

Master Ingénierie mathématique

1^{er} année

TD sous R de Statistique.

Anne PHILIPPE

Adresses email :

Anne.Philippe@univ-nantes.fr

Pages web :

Information sur le module / données / exercices

http://www.math.sciences.univ-nantes.fr/~philippe/Master_1.html

http://www.math.sciences.univ-nantes.fr/~philippe/R_freeware.html

Première partie 1. Estimation non paramétrique

EXERCICE 1. ESTIMATION D'UNE LOI DISCRÈTE FINIE

1) Récupérer les données dans le fichier

<http://www.math.sciences.univ-nantes.fr/~philippe/data/SampleDiscret.txt> Les données ont été simulées suivant la loi binomiale de paramètre $n = 5$ et $p = 0.3$

Soit X_1 une variable aléatoire discrète prenant les $\{0, \dots, 5\}$. On veut estimer la loi de X_1 à partir d'un échantillon $X_1, \dots, X_n$

On définit les deux estimateurs suivants :

-o- les fréquences empiriques

$$\hat{p}_n = (\hat{p}_n(0), \dots, \hat{p}_n(5)) = \frac{1}{n}(N_n(0), \dots, N_n(5))$$

où

$$N_n(j) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{j\}}(X_i)$$

La fonction `table` calcule les valeurs des $N_n(j)$

-o- la fonction de répartition empirique

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{]-\infty, x]}(X_i)$$

qui estime la fonction de répartition de X_1 :

$$F(x) = P(X_1 \leq x)$$

La fonction `ecdf` calcule la fonction F_n :

> Fn = ecdf(x)

> plot(Fn)

2) Justifier que $\hat{p}_n$ est un estimateur de la densité c'est à dire du vecteur de probabilités

$$(P(X_1 = 0), \dots, P(X_1 = 5))$$

3) Pour $n = 50, 500, 1000, 10000$,

a) Tracer l'estimation de la loi par l'estimateur des fréquences empiriques $\hat{p}_n$ calculées sur les n premières observations

b) Ajouter les valeurs de $\{(i, P(X_1 = i)), i = 0 \dots 5\}$

Indication : Représenter les 4 graphiques sur une même fenêtre en utilisant la commande `par(mfrow=c(2,2))` qui partage la fenêtre en 2 lignes et 2 colonnes

4) Pour $n = 50, 500, 1000, 10000$,

a) Tracer la fonction de répartition empirique F_n calculée sur les n premières observations

b) Superposer la fonction de répartition théorique

Indication : utiliser `pbinom` pour calculer les valeurs de la fonction de répartition, puis construire avec `stepfun` une fonction constante par morceaux et continue à droite.

La syntaxe est

F=stepfun(x, z) où

— le vecteur **x** contient les points de discontinuité de la fonction,

— le vecteur **z** est de la forme `c(a, y)` où **a** est la valeur prise par F avant le premier point de discontinuité et **y** les valeurs de la fonction aux points de discontinuités **x**.

Pour représenter graphique la fonction F on utilise `plot` ou `lines` (pour superposer) : avec l'argument `vertical=FALSE`

> plot(F, vertical=FALSE) ou lines(F, vertical=FALSE)

5) Commenter les résultats obtenus

EXERCICE 2. LE PROCESSUS EMPIRIQUE

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuée suivant une loi continue. On rappelle que le processus empirique est défini par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{]-\infty, t]}(X_i).$$

- 1) Illustrer la convergence presque sûre de $F_n(x)$ vers $F(x) = P(X_1 \leq x)$ et le théorème de Glivenko Cantelli sur les exemples suivants
 - X_1 suit une loi uniforme sur $[0, 1]$
 - X_1 suit une loi gaussienne standard
 - X_1 suit une loi beta de paramètre $(1/2, 1/4)$,
 - X_1 suit une loi de student à 1 degré de liberté.
- 2) Illustrer la propriété suivante :

la loi de la variable aléatoire $\sup_t |F_n(t) - F(t)|$ ne dépend pas de la loi de X_1

Reprendre les exemples de la question 1).

- 3) Par la simulation, trouver une suite (u_n) telle que

$$u_n \sup_t |F_n(t) - F(t)|$$

converge en loi vers une variable aléatoire non dégénérée.

EXERCICE 3. ESTIMATION NON PARAMÉTRIQUE SUR DES LOIS CONTINUES

Récupérer les fichiers de données suivantes :

- le fichier `SampleGauss01.txt` contient une réalisation $(x_1, \dots, x_n)$ de l'échantillon $X = (X_1, \dots, X_n)$ iid suivant la loi normale $\mathcal{N}(0, 1)$.
<http://www.math.sciences.univ-nantes.fr/~philippe/data/SampleGauss01.txt>
- le fichier `SampleGauss03.txt` contient une réalisation $(y_1, \dots, y_n)$ de l'échantillon $Y = (Y_1, \dots, Y_n)$ iid suivant la loi normale $\mathcal{N}(0, 3^2)$.
<http://www.math.sciences.univ-nantes.fr/~philippe/data/SampleGauss03.txt>

Estimation de la densité par un histogramme

Quelques fonctions R : La fonction `hist` trace l'histogramme de l'échantillon x .

La syntaxe est

- `hist(x, proba = TRUE)`, le choix du nombre de classes est optimisé pour minimiser l'erreur quadratique moyenne.
 On peut fixer les classes en ajoutant l'argument `breaks`.
- `hist(x, proba = TRUE, breaks = p)` avec `p` un entier, le nombre de classes est approximativement `p` et les classes sont de même longueur.

— `hist(x, proba = TRUE , breaks = a)` avec `a` un vecteur, les coordonnées de `a` définissent les classes de l'histogramme. Il y a donc `length(a) - 1` classes.

- 1) Dans cette question, on estime la densité par un histogramme en utilisant le nombre de classes optimal de la fonction `hist`.

Pour $n = 50, 500, 1000, 10000$:

- Tracer l'histogramme calculé sur les n premières observations $X_1, \dots, X_n$.
- Superposer la densité de la loi théorique c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
- Commenter les résultats.

Tracer les 4 graphiques sur une même fenêtre en utilisant `par(mfrow=c(2,2))`.

- 2) Refaire la question 1 sur l'échantillon Y et comparer les résultats.

- 3) Sur l'échantillon X , nous allons illustrer les effets du nombre de classes sur la qualité de l'estimateur.

On fixe le nombre d'observations $n = 500$ et on construit l'histogramme avec k classes de même longueur pour différentes valeurs de $k = 3, 5, 10, 15, 20, 25, 50, 100, 150$.

- Construire une suite arithmétique a de longueur $k + 1$ allant $\min(X_1, \dots, X_{500})$ à $\max(X_1, \dots, X_{500})$. Ces points définissent les classes de l'histogramme.
- Tracer l'histogramme de $X_1, \dots, X_{500}$ ayant k classes définies par le vecteur a .
- Superposer la densité de la loi théorique c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
- Commenter les résultats obtenus. Comparer avec le résultat obtenu à la question 1.

Tracer les 9 histogrammes sur une même page : `par(mfrow=c(3,3))`.

Estimation de la densité par l'estimateur à Noyau

Quelques fonctions R :

- La fonction `fn = density(x)` calcule sur l'échantillon x l'histogramme à noyau avec un choix automatique de h , optimisé pour minimiser l'erreur quadratique moyenne et par défaut le noyau gaussien
- Pour changer la valeur de h , on ajoute l'option `bw`
- Pour changer le noyau, on ajoute l'option `kernel`
- La commande `plot(fn)` permet ensuite de représenter graphiquement l'estimateur.

- 1) Dans cette question, on utilise l'estimateur avec la valeur de h optimisée par la fonction `density` et le gaussien (par défaut).

Pour $n = 5, 10, 50, 100, 1000, 10000$, (tracer les 6 graphiques sur une même fenêtre.)

- a) Tracer l'estimateur à noyau calculé sur les n premières observations $X_1, \dots, X_n$.
 - b) Superposer la densité de la loi théorique c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
 - c) Commenter les résultats.
- 2) On fixe $n = 1000$ et on fait varier le paramètre h .
 Pour $h = 0.01, 0.02, 0.1, 0.25, 0.5, 1, 2$. On conserve le noyau gaussien.
- a) Tracer l'estimateur à noyau calculé sur les 1000 premières observations de X .
 - b) Superposer la densité de la loi théorique, c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
 - c) Commenter les résultats.
- 3) Refaire les questions 1 et 2 avec le noyau uniforme.

Deuxième partie 2. Elements de statistique de descriptive

EXERCICE 4.

Boxplot : pourquoi 1.5 ?

Dans un boxplot, les valeurs aberrantes sont définies comme les observations en dehors de l'intervalle

$$I = [Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)].$$

D'où vient le **1.5** ?

Réponse de l'inventeur du boxplot (Turkey) :

Because 1 is too small and 2 is too large.

On suppose que les observations sont iid suivant une loi normale standard $\mathcal{N}(0, 1)$.

- 1) Calculer les quartiles théoriques Q_1 et Q_3 d'une loi gaussienne standard (en utilisant la fonction `qnorm`).
- 2) En déduire l'écart inter quartiles théorique de la loi gaussienne standard et l'intervalle I .
- 3) Calculer la probabilité de l'évènement $[X \in I]$ quand $X \in \mathcal{N}(0, 1)$
- 4) Refaire ces calculs en remplaçant 1.5 par 1 puis par 2.
- 5) Commenter les résultats.

EXERCICE 5.

Robustesse de la médiane

On dispose d'un échantillon $\mathcal{X} = (x_1, \dots, x_n)$ de taille n iid suivant la loi gaussienne standard.

Pour évaluer la robustesse, on peut par exemple ajouter m observations (appelées valeurs aberrantes) appartenant à une région de faible probabilité pour la loi des observations puis on représente l'évolution de la médiane et de la moyenne en fonction de m .

- 1) Simuler un échantillon x avec $n=100$

- 2) Construire une fonction R qui retourne la moyenne et la médiane de l'échantillon x auquel on a ajouté m valeurs aberrantes simulées suivant la loi uniforme sur l'intervalle $[-10, -5] \cup [5, 10]$.
- 3) Représenter l'évolution de la moyenne et de la médiane en fonction de m
- 4) Commenter les résultats obtenus.

Troisième partie 3. Modèle uniforme

On considère un n -échantillon $(X_1, \dots, X_n)$ d'une loi uniforme sur $[0, \theta]$.

On pose

$$\hat{\theta}_n = \max\{X_1, \dots, X_n\}.$$

On a prouvé les résultats suivants en TD :

- $\hat{\theta}_n$ est asymptotiquement sans biais.
- $\hat{\theta}_n$ converge presque sûrement et dans L^2 vers θ quand $n \rightarrow \infty$.

Comment vérifier ces propriétés par la simulation ?

- 1-1 Simuler $N = 500$ échantillons de taille $n = 1000$ de loi uniforme sur $[0, \theta_0]$ par exemple $\theta_0 = 2$. Stocker ces échantillons dans une matrice.
- 2-2 Pour chacun des échantillons simulés, calculer les valeurs de $\hat{\theta}_i$ pour $i = 1, \dots, n$. Stocker les résultats dans une autre matrice.
- 3-3 Représenter sur un même graphique, les N courbes reliant les points $\{(i, \hat{\theta}_i), i = 1, \dots, n\}$
- 4-4 Interpréter ce graphique.
- 5-5 Utiliser une méthode de Monte Carlo pour estimer le biais de l'estimateur lorsque la taille de l'échantillon varie de 1 à n .
- 6-6 Représenter le biais estimé en fonction de la taille de l'échantillon. Commenter.
- 7-7 Utiliser une méthode de Monte Carlo pour estimer la variance de l'estimateur lorsque la taille de l'échantillon varie de 1 à n .
- 8-8 Représenter la variance estimée (notée $\widehat{\text{Var}}(j)$) en fonction de la taille de l'échantillon $j = 1, \dots, n$.
- 9-9 Représenter le nuage de points

$$\{(\log(j), \log(\widehat{\text{Var}}(j))), j = 1, \dots, n\}$$

[représentation log-log]

- 10-10 Commenter les résultats et faire le lien avec la vitesse de convergence de l'estimateur.
- 11-11 Reprendre les questions précédentes avec $\theta_0 = 10$. Superposer sur un même graphique, la variance estimée pour les deux valeurs de θ_0 (utiliser la représentation log-log). Comparer les vitesses de convergence.

On veut maintenant comparer deux estimateurs sans biais de θ (vu en TD)

$$(1) \hat{\theta}'_n = \frac{n+1}{n} \max\{X_1, \dots, X_n\}$$

$$(2) \hat{\theta}''_n = \frac{2}{n} \sum_{j=1}^n X_j$$

B-1 En utilisant les questions précédentes, calculer et représenter une estimation de la variance de $\hat{\theta}'_j$ en fonction de $j = 1, \dots, n$.

B-2 Utiliser une méthode de Monte Carlo pour estimer la variance de l'estimateur $\hat{\theta}''_j$ lorsque la taille de l'échantillon varie de 1 à n .

B-3 Ajouter sur le graphe précédent, la variance estimée de $\hat{\theta}''_j$ en fonction de $j = 1, \dots, n$.

B-4 Commenter les résultats obtenus.

B-5 Comparer les vitesses de convergence à l'aide de la représentation log-log des deux variances.

B-6 Conclure.

On souhaite maintenant illustrer la convergence en loi.

C-1 Mettre en évidence que $\sqrt{n}(\hat{\theta}_n - \theta)$ ne converge pas en loi vers une variable gaussienne $\mathcal{N}(0, \sigma^2)$ ($\sigma > 0$)

C-2 Illustrer la convergence en loi de $n(\hat{\theta}_n - \theta)$ vers une variable non gaussienne.

C-3 Donner une estimation de la densité de la loi limite de $n(\hat{\theta}_n - \theta)$.

C-4 Comparer la loi limite de $n(\hat{\theta}_n - \theta)$ avec celle de $\sqrt{n}(\hat{\theta}''_n - \theta)$.

Quatrième partie 4. Maximum de vraisemblance

EXERCICE 6.

- (1) Simuler un échantillon de taille 100 suivant la loi gaussienne de moyenne zéro et de variance 4.
- (2) Créer une fonction pour calculer la vraisemblance de l'échantillon simulé.
- (3) Représenter la vraisemblance
Indications : utiliser la fonction `outer` et les fonctions graphiques `persp` ou `image`
- (4) Donner une valeur approchée de l'estimateur du MV.
- (5) Reprendre les questions précédentes en augmentant la taille de l'échantillon (1000, puis 10 000).
- (6) Commenter les résultats obtenus.

EXERCICE 7.

- (1) Charger la librairie
- MASS**
- et les données
- quine**
- .

indication : utiliser les fonctions **library**, **data**

On cherche à modéliser le nombre de jours d'absence. Les observations sont disponibles dans la variable **Days**.

- (2) Donner une estimation de la moyenne, puis de la variance de cette variable.
 (3) Représenter l'histogramme des données
 (4) Dédire des résultats précédents une modélisation possible.
 (5) Pour le modèle choisi, tracer la fonction de vraisemblance.
 (6) En déduire une valeur approchée de l'estimateur du MV
 (7) Superposer l'histogramme et la loi obtenue pour modéliser ces données.
 (8) Commenter

EXERCICE 8. LE MODÈLE LOGISTIQUE

On rappelle que la densité d'une loi logistique de paramètre $\theta \in \mathbb{R}$ est

$$f(x) = \frac{e^{-(x-\theta)}}{(1 + e^{-(x-\theta)})^2}, \quad x \in \mathbb{R}.$$

En TD nous avons montré les propriétés suivantes

- la densité de la loi est symétrique par rapport à θ donc $\mathbb{E}_\theta(X_1) = \theta$
- l'estimateur du maximum de vraisemblance existe et il est solution de

$$h_n(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log(f_\theta(X_i)) = 1 - \frac{2}{n} \sum_{i=1}^n \frac{e^{-(X_i-\theta)}}{1 + e^{-(X_i-\theta)}} = 0$$

- $h'_n(\theta) = \frac{-2}{n} \sum_{i=1}^n f_\theta(X_i)$
- $I(\theta) = 1/3$
- $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \pi^2/3)$

- 1) Montrer que la médiane $m(X_1)$ de X_1 est égale à θ ,

$$m(X_1) = F_\theta^{-1}(1/2) = \int_{-\infty}^{m(X_1)} f_\theta(x) \, dx = 1/2.$$

- 2) La médiane empirique $\hat{m}_n$ (voir ex 5) est un estimateur de la médiane qui vérifie la propriété suivante

$$\sqrt{n}(m_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4} \frac{1}{f(\theta)^2}\right)$$

Illustrer ce résultat par la simulation.

Indications :

- Simuler $N = 500$ échantillons de taille n ($= 50$ augmenter n pour visualiser la convergence) suivant la loi la loi logistique de paramètre $\theta = 1$ (utiliser la fonction `rlogis`).
- Calculer sur chaque échantillon $T_{n,i} = \sqrt{n}(m_n - \theta)$, pour $i = 1, \dots, N$.
- comparer la densité de la loi limite avec une estimation de la densité calculée sur l'échantillon $T_{n,1}, \dots, T_{n,N}$

On veut comparer les cinq estimateurs de θ suivants :

- la moyenne empirique $\bar{X}_n$,
- la médiane empirique m_n ,
- le maximum de vraisemblance $\hat{\theta}_n$
(indication : pour trouver le zéro d'une fonction, on pourra utiliser la fonction `uniroot` sous R , qui approche les zéros sur un intervalle que l'on fixe assez grand $[-10,10]$),
- l'estimateur asymptotiquement optimal associé à la moyenne empirique

$$\tilde{\theta}_n^1 = \bar{X}_n - \frac{h_n(\bar{X}_n)}{h'_n(\bar{X}_n)}$$

- l'estimateur asymptotiquement optimal associé à la médiane

$$\tilde{\theta}_n^2 = m_n - \frac{h_n(m_n)}{h'_n(m_n)}$$

- 3) Pour chacun des 5 estimateurs, construire une fonction dont le paramètre d'entrée est X un vecteur de longueur n et le paramètre de sortie est le vecteur constitué des estimations de θ calculées sur $(X_1, \dots, X_k)$ pour k variant de 10 à n par pas de $d = 10$. Indications : pour éviter des boucles !

```
est.1 = fonction(k,x)
{
  calcul de 1 estimateur sur x[1:k]
}
```

```
est.V1 =function(x,d)
{
  sapply(seq(d,n,d) , est.1 ,x)
}
```

- 4) Simuler un échantillon de taille $n = 300$, calculer et représenter l'évolution des estimateurs en fonction du nombre d'observations pour k variant de d à n par pas de $d = 10$.

On veut maintenant comparer les variances des 5 estimateurs.

- 5) Reprendre la question précédente sur $N = 200$ échantillons de taille $n = 100$ (il faudra éventuellement augmenter $n = 200$ et $N = 500$ pour bien estimer les variances des

lois limites. Si le temps de calcul est trop long augmenter le pas d entre deux valeurs calculées) .

Pour chaque estimateur, stocker les résultats dans une matrice à N colonnes. Soit η_k un des estimateurs calculé sur les k premières observations. Soit $\eta_k(1), \dots, \eta_k(N)$ les N estimations obtenues sur les N échantillons. On note $\widehat{\text{Var}}(\eta_k)$ la variance empirique de $\eta_k(1), \dots, \eta_k(N)$.

Remarque : c'est aussi la variance empirique de $\eta_k(1) - \theta, \dots, \eta_k(N) - \theta$

- 6) Pour chacun des estimateurs, calculer $\widehat{\text{Var}}(\eta_k)$ pour k variant de 1 à n par pas de d .
- 7) Superposer les courbes $(k, \widehat{\text{Var}}(\eta_k))$ des 5 estimateurs
- 8) Superposer les courbes $(\log(k), \log(\widehat{\text{Var}}(\eta_k)))$ des 5 estimateurs
- 9) Superposer les courbes $(k, k * \widehat{\text{Var}}(\eta_k))$ des 5 estimateurs
- 10) Les résultats des simulations sont-ils en accord avec la théorie ?

Cinquième partie 5. Régions et intervalles de confiance

EXERCICE 9. SITUATION GAUSSIENNE

Soit $X_1, \dots, X_n$ un n -échantillon distribué suivant la loi normale de paramètres (μ, σ^2) . On veut estimer μ .

On note

— q_α le quantile d'ordre α de la loi normale standard

— $t_{n,\alpha}$ le quantile d'ordre α de la loi de Student à n degrés de liberté.

si σ^2 est connue,

$$I_n = [\bar{X}_n - \frac{q_{1-\alpha/2}\sigma}{\sqrt{n}}; \bar{X}_n + \frac{q_{\alpha/2}\sigma}{\sqrt{n}}]$$

est un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre μ c'est à dire

$$P_\mu(\mu \in I_n) = 1 - \alpha$$

si σ^2 est inconnue, on pose

$$\sigma_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

et

$$I'_n = [\bar{X}_n - \frac{t_{n-1,1-\alpha/2}\hat{\sigma}_n}{\sqrt{n}}; \bar{X}_n + \frac{t_{n-1,\alpha/2}\hat{\sigma}_n}{\sqrt{n}}]$$

est un intervalle de confiance de niveau $1 - \alpha$.

- 1) Simuler $N = 1000$ échantillons de taille $n = 10$ iid suivant la loi normale de paramètres $(1, 1)$
- 2) Calculer pour chaque échantillon les régions de confiance I_n et I'_n au niveau 95%
- 3) Représenter les régions de confiance (on peut se limiter aux 100 premières) et utiliser des couleurs pour distinguer les intervalles qui contiennent $\mu = 1$.

- 4) Evaluer la fréquence de l'évènement μ appartient à l'intervalle de confiance. Commenter le résultat.
- 5) Reprendre les questions 1) -> 4) avec $n = 50$

EXERCICE 10. SITUATION ASYMPTOTIQUE

On suppose que $X_1, \dots, X_n$ sont iid, L^2

On construit un intervalle de confiance sur la moyenne à partir de la loi limite de

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X_1))$$

On obtient un intervalle asymptotiquement de niveau $1 - \alpha$ en prenant

$$I_n(X_1, \dots, X_n) = \left[\bar{X}_n - q_{1-\alpha/2} \frac{S_n}{\sqrt{n}} ; \bar{X}_n - q_{\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

où

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{et} \quad S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

On considère $(X_1, \dots, X_n)$ iid suivant la loi de student à 3 ddl. On veut estimer le niveau exact de l'intervalle de confiance $\beta_n = P_\theta(\theta \in I_n)$ en fonction de n par la simulation. On utilise la méthode de Monte Carlo suivante :

- étape 1 On simule N échantillons indépendants de taille n suivant la loi de Student à 3 ddl
- étape 2 Pour chacun des échantillons, on calcule l'intervalle de confiance : $I_n(i)$, $i = 1, \dots, N$
- étape 3 On approche la probabilité β_n par

$$\hat{\beta}_n(N) := \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{I_n(i)}(\theta)$$

avec N grand.

- 1) Justifier la méthode
- 2) Pour $n = 15$, simuler $N = 10000$ échantillons et représenter $\hat{\beta}_n(N)$ en fonction de N . En déduire une estimation de β_n
- 3) Reprendre la question précédente pour $n = 25, 50, 100, 500$. Superposer les différentes courbes.
- 4) Faire le lien entre les résultats numériques et le niveau asymptotique des intervalles de confiance.

EXERCICE 11. CHOIX DE L'ESTIMATEUR DE LA VARIANCE

Soit $X_1, \dots, X_n$ iid suivant la loi exponentielle de paramètre θ^{-1}

- La densité s'écrit $f_\theta(x) = \theta^{-1} e^{-x\theta^{-1}} \mathbb{I}_{\mathbb{R}^+}(x)$
- On a $E_\theta(X) = \theta$ et $\text{Var}_\theta(X) = \sigma^2 = \theta^2$
- Convergence en loi

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \theta) \xrightarrow{\text{loi}} N(0, 1)$$

On peut estimer la variance σ^2 par les estimateurs consistants suivant :

- la variance empirique S_n^2
- la méthode des moments : on estime $\sigma^2 = h(\theta)$ par $h(\bar{X}_n) = \bar{X}_n^2$

On peut donc construire les intervalles de confiance de niveau asymptotique $1 - \alpha$ suivant

$$I_n(X_1, \dots, X_n) = \left[\bar{X}_n - q_{1-\alpha/2} \frac{S_n}{\sqrt{n}} ; \bar{X}_n - q_{\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

ou

$$I'_n(X_1, \dots, X_n) = \left[\bar{X}_n - q_{1-\alpha/2} \frac{\bar{X}_n}{\sqrt{n}} ; \bar{X}_n - q_{\alpha/2} \frac{\bar{X}_n}{\sqrt{n}} \right]$$

On note

$$\beta_n = P_\theta(\theta \in I_n)$$

$$\beta'_n = P_\theta(\theta \in I'_n)$$

- 1) Pour $\theta = 1$ et $n = 25, 50, 100, 200$, simuler $N = 10000$ échantillons et superposer les courbes $N \mapsto \hat{\beta}_n(N)$ et $N \mapsto \hat{\beta}'_n(N)$. En déduire des valeurs approchées de β_n et β'_n .
- 2) Même question pour $\theta = 2$
- 3) Commenter les résultats obtenus

Sixième partie 6. Test paramétrique**EXERCICE 12.**

Écrire des fonctions permettant de tester l'égalité des moyennes et l'égalité des variances de deux échantillons gaussiens.

EXERCICE 13.

On désire tester si un médicament a une influence sur le comportement psychomoteur. On choisit au hasard 20 sujets qu'on répartit au hasard en deux groupes : le groupe témoin et le groupe expérimental. On leur fait subir la même expérience psychomotrice. On a administré auparavant le médicament aux sujets du groupe expérimental et un placebo au groupe témoin. Les résultats sont les suivants :

Groupe témoin	166	167	169	170	174	173	172	170	166	173
Groupe expérimental	167	162	165	168	162	164	162	160	165	169

On suppose que dans chaque groupe les résultats sont distribués selon une loi gaussienne et que les performances des sujets sont indépendantes.

- (1) Comparer les variances des deux échantillons
- (2) Tester au niveau 0.05 l'hypothèse selon laquelle le médicament n'a aucun effet sur le comportement psychomoteur en utilisant la fonction créée à l'exercice précédent, puis à l'aide de la fonction `t.test`.

EXERCICE 14.

On a fait une numération globulaire à un groupe de 20 personnes à deux périodes différentes de l'année. Pour chaque sujet on note les résultats des deux numérations (à multiplier par 10^5) :

Sujet	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Janvier-février	46	38	42	47	48	40	40	43	42	49	45	51	47	52	50	48	47	47	47	45
Sept.-oct.	48	47	44	45	51	44	47	48	47	57	49	55	48	48	46	48	54	54	44	48

On suppose que les sujets sont mutuellement indépendants et suivent une loi gaussienne. Tester au niveau 0.05 l'hypothèse selon laquelle les résultats de la numération sont les mêmes aux deux périodes.

EXERCICE 15.

On jette une pièce 5 fois et on note N le nombre de fois où la pièce tombe sur PILE. Soit p la probabilité que la pièce tombe sur PILE. On veut tester $p = 1/2$ contre $p > 1/2$.

- 1) Peut on construire un test UPP(5%) ?
- 2) A quel niveau α peut on obtenir un test UPP(α) ?
- 3) Choisir le niveau le plus proche de 5%.
- 4) Écrire une fonction qui retourne la décision du test au niveau choisi.
- 5) Simuler $N_1, \dots, N_n$ un n -échantillon distribué suivant la loi binomiale de paramètres $(5, \frac{1}{2})$ (prendre $n=1000$)
- 6) Calculer la décision du test pour chacune des observations N_k , $k = 1, \dots, n$.
- 7) Calculer la fréquence de l'évènement *on rejette l'hypothèse nulle* et commenter le résultat obtenu.
- 8) Reprendre les questions 6) et 7) lorsque la loi de l'échantillon $N_1, \dots, N_n$ est la loi binomiale de paramètres $(5, p)$ avec $p = 0.6; 0.7; 0.8; 0.9$
- 9) Commenter les résultats obtenus
- 10) Reprendre les questions précédentes pour $K = 50; 100$.
- 11) Commenter les résultats.