theta p + \theta^2/8 \tag{150}$$

7°. À l'aide des questions précédentes, établir l'inégalité

$$\mathbb{P}[\overline{Y}_n - p \geq \epsilon] \leq e^{-2n\epsilon^2} \tag{151}$$

8°. On pose

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i) \tag{152}$$

Majorer $\mathbb{P}\left[|\overline{W}_n - q| \geq \epsilon\right]$.

9°. Déduire des questions précédentes l'inégalité

$$\mathbb{P}[|\overline{W}_n - p| \geq \epsilon] \leq e^{-2n\epsilon^2} \tag{153}$$

10°. Pour $\epsilon = \sqrt{1,844n}$, en déduire une nouvel intervalle de confiance au niveau $0,95$ et le comparer à l'intervalle de confiance obtenu précédemment. Conclure.

3.6

Soit $(X_1, ..., X_n)$ un n-échantillon d'une variable aléatoire X de densité

$$g(x) = \frac{1}{2}(1 + \theta x)\mathbb{1}_{]-1,1[}(x) \tag{154}$$

$\theta \in]-1, 1[$.

1°. Construire un estimateur $\widehat{\theta}$ de θ en utilisant la méthode des moments. Calculer son biais et son risque quadratique moyen.

2°. Déterminer la loi limite de $\widehat{\theta}$ et en déduire un intervalle de confiance asymptotique pour θ au niveau $1 - \alpha$.

3.7

Soit $(X_1, ..., X_n)$ un n-échantillon d'une variable aléatoire X de densité exponentielle de paramètre λ.

1°. Montrer que $M_n = \lambda \max_{i=1}^{n} X_i$ est une variable pivotale pour λ. En déduire un intervalle de confiance au niveau $1 - \alpha$ pour ce paramètre.

2°. Construire un intervalle de confiance pour λ au niveau $1 - \alpha$ en utilisant la moyenne empirique $\overline{X}$.

3°. Déterminer la loi limite et la vitesse de convergence vers cette loi de l'estimateur $1/\overline{X}$. En déduire un intervalle de confiance asymptotique pour le paramètre λ , au niveau $1 - \alpha$.

3.8

Soit $(X_1, ..., X_n)$ un n-échantillon d'une variable aléatoire X dont la densité est

$$g(x) = \frac{x \ln\theta}{\theta^{x^2/2}}\mathbb{1}_{[0,+\infty[}(x) \tag{155}$$

où $\theta > 1$. On pose également

$$\widehat{\theta} = \exp\left(\frac{2n}{\sum_{i=1}^{n} X_i^2}\right) \tag{156}$$

1°. Calculer les moments d'ordre 2 et 4 de la loi de X.

2°. Montrer que $\widehat{\theta}$ est biaisé et consistant.

3°. Déterminer sa loi limite et la vitesse de convergence vers cette loi.

4°. Déterminer un intervalle de confiance asymptotique pour θ au niveau $1 - \alpha$.

3.9

Comme test de fiabilité de composants électroniques, on effectue n tirages indépendants avec remise, dans différents lots, jusqu'à sélectionner un composant défectueux. On passe alors au lot suivant. On cherche à estimer le pourcentage de composants défectueux.

1°. Déterminer le modèle statistique associé et construire un estimateur de la proportion de composants défectueux. En déduire un intervalle de confiance

asymptotique (sur le nombre d'expériences) pour la proportion de composantes défectueux.

2°. Donner un estimtateur sans biais du nombre moyen de composants tirés au cours des n expériences. En déduire un intervalle de confiance asymptotique (sur le nombre d'expériences) pour le nombre moyen de composants tirés.

3.10

Soit $(X_1,...,X_n)$ un n-échantillon de loi exponentielle de moyenne λ.

1°. Déterminer un pivot fondé sur la statistique $S = \sum_{i=1}^{n} X_i$.

2°. En déduire un intervalle de confiance de niveau $1-\alpha$ pour λ.

3.11

Soient $(X_1,...,X_n)$ et $(Y_1,...,Y_m)$ deux échantillons de v.a.i.i.d. indépendants entre eux, de lois normales respectives $\mathcal{N}(m_1,\sigma_1^2)$ et $\mathcal{N}(m_2,\sigma_2^2)$. On souhaite estimer le rapport des variances $r = \sigma_1^2/\sigma_2^2$, les moyennes étant également inconnues.

1°. Donner un pivot suivant une loi de Fisher.

2°. En déduire un intervalle de confiance au niveau $1-\alpha$ pour r.

3.12

On considère une v.a. X de densité

$$f(x,\theta) = \theta x^{\theta-1}\mathbb{1}_{]0,1[}(x) \tag{157}$$

1°. Construire un intervalle de confiance pour $\theta > 0$ en utilisant un pivot calculé à partir de la fonction de répartition de X.

2°. Montrer que l'intervalle le plus petit de niveau $1-\alpha$ a nécessairement une borne inférieure nulle.

3.13

On considère une variable aléatoire X de loi de Poisson de paramètre λ, une variable aléatoire Y suivant une loi gamma $\Gamma(y+1,z)$ et une variable aléatoire Z suivant une loi du χ^2 à $2(y+1)$ degrés de liberté.

1°. Démontrer que

$$\mathbb{P}_X[0,y] = \mathbb{P}_Z[\lambda z,+\infty[\tag{158}$$

en déduire que

$$\mathbb{P}_X[0,y] = \mathbb{P}_Z[2\lambda,+\infty[\tag{159}$$

2°. Soit $(X_1,...,X_n)$ un n-échantillon de X dont le paramètre λ est inconnu. Déterminer un intervalle de confiance de niveau $1-\alpha$ pour ce paramètre en utilisant un pivot fondé sur la statistique $S = \sum_{i=1}^{n} X_i$.

3°. Montrer qu'un intervalle $(1-\alpha)$ crédible pour λ peut-être aussi fondé sur les quantiles d'une loi du χ^2 lorsque l'on prend pour loi *a priori* la loi exponentielle de paramètre 1.

4°. A.N. au cas où $n = 50$, S a pour valeur observée $s = 5$ et $\alpha = 0.1$.

5°. Soient $f(x,d)$ la densité d'une loi du χ^2 à d degrés de liberté. Montrer que pour tout $a > 0$, si $p > q$ alors $f(a,p) > f(a,q)$.

6°. En déduire que si $p > q$, alors les quantiles supérieurs d'ordre α de ces deux lois sont dans le même ordre.

7°. Comparer les longueurs des intervalles de confiance et des intervalles de crédibilité correspondants.

3.14

Soit X une v.a. de Bernoulli de paramètre $p \in]0,1[$. On considère une suite de v.a.i.i.d. $(X_n)_n$ de même loi que X. On note également $S_n = \sum_{k=0}^{n} X_k$.

1°. Soit $\phi(s) = \mathbb{E}[e^{sX}]$ la fonction génératrice de X. Calculer ϕ en justifiant de son existence.

2°. Déterminer la loi de S_N/N. Pour un réel s, montrer que

$$\mathbb{E}\left[\exp\left(s\frac{S_N}{N}\right)\right] = \left(\phi\left(\frac{s}{N}\right)\right)^N \tag{160}$$

3°. Soit $a \in]0,1[$. Soi $s > 0$. Montrer que

$$\mathbb{E}\left[\exp\left(s\frac{S_N}{N}\right)\right] \geq e^{as} \times \mathbb{P}\left[\frac{S_N}{N} \geq a\right] \tag{161}$$

4°. Montrer que, pour tout $s \geq 0$,

$$\mathbb{P}\left[\frac{S_N}{N} \geq a\right] \leq \left(\phi\left(\frac{s}{N}\right)\right)^N e^{-as} \tag{162}$$

5°. On suppose que $a > p$. Étudier les variations de $l_a(s) = as - \ln\phi(s)$ et donner la valeur du maximum strictement positif $h(a,p)$. Montrer que

$$\mathbb{P}\left[\frac{S_N}{N} \geq a\right] \leq e^{-Nh(a,p)} \tag{163}$$

6°. On suppose que $a < p$. Déterminer la loi de la v.a. $N - S_N$. Montrer que

$$\mathbb{P}\left[\frac{S_N}{N} \leq a\right] \leq e^{-Nh(1-a,1-p)} = e^{-Nh(a,p)} \tag{164}$$

7°. Soit $\epsilon > 0$. Déduire des questions précédentes que

8°. Déterminer

$$\lim_{N\to+\infty} \mathbb{P}\left[\left|\frac{S_N}{N} - p\right| \geq \epsilon\right] \tag{165}$$

9°. On effectue un test de fiabilité pour des composants dont la probabilité d'être défectueux est p. On prélève un échantillon de $n \geq 1$ composants. Pour tout $i = 1,...,n$, on définit X_i qui vaut 1 lorsque l'objet est défectueux et 0 sinon. On suppose que les X_i sont indépendants. Montrer que $F_N = S_N/N$ est un estimateur sans biais de p. Calculer le risque quadratique

$$r_N = \mathbb{E}\left[(F_N - p)^2\right] \tag{166}$$

et déterminer $\lim_{N\to+\infty} r_N$.

10°. Soit $\alpha \in]0,1[$. On souhaite définir un intervalle de confiance du paramètre p inconnu, au niveau de confiance $1-\alpha$, à partir de l'échantillon. Quelle est la limite en loi de la suite

$$\left(\sqrt{N}\frac{F_n - p}{\sqrt{p(1-p)}}\right)_{n>0} \tag{167}$$

Soit f_n la réalisation de F_N sur l'échantillon considéré. Soit t le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite. Montrer qu'un intervalle de confiance de p au niveau $1-\alpha$ est donné par $[U_N, V_N]$, avec
vskip-5mm

$$U_N = f_N - \frac{t}{2\sqrt{N}} \text{ et } V_N = f_N + \frac{t}{2\sqrt{N}} \tag{168}$$

3.15 Inégalité de Hoeffding

Dans cet exercice, n désigne un entier naturel non nul. Soient a et b deux réels tels que $a < b$. Toutes les variables aléatoires sont définies sur un même espace probabilisé $(\Omega, \mathscr{F}, \mathbb{P})$. Soit X une variable aléatoire à valeurs dans $[a,b]$ et soit $(X_1, ..., X_n)$ un n-uplet de variables aléatoires réelles mutuellement indépendantes et identiquement distribuées, de même loi que X. On note $\overline{X}_n$ la moyenne empirique de l'échantillon :

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

L'objectif de cet exercice est de démontrer l'inégalité de Hoeffding :

$$\forall t > 0,\ \mathbb{P}\left[\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq t\right] \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

On note ψ_X la fonction définie pour tout $s \in \mathbb{R}$ par

$$\psi_X(s) = \mathbb{E}[e^{sX}].$$

On suppose que X est centrée.

1°. Démontrer que pour tout $x \in]a,b[$ fixé, $s \mapsto e^{sx}$ est convexe et en déduire que

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}.$$

2°. Démontrer que pour tout $s \in \mathbb{R}$,

$$\psi_X(s) \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}.$$

3°. On pose $p = b/(b-a)$, $q = 1-p$ et $u = (b-a)s$. On considère la fonction

$$u \mapsto \phi(u) = \ln\left(pe^{sa} + qe^{sb}\right).$$

Pour tout réel $u > 0$, expliciter $\phi(u)$ en fonction de u et en déduire qu'il existe $\theta \in]0,u[$ tel que :

$$\phi(u) = \phi(0) + \phi'(0)u + \frac{1}{2}\phi''(\theta)u^2.$$

4°. Déterminer $\phi(0)$, $\phi'(0)$, $\phi''(u)$ et démontrer que $\phi''(\theta) \leq \frac{1}{4}$.

5°. Démontrer que,

$$\forall s > 0,\ \psi_X(s) \leq \exp\left(\frac{s^2}{8}(b-a)^2\right).$$

6°. Rappeler et démontrer l'inégalité de Markov.

7°. On suppose toujours X centrée. Démontrer que,

$$\forall t > 0,\ \forall s > 0,\ \mathbb{P}\left[\overline{X}_n \geq t\right] \leq e^{-st}\left(\psi_X\left(\frac{s}{n}\right)\right)^n.$$

8°. On ne suppose plus que X est centrée. Déduire des questions précédentes que :

$$\mathbb{P}\left[\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq t\right] \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

puis que

$$\mathbb{P}\left[\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq t\right] \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

9°. Soit $\delta \in]0,1[$. Démontrer que :

$$\forall n \geq 1,\ \mathbb{P}\left[\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \leq \frac{|a-b|}{\sqrt{2n}}\sqrt{\ln(2/\delta)}\right] \geq 1-\delta.$$

3.16 Intervalles de confiance et de crédibilité pour une loi exponentielle

Soient $X_1, ..., X_n$ un n-échantillon de X suivant une loi exponentielle de paramètre θ inconnu.

1°. Décrire le modèle statistique. Démontrer que $S_n = \sum_{i=1}^n X_i$ est une statistique exhaustive de θ. Rappeler sa loi.

2°. On veut estimer $\mathbb{E}[X]$. Rappeler l'expression de cette espérance en fonction de θ. Calculer le risque quadratique associé à $\overline{X}$, moyenne empirique des X_i.

3°. En déduire un estimateur sans biais.

4°. Déterminer un intervalle de confiance de θ au niveau $1-\alpha$ pour $n = 15$ et $\alpha = 0,05$ (on rappelle que si Y suit une loi gamma de paramètres (a,θ), alors θY suit une loi Gamma de paramètres $(a,1)$).

5°. On utilise une loi *a priori* Gamma de paramètres (b,η). En déduire un estimateur bayésien de θ. Que retrouve-t-on approximativement pour de grandes tailles de l'échantillon?

7°. En quoi les études précédentes auraient-elles été modifiées si on avait observé n v.a. de loi Gamma de paramètres (a,θ) avec a connu?

3.17 Intervalles de confiance pour un échantillon uniforme

Soient $X_1, ..., X_n$ un n-échantillon de X suivant une loi uniforme sur $[0,\theta]$ avec θ inconnu.

1°. Décrire le modèle statistique. Démontrer que $Y = \sup(X_1, ..., X_n)$ est une statistique exhaustive de θ. Préciser sa loi, son espérance et sa variance.

2°. Déterminer l'estimateur du maximum de vraisemblance. Est-il sans biais? En déduire un

estimateur sans biais, calculer les deux risques quadratiques associés.

3°. Trouver un estimateur sans biais fondé sur la moyenne empirique et le comparer aux deux précédents.

4°. Pour $\alpha = 0,05$ et $n = 15$, déterminer un intervalle de confiance de niveau $1-\alpha$ pour θ (utiliser le fait que $\theta^n Y$ suit une loi libre de θ).

3.18 Intervalles de confiance pour un échantillon gamma

On considère un n-échantillon de X suivant une loi $\gamma(a,b)$ dont la densité est donnée par

$$f(x) = \frac{b}{\Gamma(a)}(bx)^{a-1}e^{-bx}\mathbb{1}_{[0,\infty[}(x) \tag{169}$$

où Γ est la fonction Gamma d'Euler. On note $\overline{X}$ la moyenne empirique de l'échantillon et S^2 sa variance empirique non biaisée. On suppose a connu.

1°. Déterminer l'estimateur $\widehat{b}$ du maximum de vraisemblance de b.

2°. Montrer que la fonction caractéristique d'une loi $\gamma(a,b)$ est

$$\phi(t) = \left(\frac{b}{b-it}\right)^a \tag{170}$$

En déduire la loi de la somme des observations. Calculez le biais de l'estimateur du maximum de vraisemblance. Peut-on en déduire un estimateur sans biais ?

3°. Établir la $\sqrt{n}$-consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance. En déduire que la variable aléatoire

$$\sqrt{an}\left(\widehat{b}/b - 1\right) \tag{171}$$

est un pivot asymptotique.

4°. L'estimateur du maximum de vraisemblance est-il asymptotiquement efficace ? Déterminer un estimateur admissible pour le risque quadratique parmi les estimateurs sans biais ?

On suppose maintenant également a inconnu.

5°. Montrer que l'estimateur des moments de (a,b) est

$$\left(\frac{\overline{X}^2}{S^2}, \frac{\overline{X}}{S^2}\right) \tag{172}$$

6°. Montrer que l'estimateur des moments est consistant.

7°. Déduire des questions précédentes que la variable aléatoire

$$\sqrt{n}\frac{\overline{X}}{S}\left(\frac{\overline{X}^2/S^2}{b\overline{X}} - 1\right) \tag{173}$$

est un pivot asymptotique. En déduire un intervalle de confiance asymptotique de niveau $1-\alpha$.

3.19 Intervalles de crédibilité pour un échantillon uniforme

On considère un n-échantillon de X suivant une loi uniforme sur $[0,\theta]$. θ est inconnu et suit une loi *a priori* de densité

$$\pi(\theta) = \frac{1}{\theta^2}\mathbb{1}_{[1,\infty[}(\theta) \tag{174}$$

1°. Montrer que la loi *a posteriori* du modèle s'écrit sous la forme

$$L(\theta|x) = \frac{c(x)}{\theta^{b(x)}}\mathbb{1}_{\theta > s(x)} \tag{175}$$

avec $x = (x_1, ..., x_n)$.

2°. Déterminer l'estimateur de Bayes de θ pour le risque quadratique.

3°. Construire un intervalle de crédibilité pour θ à 90%.

4 Tests statistiques

4.1

On souhaite vérifier que la contenance de bouteilles en provenance d'un producteur respecte bien en moyenne la limite légale de 75 cL. On sélectionne au hasard un échantillon de 10 bouteilles et l'on obtient une contenance moyenne de 74,42 cL.

On suppose que la contenance des bouteilles (en cL) suit une loi normale d'espérance θ inconnue et d'écart type égal à 1.

1°. Décrire le modèle statistique correspondant.

2°. On effectue le test

$$\begin{cases} \mathscr{H}_0 : \theta = 75 \\ \mathscr{H}_1 : \theta < 75 \end{cases} \tag{176}$$

Quel point de vue adopte-t-on en choisissant ces hypothèses ?

3°. Construire, à l'aide d'une règle de décision intuitive basée sur la moyenne empirique, un test pur de niveau $\alpha = 1\%$ de $\mathscr{H}_0$ contre $\mathscr{H}_1$. Quelle est la conclusion de ce test ?

4°. Tracer l'allure de la courbe de puissance et de la courbe d'efficacité de ce test.

5°. On veut pouvoir détecter, avec une probabilité de 99%, une contenance moyenne de 74,8 cL, tout en gardant un test de niveau $\alpha = 0,1\%$. Que doit-on faire ?

6°. Quelles sont les caractéristiques et la conclusion du test suivant :

$$\begin{cases} \mathscr{H}_0 : \theta \geq 75 \\ \mathscr{H}_1 : \theta < 75 \end{cases} \tag{177}$$

7°. On suppose maintenant que la contenance des bouteilles suit une loi normale de moyenne 75 cL et d'écart type inconnu θ. Décrire le modèle correspondant, le test, et donner sa conclusion.

4.2

Avant le second tour d'une élection présidentielle, un candidat commande un sondage à une société spécialisée, pour savoir s'il a une chance d'être élu.

Soit p, la proportion d'électeurs qui lui est favorable dans la population. On pose

$$\begin{cases} \mathscr{H}_0 : p = 0,48 \\ \mathscr{H}_1 : p = 0,52 \end{cases} \tag{178}$$

1°. Décrire le modèle statistique correspondant. Quelle est la signification du choix $p = 0,48$ comme hypothèse nulle ? Quelle statistique de test peut-on considérer ?

2°. Construire un test de niveau 10%, puis un autre de niveau asymptotique 10% lorsque le sondage est effectué auprès de $n = 100$ personnes.

3°. Combien d'électeurs devra-t-on interroger si l'on souhaite avoir un seuil asymptotique α et un risque de second espèce asymptotique inférieur à β, avec α et β donnés ?

4°. Le candidat souhaite maintenant tester

$$\begin{cases} \mathscr{H}_0 : p \leq 0,5 \\ \mathscr{H}_1 : p > 0,5 \end{cases} \tag{179}$$

Que peut-on conclure ?

4.3

Le nombre annuel de pannes sur une voiture d'un modèle donné, peut être modélisé par une v.a. qui suit une loi de Poisson de paramètre $\lambda = 2$. Après avoir souscrit un contrat d'entretien, on s'attend à ce que la valeur du paramètre diminue.

1°. Construire un test, basé sur le nombre total de pannes pour 6 ans de contrat, permettant de le vérifier.

2°. Que décide-t-on au seuil de 10% si le nombre total de pannes sur les six dernières années est de 10 ?

3°. Tracer l'allure de la courbe d'efficacité du test.

4.4

Un programme de simulation d'une loi uniforme sur $[0,\theta]$ a généré les nombres suivants : 95, 24, 83, 52, 68.

1°. Donner l'estimateur du maximum de vraisemblance de θ et déterminer sa loi.

2°. En déduire un test de niveau 5% de la forme

$$\begin{cases} \mathscr{H}_0 : \theta = 100 \\ \mathscr{H}_1 : \theta > 100 \end{cases} \tag{180}$$

Que peut-on conclure ?

4.5

On admet que la durée de vie d'un matériel est modélisé par une v.a. X de loi exponentielle « à l'anglo-saxonne », de paramètre θ. On considère un n-échantillon $(X_1,...,X_n)$ de X et une observation $(x_1,...,x_n)$ de cet échantillon.

1°. Déterminer l'estimateur $\widehat{\theta}_n$ du maximum de vraisemblance de θ.

2°. On rappelle que la densité d'une loi du chi-deux à $2k$ degrés de libertés est donnée par

$$g_k(x) = \frac{1}{2^k(k-1)!} x^{k-1} e^{-x/2} \mathbb{1}_{[0,\infty[}(x) \tag{181}$$

Montrer que la variable $2X/\theta$ suit une loi du chi-deux à 2 degrés de libertés. En déduire la loi de

$$Z = \frac{2}{\theta} \sum_{i=1}^{n} X_i \tag{182}$$

3°. Des études passées avaient permis d'attribuer au paramètre θ la valeur θ_0. L'évolution des méthodes de fabrication pouvant avoir entraîné une augmentation de θ, on considère le test suivant :

$$\begin{cases} \mathscr{H}_0 : \theta = \theta_0 \\ \mathscr{H}_1 : \theta > \theta_0 \end{cases} \tag{183}$$

Construire un test de niveau α à partir de ces hypothèses.

4°. On a relevé les durées de vie de $n = 31$ matériels et l'on trouve $\sum x_i = 67,68$. Quelle est la conclusion du test pour un niveau $\alpha = 5\%$ et $\theta_0 = 2$ (durée de la garantie) ?

5°. On suppose maintenant que l'on n'observe pas les durées de vie X_i directement, mais seulement les variables

$$Y_i = \mathbb{1}_{[X_i \geq 2]} \tag{184}$$

pour $i = 1,...,n$. Proposer un nouveau test.

4.6

Les notes à l'examen de « statistique mathématique » sont aléatoires et suivent une loi normale. On a relevé les notes de 15 élèves deux années consécutives durant lesquelles l'enseignant a changé et l'on souhaite savoir si ce changement a eu un effet sur les résultats.

1°. Proposer un modèle statistique et un test.

2°. Quelles sont les conclusions du test si les notes l'année $n-1$ et l'année n sont respectivement

12,8	15	8,5	12,7	10,4
15,5	9,6	10,3	8,5	8,1
7,8	14	12,5	8,6	7

et

10,1	8,9	6,1	4,8	9,1
11,9	14,2	13,5	16	12,9
11,1	11	8,8	10	9,2

On rappelle que si S_1^2, S_2^2, σ_1^2 et σ_2^2 représentent les variances empiriques et théoriques des deux échantillons de notes, alors la statistique

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{185}$$

suit, sous l'hypothèse $\mathscr{H}_0$, une loi de Fisher $\mathscr{F}(14,14)$.

On rappelle également que si $\sigma_1 = \sigma_2$, la statistique suivante suit une loi de Student de paramètre 28 :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{2/15} \times S(X_1, X_2)} \tag{186}$$

avec $\overline{X}_1$ et $\overline{X}_2$ moyenne empirique respective de chaque échantillon de note et $S(X_1, X_2)$ écart type empirique de l'échantillon global des 30 notes.

4.7

En juillet 2010, un sondage Ifop sur un échantillon représentatif de 958 personnes donnait le résultat suivant : 8 sympathisants PS sur 100 contre 6 sympathisants UMP sur 100 sont tatoués. Sur les 958 personnes interrogées, 249 se sont déclarées sympathisantes PS, dont 20 tatoués, et 297 sympathisantes UMP, dont 19 tatouées, que peut-on penser de la déclaration suivante faite dans les journaux : « les sympathisants PS sont plus tatoués que les sympathisants UMP » ?

4.8

Montrer que les modèles statistiques suivants sont à rapport de vraisemblance monotone :

1°. Le modèle binomial $\mathscr{B}(n,\theta)$.

2°. Le modèle d'échantillonnage gaussien $\mathscr{N}(\theta,1)$.

3°. Le modèle d'échantillonnage de Poisson $\mathscr{P}(\theta)$.

4°. Les modèles exponentiels.

4.9

Soit X une v.a. réelle dont la loi a pour densité

$$f(x) = \frac{1}{2\theta\sqrt{x}} e^{-\sqrt{x}/\theta} \mathbb{1}_{[0,\infty[}(x) \tag{187}$$

avec $\theta > 0$ et $X_1, ..., X_n$ n-échantillon de cette loi.

1°. Montrer que $Y = 2\sqrt{X}/\theta$ suit une loi du chi-deux à 2 degrés de libertés. En déduire la loi de

$$S_n = \frac{2}{\theta} \sum_{i=1}^{n} \sqrt{X_i} \tag{188}$$

2°. On souhaite tester pour $0 < \theta_0 < \theta_1$

$$\begin{cases} \mathscr{H}_0 : \theta = \theta_0 \\ \mathscr{H}_1 : \theta = \theta_1 \end{cases} \tag{189}$$

Déterminer un test UPPα parmi les tests de niveau α et expliciter la puissance de ce test.

2°. Décrire tous les tests UPP parmi les tests de seuil α.

3°. On souhaite maintenant tester

$$\begin{cases} \mathscr{H}_0 : \theta \leq \theta_0 \\ \mathscr{H}_1 : \theta > \theta_1 \end{cases} \tag{190}$$

Existe-t-il un test UPPα pour ce nouveau problème ?

4.10

La limite du taux X de présence d'un polluant contenu dans les déchets d'usine est de 6 mg/kg. On effectue un dosage sur 12 prélèvements de 1 kg, pour lesquels on observe les taux x_i, $i = 1, .., 12$ de présence du polluant. On trouve

$$\sum_{i=1}^{12} x_i = 84 \text{ et } \sum_{i=1}^{12} x_i^2 = 1413 \tag{191}$$

On admet que X suit une loi normale $\mathscr{N}(m, \sigma^2)$ avec $\sigma = 8$.

1°. Donner un test UPP parmi les tests de seuil 5% de

$$\begin{cases} \mathscr{H}_0 : m \leq 6 \\ \mathscr{H}_1 : m > 6 \end{cases} \tag{192}$$

Détermine la puissance de ce test. L'usine est-elle conforme à la législation ?

2°. Envisager le cas où l'écart-type σ est inconnu.

4.11

On dispose de l'observation $(x_1, ..., x_n)$ d'un échantillon de taille $n = 15$ d'une loi normale $\mathscr{N}(0, 1/\theta)$.

1°. Construire un test UPP parmi les tests de seuil $\alpha = 5\%$ de

$$\begin{cases} \mathscr{H}_0 : \theta = 1 \\ \mathscr{H}_1 : \theta > 1 \end{cases} \tag{193}$$

et déterminer la puissance de ce test.

2°. Quelle décision prend-on si $\sum_i x_i^2 = 6,8$? Pour quelles valeurs de α prendrait-on la décision contraire ? Qu'a-t-on alors calculé ?

3°. Existe-t-il un test UPP parmi les tests de seuil $\alpha = 5\%$ pour le problème

$$\begin{cases} \mathscr{H}_0 : \theta = 1 \\ \mathscr{H}_1 : \theta \neq 1 \end{cases} \tag{194}$$

Expliquer.

4.12

Le revenu annuel des individus d'une population est distribué suivant une loi de Pareto de densité

$$f(x) = \frac{ak^a}{x^{a+1}} \mathbb{1}_{[k,\infty[}(x) \tag{195}$$

Les paramètres $k > 0$ (revenu minimum) et $a > 0$ (paramètre de forme) sont inconnus.

1°. Sur la base d'un échantillon de taille n, estimer (k, a) par la méthode du maximum de vraisemblance.

2°. On voudrait tester

$$\begin{cases} \mathscr{H}_0 : a = 1 \\ \mathscr{H}_1 : a \neq 1 \end{cases} \tag{196}$$

Montrer que tout test de rapport de vraisemblance admet une région de rejet de la forme

$$\mathscr{R} = [\mathrm{T} \leq s_1] \cup [\mathrm{T} \geq s_2] \tag{197}$$

où T est la statistique définie par

$$\mathrm{T}(\mathrm{X}) = \mathrm{T}(\mathrm{X}_1, ..., \mathrm{X}_n) = \ln\left(\frac{\prod_{i=1}^n \mathrm{X}_i}{\left(\min_{i=1}^n \mathrm{X}_i\right)^n}\right) \tag{198}$$

4.13

On considère une variable aléatoire $\mathrm{X} \sim \mathscr{N}(m, \sigma^2)$ et $(\mathrm{X}_1, ..., \mathrm{X}_n)$ un n-échantillon de X.

1°. En utilisant la méthode de Neyman-Pearson, résoudre le problème suivant :

$$\begin{cases} \mathscr{H}_0 : m = m_0 \\ \mathscr{H}_1 : m = m_1 \end{cases} \tag{199}$$

2°. Calculer la puissance du test.

3°. Application numérique : $n = 25$, $\overline{x} = 2,7$, $m_0 = 2$, $m_1 = 4$. Conclure. Calculer la p-valeur.

4.14

On veut vérifier la précision d'une balance après un an de fonctionnement. On suppose que la pesée d'un objet de 1 gramme suit une loi normale $\mathscr{N}(m, \sigma^2)$ avec $m = 1$. Initialement, la précision de la balance est $\sigma_0 = 1,5$ mg. Si cette précision a augmenté en $\sigma_1 > \sigma_0$, on concluera que la balance a perdu en précision. Le test proposé est donc :

$$\begin{cases} \mathscr{H}_0 : \sigma = \sigma_0 = 1,5 \\ \mathscr{H}_1 : \sigma = \sigma_1 = 2 \end{cases} \tag{200}$$

1°. En utilisant la méthode de Neyman-Pearson, déterminer la région critique du test.

2°. Application numérique : conclure lorsque $n = 10$ pour un risque $\alpha = 0,1$.

3°. Calculer la puissance du test.

4.15

Dans une population donnée, une proportion inconnue p d'individus possède un caractère C. On effectue le test suivant :

$$\begin{cases} \mathscr{H}_0 : p = p_0 \\ \mathscr{H}_1 : p = p_1 > p_0 \end{cases} \tag{201}$$

1°. Déterminer la région critique du test.

2°. Application numérique : conclure si $n = 625$

4.16

Déterminer la statistique du rapport de vraisemblance lorsque l'on teste, pour une valeur m_0 fixe,

$$\begin{cases} \mathscr{H}_0 : m \leq m_0 \\ \mathscr{H}_1 : m > m_0 \end{cases} \tag{202}$$

sur la base d'un échantillon i.i.d. de taille n de la loi normale $\mathscr{N}(m, \sigma^2)$, avec σ^2 inconnu.

Montrer que ce test est fondé sur une statistique de loi de Student.

4.17

Sur la base d'un échantillon de taille n de densité

$$f(x) = \frac{1}{\sigma} \exp\left(-\frac{x-\theta}{\sigma}\right) \mathbb{1}_{[\theta, +\infty[}(x) \tag{203}$$

où θ et $\sigma > 0$ sont inconnus, on désire effectuer le test

$$\begin{cases} \mathscr{H}_0 : \theta \leq \theta_0 \\ \mathscr{H}_1 : \theta > \theta_0 \end{cases} \tag{204}$$

Déterminer la forme du test de rapport de vraisemblance.

4.18

On considère un échantillon $(\mathrm{X}_1, ..., \mathrm{X}_n)$ d'une loi de densité

$$f(x) = e^{\theta - x} \mathbb{1}_{[\theta, +\infty[}(x) \tag{205}$$

1°. Montrer que la statistique $\mathrm{X}_{(1)} = \min_i \mathrm{X}_i$ est exhaustive pour θ.

2°. En se fondant sur cette statistique, déterminer la forme de tout test de rapport de vraisemblance de

$$\begin{cases} \mathscr{H}_0 : \theta \leq \theta_0 \\ \mathscr{H}_1 : \theta > \theta_0 \end{cases} \tag{206}$$

3°. Exprimer la puissance d'un tel test.

4°. Pour $\alpha \in [0, 1]$, quel test est de niveau α ?

4.19

Soit $\mathrm{X} \sim \mathscr{P}(\lambda)$. On effectue un test bayésien

$$\begin{cases} \mathscr{H}_0 : \lambda \leq 1 \\ \mathscr{H}_1 : \lambda > 1 \end{cases} \tag{207}$$

Autrement dit, on ne rejètera pas $\mathscr{H}_0$ si sa probabilité *a posteriori* est supérieure à celle de $\mathscr{H}_1$.

1°. Calculer la loi *a posteriori* de $\mathscr{H}_0$ pour $x = 1$ avec pour loi *a priori* $\lambda \sim \Gamma(\alpha, \beta)$.

2°. Montrer que lorsque α et β tendent vers 0, on obtient la même loi *a posteriori* qu'en posant $\Pi(\lambda) = 1/\lambda$. Quelle est alors la conclusion du test ? Montrer que cet *a priori* n'est pas toujours valide selon l'observation x.

4.20

Soient $\mathrm{X}_1, ..., \mathrm{X}_n$ un n-échantillon de X de loi uniforme sur $[0, \theta]$. Soit $\mathrm{M} = \mathrm{X}_{(n)} = \max_{i=1}^n \mathrm{X}_i$. On cherche à tester

$$\begin{cases} \mathscr{H}_0 : \theta = 1 \\ \mathscr{H}_1 : \theta > 1 \end{cases} \tag{208}$$

1°. Pourquoi ne peut-on pas utiliser le test du rapport de vraisemblance (donner deux raisons) ?

2°. On propose le test suivant : on rejette $\mathscr{H}_0$ lorsque $\mathrm{M} > s$, où s est une constante donnée. Calculer alors la fonction de puissance.

3°. Quelle valeur doit prendre s pour un seuil de 5%.

4°. Si $n = 2$ et que la valeur observée m de M est $m = 0,96$, que vaut la p-valeur ? Quelle conclusion sur les hypothèses ? Même question si $m = 1,04$.

5°. On se propose d'utiliser une approche bayésienne et de poser pour loi *a priori* (impropre) :

$$\Pi(\lambda) = \frac{1}{\theta} \tag{209}$$

Ce choix est-il justifié ? Calculer alors la probabilité *a posteriori* que $\theta > 1$.

4.21

On considère veux v.a. indépendantes U et V, de loi normale centrée réduite et l'on pose $Z = U/V$.

1°. Déterminer la densité de Z et reconnaître cette loi. Déterminer sa fonction caractéristique.

2°. Soit $\overline{Z}$ la moyenne empirique d'un n-échantillon de Z. Déterminer sa loi.

3°. On appelle loi de Cauchy générale de paramètre de position $\theta \in \mathbb{R}$ et de paramètre d'échelle $\sigma > 0$ la loi de la v.a. $X = \theta + \sigma Z$. On notera cette loi $\mathscr{C}(\theta, \sigma)$. Étant donné un n-échantillon $(X_1, ..., X_n)$ de X, déterminer les lois des variables aléatoires

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } Y = \sum_{i=1}^{n} a_i X_i \tag{210}$$

$a_i \in \mathbb{R}\ \forall i$.

4°. Quelle est la signification de θ ?

5°. On suppose $\sigma = 1$ et l'on veut estimer θ. Utiliser la méthode du maximum de vraisemblance pour construire un estimateur $\widehat{\theta}$ et calculer sa variance asymptotique. Peut-on estimer θ par la méthode des moments ? Proposer un estimtateur fondé sur l'interprétation de θ et calculer son efficacité asymptotique.

6°. Soit $Y_i = \mathbb{1}_{[X_i \leq 0]}$. Quelle est la loi de Y_i ? Quelle est la loi de N_n, nombre d'observations négatives ou nulles dans l'échantillon initial ? En déduire de la vraisemblance de N_n un estimateur de θ.

7°. Soient $U \sim \mathscr{N}(0, \sigma_1^2)$ et $V \sim \mathscr{N}(0, \sigma_1^2)$, deux v.a. indépendantes. On pose $\lambda = \sigma_1/\sigma_2$ et l'on l'on définit la v.a. Z par $Z = U/V$. déterminer la densité de Z ainsi que sa fonction de répartition.

8°. On dispose de deux échantillons $(U_1, ..., U_n)$ et $(V_1, ..., V_n)$ de U et V. Déterminer la loi de

$$T = \frac{\sum_{i=1}^{n} U_i}{\sum_{i=1}^{n} V_i} \tag{211}$$

En déduire l'estimateur $\widehat{\lambda}$ du maximum de vraisemblance de λ.

9°. On considère maintenant un vecteur gaussien (U, V) avec $U \sim \mathscr{N}(m_1, \sigma_1^2)$ et $V \sim \mathscr{N}(m_2, \sigma_2^2)$. U et V ne sont pas indépendantes et leur coefficient de corrélation linéaire est noté ρ. On pose

$$Z = \frac{(U - m_1)/\sigma_1}{(V - m_2)/\sigma_2} \tag{212}$$

Déterminer la loi de Z. Interpréter ρ. En déduire, à partir d'un échantillon $(U_n, V_n)_n$, un test asymptotique sur ρ de la forme :

$$\begin{cases} \mathscr{H}_0 : \rho = \rho_0 \\ \mathscr{H}_1 : \rho \neq \rho_0 \end{cases} \tag{213}$$

4.22

On considère une v.a. X de loi exponentielle (sous forme anglo-saxonne) de paramètre θ.

1°. Calculer l'information de Fisher apportée par un n-échantillon de X et déterminer la borne de FDCR associée à θ.

2°. Estimer θ par maximum de vraisemblance et étudier l'e.m.v. $\widehat{\theta}_1$. Faire de même avec l'estimateur des moments $\widehat{\theta}_2$, puis avec un estimateur $\widehat{\theta}_3$ fondé sur la v.a. Z égale au nombre d'observations supérieures ou égales à 2 dans l'échantillon.

3°. Donner un intervalle de confiance au niveau 0.95 pour θ.

4°. Une observation du n-échantillon pour $n = 31$ donne $\sum x_i = 64.83$ et $\sum x_i^2 = 170.92$. On pense que θ peut être égale à 1,2 ou 3. Résoudre les trois problèmes de tests suivants :

$$\begin{cases} \mathscr{H}_0 : \theta = 2 \\ \mathscr{H}_1 : \theta = 1 \end{cases} \tag{214}$$

$$\begin{cases} \mathscr{H}_0 : \theta = 2 \\ \mathscr{H}_1 : \theta = 3 \end{cases} \tag{215}$$

$$\begin{cases} \mathscr{H}_0 : \theta = 2 \\ \mathscr{H}_1 : \theta \neq 2 \end{cases} \tag{216}$$

5°. θ étant toujours inconnu, on s'intéresse au paramètre $d = \mathbb{P}[X \geq 2]$. Exprimer d en fonction de θ et déterminer l'e.m.v. $\widehat{d_1}$ de d. Quelle est sa loi asymptotique ? Donner un intervalle de confiance pour d au niveau de confiance 0.95.

6°. Pour un n-échantillon $(X_1, ..., X_n)$ de X, on considère maintenant l'estimateur δ défini par

$$\delta = \mathbb{1}_{[X_1 \geq 2]} \tag{217}$$

Calculer $\mathbb{E}[\delta]$ et $\mathbb{V}(\delta)$.

7°. On pose $S = \sum_{i=1}^{n} X_i$ et $T = \mathbb{E}[\delta|S]$. Calculer $\mathbb{E}[T]$ et montrer que

$$T(X_1, ..., X_n) = \left(1 - \frac{2}{S}\right)^{n-1} \mathbb{1}_{[S \geq 2]} \tag{218}$$

4.23 Loi de Fréchet

La loi de Fréchet de paramètres (m, σ, α) avec $x \geq m$, $\sigma > 0$, $\alpha > 0$ a pour fonction de répartition

$$F(x) = \exp\left(-\left(\frac{x - m}{\sigma}\right)^{-\alpha}\right) \mathbb{1}_{[m, +\infty[}(x) \tag{219}$$

1°. Montrer que la densité correspondante est

$$F(x) = \frac{\alpha}{\sigma}\left(\frac{x - m}{\sigma}\right)^{-\alpha-1} \exp\left(-\left(\frac{x - m}{\sigma}\right)^{-\alpha}\right) \mathbb{1}_{[m, +\infty[}(x) \tag{220}$$

Soit $X_1, ..., X_n$ un n-échantillon de X suivant une loi de Fréchet définie précedemment. On suppose que $m = 0$ et $\alpha = 1$ et l'on veut estimer $\theta = \sigma$.

1°. Le modèle est-il exponentiel?

2°. Montrer que

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i} \tag{221}$$

est un estimateur sans biais efficace de θ.

3°. Donner l'e.m.v. de θ et préciser son comportement asymptotique (variance asymptotique, etc.).

4°. Montrer que cet estimateur est biaisé et qu'il n'existe pas d'estimateur sans biais pour $n = 1$.

5°. Montrer que la famille de lois Gamma

$$f(s) = s^{a-1}e^{-bs}\mathbb{1}_{[0,\infty[}(s) \tag{222}$$

pour $a, b > 0$ est conjuguée pour ce modèle.

On suppose maintenant que $\alpha = 1$, $\sigma = 1$ et l'on cherche à estimer $\theta = m$.

6°. Est-ce que les conditions de régularité nécessaires au comportement asymptotique standard de l'e.m.v. sont satisfaites ici?

7°. Posons $\widehat{m}_1 = \min_{i=1}^{n} X_i$. Montrer que pour tout $c > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\left[n^a(\widehat{m}_1 - m) > c\right] = \begin{cases} 0 \text{ si } a = 0 \\ 1 \text{ si } a > 0 \end{cases} \tag{223}$$

8°. On cherche à construire un second estimateur plug-in $\widehat{m}_2$ basé sur la médiane m_e. Montrer que

$$\sqrt{n}(\widehat{m}_2 - m_e) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{4f(m_e)}\right) \tag{224}$$

puis construire à partir de ce résultat un estimateur estimateur asymptotiquement normal de m.

9°. À partir de la question précédente, construire un test de $\mathscr{H}_0 : m = 0$ *vs* $\mathscr{H}_1 : m \neq 0$. Donner la statistique de test et la région de confiance.

10°. Dans cette dernière question, on suppose que $m = 0$ et l'on cherche à estimer $\theta = (\sigma, \alpha)$.

Proposer un test pour les hypothèses $\mathscr{H}_0 : \alpha = 1$ *vs* $\mathscr{H}_1 : \alpha \neq 1$. Donner la statistique de test et la région de confiance.

5 Applications

5.1 Un modèle de réseau social : stochastic block models

Les stochastic block models (SBM) sont des modèles de graphes aléatoires utilisés pour modéliser des réseaux sociaux contenant des communautés (facebook, etc.).

5.2 Détection des virus informatiques par analyse comportementale

Ce problème nécessite la lecture d'un chapitre du livre d'Eric Filiol « techniques virales avancées », édité chez Springer.

5.3 Quantité d'information et entropie au sens de Shannon

Dans tout cet exercice, n est un entier naturel non nul. Toutes les variables aléatoires sont définies sur un même espace probabilisé $(\Omega, \mathscr{F}, \mathbb{P})$.
$\log_2$ représente le logarithme de base 2 et est défini, pour tout $x > 0$, par

$$\log_2(x) = \frac{\ln x}{\ln 2}. \tag{225}$$

On considère une variable aléatoire discrète X à support dans $\mathbb{N}$. Le cardinal de $X(\Omega)$ peut être fini ou non. On définit l'entropie de X, lorsqu'elle existe, par la formule

$$H(X) = -\sum_{x\in X(\Omega)} \mathbb{P}[X = x]\log_2 \mathbb{P}[X = x]. \tag{226}$$

1°. Démontrer que pour tout $x > 0$, $\ln x \leq x - 1$ et préciser les cas d'égalité.

2°. Soit ϕ la fonction définie sur $[0, 1]$ par $\phi(x) = -x\log_2(x)$ si $x > 0$ et $\phi(0) = 0$. Effectuer l'étude de cette fonction en précisant sa monotonie et ses *extrema*. Démontrer que ϕ est concave, puis donner l'allure de sa courbe représentative.

Dans les questions 3° à 10°, on suppose $X(\Omega)$ et $Y(\Omega)$ de cardinaux finis.

3°. Démontrer que, quel que soit X, $H(X) \geq 0$. À quelle condition sur X a-t-on $H(X) = 0$?

Soit X une variable aléatoire de Bernoulli de paramètre $p \in]0, 1[$.

4°. Calculer H(X). H(X) est une fonction de p que l'on notera h. Effectuer l'étude de h. En quelle valeur h atteint-elle son maximum? Interpréter le résultat.

5°. Déterminer H(X) lorsque X est une variable aléatoire de loi uniforme sur $X(\Omega) = \{1, ..., n\}$.

6°. À l'aide de la question 1°, démontrer l'inégalité de Gibbs : si $(p_1, ..., p_n)$ et $(q_1, ..., q_n)$ sont des lois de probabilités à support dans $\{1, ..., n\}$ alors

$$\sum_{k=1}^{n} p_k \log_2(q_k/p_k) \leq 0. \tag{227}$$

avec égalité si, et seulement si $p_k = q_k$ pour tout k.

7°. Démontrer que pour toute variable aléatoire X sur $\{1, ..., n\}$, $H(X) \leq \log_2 n$. Interpréter ce résultat.

8°. L'entropie conjointe de deux variables aléatoires X et Y se définit par la formule

$$H(X, Y) = -\sum_{(x,y)\in X(\Omega)\times Y(\Omega)} \mathbb{P}[X = x, Y = y]\log_2 \mathbb{P}[X = x, Y = y]. \tag{228}$$

Si X et Y sont deux variables aléatoires indépendantes, démontrer que $H(X, Y) = H(X) + H(Y)$.

9°. Les variables X et Y ne sont plus supposées indépendantes. On définit l'entropie conditionnelle de Y sachant X par la formule

$$H(Y|X) = -\sum_{(x,y)\in X(\Omega)\times Y(\Omega)} \mathbb{P}[X=x, Y=y]\log_2 \mathbb{P}[Y=y|X=x]. \tag{229}$$

Démontrer que

$$H(X,Y) = H(Y|X) + H(X) = H(X|Y) + H(Y). \tag{230}$$

Démontrer que $H(X)+H(Y) \leq 2H(X,Y)$. En utilisant la concavité de ϕ, démontrer que

$$H(X) \geq H(X|Y). \tag{231}$$

En déduire que $H(Y) \geq H(Y|X)$, puis démontrer que $H(X,Y) \leq H(X)+H(Y)$.

10°. Pour toute fonction f définie sur $X(\Omega)$, démontrer que $H(f(X)|X) = 0$ et que $H(X) \geq H(f(X))$.

On suppose maintenant $X(\Omega)$ de cardinal infini. On admet que si $\mathbb{E}[X] < \infty$ alors $H(X)$ existe. On admet également que l'inégalité de Gibbs s'étend au cas où $X(\Omega)$ est dénombrable, sous réserve de convergence de la somme.

11°. Calculer l'entropie d'une variable aléatoire G de loi géométrique de paramètre $p \in]0,1[$, en justifiant son existence. Montrer que pour toute variable aléatoire X discrète telle que $\mathbb{E}[X] \leq \mathbb{E}[G]$, on a $H(X) \leq H(G)$.

On considère maintenant que $X(\Omega)$ est un intervalle de $\mathbb{R}$ et X une variable aléatoire à densité continue f sur $X(\Omega)$. On définit l'entropie différentielle de X, sous réserve d'existence, par la formule :

$$H(X) = -\int_{-\infty}^{+\infty} f(x)\ln(f(x))\,dx. \tag{232}$$

12°. Calculer l'entropie différentielle d'une variable aléatoire X de loi uniforme sur $[a,b]$.

13°. Calculer l'entropie différentielle d'une variable aléatoire suivant une loi normale $\mathcal{N}(m,\sigma^2)$.

14°. Soit X suivant une loi $\mathcal{N}(0,\sigma^2)$ dont la densité sera notée ψ et soit Y une variable aléatoire réelle, centrée, de variance finie σ^2, dont la densité sera notée f. En supposant les deux intégrales suivantes convergentes, démontrer que

$$H(Y) = \int_{-\infty}^{+\infty} f(x)\ln\frac{\psi(x)}{f(x)}\,dx - \int_{-\infty}^{+\infty} f(x)\ln\psi(x)\,dx. \tag{233}$$

En déduire que $H(Y) \leq H(X)$. Interpréter ce résultat.

5.4 Introduction aux turbocodes

A. Rapport de vraisemblance logarithmique

Nous souhaitons transmettre des données binaires sur un canal de communication bruité. Du point de vue du canal ou du récepteur, on ne connait pas à l'avance la valeur du bit qui va être transmis. Cette valeur peut donc se modéliser par une variable aléatoire U de Bernoulli pouvant prendre comme valeur 0 ou 1. Pour des raisons d'efficacité, on préfère souvent transmettre comme valeurs 1 et -1 au lieu de 0 ou 1 (c'est de la modulation antipodale). On pose alors $X = (-1)^U$ qui est également une variable aléatoire de Bernoulli, mais prenant comme valeurs 1 ou -1 selon que que U vaut 0 ou 1.

Pour une variable de Bernoulli, on appelle rapport de vraisemblance logarithmique ou LLR (log-likehood ratio) la quantité

$$L(X) = \log\left(\frac{\mathbb{P}[X=1]}{\mathbb{P}[X=-1]}\right)$$

Lorsque U est une variable aléatoire qui prend comme valeur 0 ou 1, nous noterons de même

$$L(U) = \log\left(\frac{\mathbb{P}[U=0]}{\mathbb{P}[U=1]}\right)$$

Comme nous allons le voir, cette grandeur réelle permet de connaître la valeur la plus probable de la variable aléatoire et d'en mesurer la fiabilité. Dans toute la suite, nous noterons $p = \mathbb{P}[X=-1]$.

1°. Effectuer l'étude complète de la fonction $f(x) = \log\left(\frac{x}{1-x}\right)$

2°. Démontrer que $L(X) > 0 \iff \mathbb{P}[X=1] > \mathbb{P}[X=-1]$

3°. Lorsque p varie de 0 à 1, déterminer les variations de $L(X)$. En quoi $L(X)$ mesure-t-il la fiabilité des valeurs que peut prendre X ?

4°. On définit la fonction sgn par

$$\mathrm{sgn}(x) = \begin{cases} 1 \text{ si } x > 0 \\ 0 \text{ si } x = 0 \\ -1 \text{ si } x < 0 \end{cases}$$

Démontrer que $L(X) = \mathrm{sgn}(L(X)) \times |L(X)|$, puis que $\mathrm{sgn}(L(X))$ est égal à la valeur la plus probable de X.

Lorsque l'on remplace $L(X)$ par $\mathrm{sgn}(L(X))$ pour donner la valeur la plus probable de X, on dit que l'on a effectué une décision ferme. La valeur obtenue est un entier. Lorsque l'on travaille avec $L(X)$ (qui est un nombre réel) on dit que l'on effectue un décodage souple, car $L(X)$ conserve une information supplémentaire sur la fiabilité de la valeur de X.

5°. Démontrer que

$$\begin{cases} \mathbb{P}[X=1] = \dfrac{e^{L(X)}}{1+e^{L(X)}} \\ \mathbb{P}[X=-1] = \dfrac{e^{-L(X)}}{1+e^{-L(X)}} = \dfrac{1}{1+e^{L(X)}} \end{cases}$$

6°. En déduire que $\mathbb{E}[X] = \tanh\dfrac{L(X)}{2}$ où tanh représente la fonction tangente hyperbolique.

7°. On considère maintenant deux variables aléatoires indépendantes U_1 et U_2, pouvant prendre comme valeurs 0 ou 1. A ces deux variables correspondent deux autres variables aléatoires X_1 et X_2 pouvant prendre comme valeurs -1 et 1 et définies par $X_1 = (-1)^{U_1}$ et $X_2 = (-1)^{U_2}$. Expliquer rapidement pourquoi X_1 et X_2

sont indépendantes. Démontrer que l'addition modulo deux de U_1 et U_2 (que nous noterons $U_1 \oplus U_2$) correspond au produit des variables aléatoires X_1 et X_2.

8°. Déduire de la question 6° la règle des tangentes hyperboliques :

$$\tanh\frac{L(X_1X_2)}{2} = \tanh\frac{L(X_1)}{2} \times \tanh\frac{L(X_2)}{2}$$

Puis que

$$L(X_1X_2) = 2\,\mathrm{arg\,tanh}\left(\tanh\frac{L(X_1)}{2} \times \tanh\frac{L(X_2)}{2}\right)$$

où arg tanh représente la réciproque de la fonction tangente hyperbolique. Afin de simplifier la formule précédente, nous noterons plutôt $L(X_1X_2) = L(X_1) \boxplus L(X_2)$

Pour des variables aléatoires U_1 et U_2 prenant comme valeurs 0 ou 1 nous noterons de la même façon $L(U_1 \oplus U_2) = L(U_1) \boxplus L(U_2)$.

9°. Dans le cas où $p = \mathbb{P}[X_1 = 1] = \mathbb{P}[X_2 = 1] = 3/4$, calculer $L(X_1)$, $L(X_2)$ et $L(X_1X_2)$.

10°. Montrer par ailleurs que $L(X_1X_2)$ est égal à

$$\log\left(\frac{\mathbb{P}[X_1 = 1]\mathbb{P}[X_2 = 1] + \mathbb{P}[X_1 = -1]\mathbb{P}[X_2 = -1]}{\mathbb{P}[X_1 = 1]\mathbb{P}[X_2 = -1] + \mathbb{P}[X_1 = -1]\mathbb{P}[X_2 = 1]}\right)$$

B. Information extrinsèque.

Soit X la variable aléatoire représentant le bit émis à l'entrée du canal et Y la variable aléatoire représentant le bit reçu à la sortie du canal. Le lien entre X et Y est donné par l'équation $Y = X + B$ où B représente le bruit du canal. Il s'agit souvent d'une variable aléatoire qui peut prendre des valeurs discrètes (cas d'un canal binaire symétrique) ou bien continues (cas d'un canal à bruit blanc gaussien). Ainsi, la valeur du bit reçu pourra être différente de la valeur du bit émis et n'est d'ailleurs pas forcément égale à 1 ou -1 ; elle peut prendre comme valeur 0.8, 1.2, -0.33, etc. Le récepteur aura pour tâche d'essayer de retrouver la valeur émise $[X = x]$ en ne connaissant que la valeur reçue $[Y = y]$. Nous introduisons à cet effet un peu de vocabulaire :

La vraisemblance *a priori* est la quantité
$L(X) = \log\left(\frac{\mathbb{P}[X=1]}{\mathbb{P}[X=-1]}\right)$. Elle ne dépend que du bit émis.

La vraisemblance *a postériori* est la quantité
$L(X|Y) = \log\left(\frac{\mathbb{P}[X=1|Y=y]}{\mathbb{P}[X=-1|Y=y]}\right)$

Elle donne une information sur le bit réellement émis sachant la valeur observée à la sortie du canal. C'est cette quantité qui nous servira pour décoder les valeurs reçues et décider des valeurs réellement émises.

La vraisemblance du canal est la quantité
$L(Y|X) = \log\left(\frac{\mathbb{P}[Y=y|X=1]}{\mathbb{P}[Y=y|X=-1]}\right)$

1°. En utilisant la formule de Bayes, démontrer que
$L(X|Y) = L(Y|X) + L(X) \quad (\star)$

L'utilisation d'un code correcteur va permettre de protéger l'information et d'essayer de retrouver les valeurs émises, mêmes si celles-ci ont été modifiées par le canal. Nous allons montrer que dans ce cas, apparait dans le membre de droite de l'équation $(\star)$ un troisième terme L_e appelé information extrinsèque et représentant le gain d'information, pour un bit reçu, apporté par le décodage.

Pour illustrer le turbo décodage, nous allons utiliser un code correcteur très simple que vous manipulez déjà : Le bit de parité. Lorsque l'on doit transmettre deux bits U_1 et U_2 (qui peuvent prendre comme valeurs 0 ou 1), on transmet également le bit $U_3 = U_1 \oplus U_2$, qui permet de vérifier que l'on a toujours $U_1 \oplus U_2 \oplus U_3 = 0$. Ce faisant, on crée une dépendance entre ces trois variables. En terme de probabilité conditionnelle, la connaissance de l'une d'elles modifiera donc la connaissance que l'on a des autres.

Dans toute la suite, nous noterons E_U l'évènement « l'équation de parité (E_U) est satisfaite », E_V l'évènement « l'équation de parité (E_V) est satisfaite », etc. Par exemple $E_U = [U_1 \oplus U_2 \oplus U_3 = 0]$.

2°. Exprimer par une phrase en français l'évènement $E_U\,|[U_1 = 1]$. Faire de même avec $E_U|[U_1 = 0]$ et $[U_1 = 1]|E_U$. la barre verticale signifie « sachant que ».

3°. Démontrer que

$$\mathbb{P}(E_U|[U_1 = 1]) = \mathbb{P}([U_2 = 1] \cap [U_3 = 0]|[U_1 = 1]) + \ldots$$
$$\ldots + \mathbb{P}([U_2 = 0] \cap [U_3 = 1]|[U_1 = 1])$$

et que

$$\mathbb{P}(E_U/[U_1 = 0]) = \mathbb{P}([U_2 = 1] \cap [U_3 = 1]|[U_1 = 0]) + \ldots$$
$$\ldots + \mathbb{P}([U_2 = 0] \cap [U_3 = 0]|[U_1 = 0])$$

Considérons les trois variables Y_1, Y_2, Y_3 correspondants aux valeurs reçues lorsque U_1, U_2, U_3 sont émis sur le canal. Les valeurs de Y_1, Y_2, Y_3 sont donc les valeurs observées à la sortie du canal. Afin de simplifier les notations, nous noterons sous forme de vecteur $Y = (Y_1, Y_2, Y_3)$ la loi conjointe de ces trois variables aléatoires. Ainsi, si $y = (y_1, y_2, y_3)$ est un vecteur de $\mathbb{R}^3$,

$$[Y = y] = [Y_1 = y_1] \cap [Y_2 = y_2] \cap [Y_3 = y_3]$$

Notons maintenant $L(U_1|Y)$ la vraisemblance *a postériori* du bit U_1, sachant les valeurs observées Y_1, Y_2, Y_3 à la sortie du canal et sachant que U_1 vérifie l'équation de parité $U_1 \oplus U_2 \oplus U_3$ du code. La définition de $L(U_1|Y)$ est la suivante :

$$L(U_1|Y) = \log\left(\frac{\mathbb{P}[U_1 = 0|E_U; Y = y]}{\mathbb{P}[U_1 = 1|E_U; Y = y]}\right)$$

4°. A l'aide de la formule de Bayes et des questions précédentes, démontrer que :

$$L(U_1|Y) = L(Y_1|U_1) + L(U_1) + L_e(U_1; E_U)$$

avec

$$L(Y_1|U_1) = \log\left(\frac{\mathbb{P}[Y_1 = y_1|U_1 = 0]}{\mathbb{P}[Y_1 = y_1|U_1 = 1]}\right)$$

$$L(U_1) = \log\left(\frac{\mathbb{P}[U_1 = 0]}{\mathbb{P}[U_1 = 1]}\right)$$

Et le dernier terme qui représente l'information extrinsèque apportée à U_1 par U_2 et U_3 grâce à la présence du code correcteur :

$$L_e(U_1;E_U) =$$

$$\log\left(\frac{\mathbb{P}[Y_2=y_2;U_2=1]\times\mathbb{P}[Y_3=y_3;U_3=1]+\mathbb{P}[Y_2=y_2;U_2=0]\times\mathbb{P}[Y_3=y_3;U_3=0]}{\mathbb{P}[Y_2=y_2;U_2=1]\times\mathbb{P}[Y_3=y_3;U_3=0]+\mathbb{P}[Y_2=y_2;U_2=0]\times\mathbb{P}[Y_3=y_3;U_3=1]}\right)$$

5°. Si l'on omet les évènements relatifs aux valeurs observées de Y_2 et Y_3, faites le lien entre cette formule et celle de $L(U_2 \oplus U_3)$ de la première partie et en déduire que

$$L_e(U_1;E_U) = [L(Y_2|U_2) + L(U_2)] \boxplus [L(Y_3|U_3) + L(U_3)]$$

C. Le canal à bruit blanc gaussien additif.

C'est un modèle de canal de communication très utilisé. On le note BABG en français et AWGN en anglais (additive white gaussian noise). Le principe est le suivant : On émet un bit, modélisé par une variable de Bernoulli X, en entrée du canal et on suppose que la sortie est donnée par une variable aléatoire $Y = X + B$ où B est une variable aléatoire qui représente le bruit du canal. Dans le cas d'un canal BABG, B suit une loi gaussienne dont la moyenne est nulle et la variance σ^2 est proportionnelle au rapport signal sur bruit du canal : Puis le canal est brouillé, plus σ^2 est important. Ceci signifie que la valeur reçue après transmission sur le canal sera un nombre réel plus ou moins éloigné de la valeur émise.

1°. Si X peut prendre comme valeur ± 1, la densité $\phi_1(y)$ de B sachant $X = 1$ suit une loi $\mathcal{N}(1,\sigma^2)$ et la densité $\phi_{-1}(y)$ de B sachant que $X = -1$ suit une loi $\mathcal{N}(-1,\sigma^2)$ (nous admettons ce résultat). Tracer soigneusement les densités de ces deux variables aléatoires sur un même graphique. Donner une interprétation géométrique à l'évènement $R_1|E_{-1}$ =« le bit 1 a été reçu alors que -1 avait été émis ». Faire de même avec $R_{-1}|E_1$ =« le bit -1 a été reçu alors que 1 avait été émis ». En déduire une interprétation géométrique de l'évènement E =« il y a eu erreur lors de la transmission ».

2°. Dans le cas où $\sigma = 1$, calculer, à l'aide de la table de la loi normale, la probabilité qu'une erreur se produise.

3°. La formule de Bayes s'applique également aux lois continues en remplaçant simplement les probabilités par leur densité. Démontrer alors que

$$\begin{cases} \mathbb{P}[X = 1/Y = y] = \dfrac{1}{1+e^{-2y/\sigma^2}} \\ \mathbb{P}[X = -1/Y = y] = \dfrac{1}{1+e^{2y/\sigma^2}} \end{cases}$$

Démontrer qu'il s'agit d'une loi de Bernoulli. Cette loi conditionnelle permet de décider, au vue de la valeur de y reçue sur le canal, si c'est un 1 ou un -1 qui a été émis.

4°. Démontrer que pour un canal BABG de variance σ^2, la vraisemblance du canal est

$$L(X|Y) = \frac{2y}{\sigma^2}$$

D. Principe du turbo-décodage.

Le principe du turbo-code est simple. On va présenter les données en ligne et en colonne dans un tableau à deux dimensions. On effectue un premier décodage en ligne en réinjectant l'information extrinsèque obtenue par ce décodage dans chaque bit. On effectue ensuite un décodage en colonne en opérant de même. On itère ce procédé en continuant à alterner décodage en ligne et décodage en colonne jusqu'à que toutes les équations de parité soient vérifiées. Une très bonne analogie est donnée par un tableau de mots croisés. Vous pouvez lire à cet effet l'article qui accompagne ce devoir.

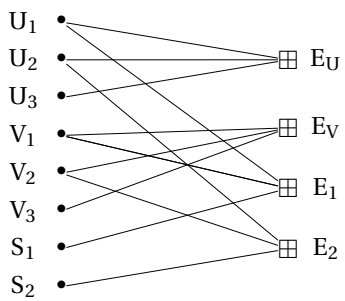
Considérons alors les variables aléatoires de Bernoulli indépendantes U_1, U_2, V_1 et V_2 qui représentent les données à transmettre sur le canal. Définissons les variables aléatoires U_3, V_3, S_1 et S_2 de telle sorte que :

$$\begin{cases} U_1 \oplus U_2 = U_3 & (E_U) \\ V_1 \oplus V_2 = V_3 & (E_V) \\ U_1 \oplus V_1 = S_1 & (E_1) \\ U_2 \oplus V_2 = S_2 & (E_2) \end{cases}$$

Ainsi, U_3 est le bit de parité de U_1 et U_2, V_3 est le bit de parité de V_1 et V_2, S_1 est le bit de parité de U_1 et V_1, S_2 est le bit de parité de U_2 et V_2. Les équations précédentes s'appellent des équations de parité. Nous les notons respectivement (E_U), (E_V), (E_1) et (E_2). Disposons ces 8 variables aléatoires dans un tableau à double entrée :

U_1	U_2	U_3
V_1	V_2	V_3
S_1	S_2	

Les équations de parité apparaissent à la fois en ligne et en colonne. Nous pouvons également visualiser les relations entre ces variables aléatoires à l'aide d'un graphe, appelé graphe de Tanner du code :



Les symboles • indiquent les nœuds de variables et représentent les bits, les symboles $\boxplus$ indiquent les nœuds de parité et représentent les équations de parité. La somme modulo 2 des bits reliés à une équation de parité doit être nulle.

Aux 8 variables aléatoires émises correspondent 8 variables aléatoires reçues que nous noterons sous la forme d'un tableau de nombre réels identique au tableau précédent :

Y_1	Y_2	Y_3
Y'_1	Y'_2	Y'_3
Y''_1	Y''_2	

Les liens entre les vraisemblances des bits de données

U_1, U_2, V_1, V_2 dûs aux codes de parité horizontaux sont donnés par les relations suivantes :

$$\begin{cases} L(U_1|Y_1) = L(Y_1|U_1) + L(U_1) + L_e(U_1; E_U) \\ L(V_1|Y_1') = L(Y_1'|V_1) + L(V_1) + L_e(V_1; E_V) \\ L(U_2|Y_2) = L(Y_2|U_2) + L(U_2) + L_e(U_2; E_U) \\ L(V_2|Y_2') = L(Y_2'|V_2) + L(V_2) + L_e(V_2; E_V) \end{cases}$$

avec

$$\begin{cases} L_e(U_1; E_U) = [L(Y_2|U_2) + L(U_2)] \boxplus L(Y_3|U_3) \\ L_e(V_1; E_V) = \left[L(Y_2'|V_2) + L(V_2)\right] \boxplus L(Y_3'|V_3) \\ L_e(U_2; E_U) = [L(Y_1|U_1) + L(U_1)] \boxplus L(Y_3|U_3) \\ L_e(V_2; E_V) = \left[L(Y_1'|V_1) + L(V_1)\right] \boxplus L(Y_3'|V_3) \end{cases}$$

Les liens entre les vraisemblances des bits de données U_1, U_2, V_1, V_2 dûs aux codes de parité verticaux sont donnés par les relations suivantes :

$$\begin{cases} L(U_1|Y_1) = L(Y_1|U_1) + L(U_1) + L_e(U_1; E_1) \\ L(V_1|Y_1') = L(Y_1'|V_1) + L(V_1) + L_e(V_1; E_1) \\ L(U_2|Y_2) = L(Y_2|U_2) + L(U_2) + L_e(U_2; E_2) \\ L(V_2|Y_2') = L(Y_2'|V_2) + L(V_2) + L_e(V_2; E_2) \end{cases}$$

avec

$$\begin{cases} L_e(U_1; E_1) = \left[L(Y_2'|V_2) + L(V_2)\right] \boxplus L(Y_1''|S_1) \\ L_e(V_1; E_1) = [L(Y_1|U_1) + L(U_1)] \boxplus L(Y_1''|S_1) \\ L_e(U_2; E_2) = \left[L(Y_2'|V_2) + L(V_2)\right] \boxplus L(Y_2''|S_2) \\ L_e(V_2; E_2) = [L(Y_1|U_1) + L(U_1)] \boxplus L(Y_2''|S_2) \end{cases}$$

On remarque l'absence de $L(U_3), L(V_3), L(S_1), L(S_2)$ dans ces formules. En fait, comme nous allons le voir, ces quantités sont nulles avec les hypothèses que nous allons poser.

Il est temps de présenter l'algorithme de turbo-décodage :

• 1°. Pour chaque variable T calculer la vraisemblance à priori L(T), avec $T = U_1, U_2, U_3, V_1, V_2, V_3, S_1, S_2$.

• 2°. Effectuer un décodage de chacune des deux lignes en calculant les informations extrinsèques horizontales $L_{eh}(T)$ pour $T = U_1, U_2, V_1, V_2$.

• 3°. Mettre à jour l'information à priori de chaque variable $T = U_1, U_2, V_1, V_2$ en posant $L(T) = L_e(T)$.

• 4°. Effectuer un décodage vertical de chacune des deux colonnes en calculant les informations extrinsèques verticales $L_{ev}(T)$ pour $T = U_1, U_2, V_1, V_2$.

• 5°. Mettre à jour l'information à priori de chaque variable $T = U_1, U_2, V_1, V_2$ en posant $L(T) = L_e(T)$.

• 6°. Mettre à jour la vraisemblance à posteriori des variables aléatoires $T = U_1, U_2, V_1, V_2$ en posant $L(Y|T) = L(T|Y) + L_{eh}(T) + L_{ev}(T)$.

• 7°. Calculer les valeurs les plus probables de $T = U_1, U_2, V_1, V_2$ en prenant une décision ferme sur $L(Y|T)$.

• 8°. Si toutes les équations de parité sont vérifiées, l'algorithme est fini, sinon retourner en 2°.

Passons maintenant aux questions de cette dernière partie :

1°. On suppose les variables aléatoires U_1, U_2, V_1, V_2 équiprobables. Démontrer qu'alors les vraisemblances a priori sont nulles :

$$L(U_i) = L(V_j) = L(S_k) = 0$$

$\forall i, j, k$.

2°. On souhaite émettre le message M = 1001 sur un canal BABG dont la variance sera supposée égale à $\sigma^2 = 1$. On dispose ces quatre bits sous la forme d'un tableau à deux lignes et deux colonnes en posant $U_1 = 1$, $U_2 = 0$, $V_1 = 0$, $V_2 = 1$. Les transformer, par modulation antipodale, en variables qui prennent comme valeurs ± 1. Calculer les bits de parité en ligne et en colonne et construire le tableau correspondant (c'est une question facile).

3°. Après transmission sur le canal, les valeurs reçues sont les suivantes :

−0.75	−0.05	−1.25
−0.1	−0.15	−1.0
−3.0	−0.5	

Construire le tableau des vraisemblances à priori du canal en utilisant les formules de la troisième partie.

4°. En faisant un petit programme sous Matlab, en langage C ou Python, décoder le tableau reçu après transmission et en déduire le message qui a été émis. Vous donnerez l'évolution des valeurs du tableau de vraisemblance au fur et à mesure des décodages horizontaux et verticaux.