UGA – L2/STA401

TP5 : Estimation paramétrique

Exercice pour l'extraction de données :

1. Charger le jeu de données cardiaque.csv et l'affecter à `cardiaque`. En extraire l'échantillon des pressions systoliques chez les patients ayant un BMI supérieur ou égal à 23 et en donner les résumés numériques usuels : taille, moyenne, écart-type empirique corrigé (noté s' dans le cours) et quartiles.
2. Vérifiez que vous pouvez charger correctement l'ensemble des données présentes dans Chamilo Données-TP, regardez pour cela `chargementData.R`.

Objectifs : *Etudier numériquement les propriétés d'estimateurs d'un ou plusieurs paramètres de la loi d'une variable continue.*

1 Modèle gaussien

Dans cette partie on considère un échantillon i.i.d. de X pour X de loi normale $\mathcal{N}(\mu, \sigma^2)$ et de taille n où μ et σ sont les paramètres inconnus du modèle.

Exercice 1 : Estimer l'espérance dans le Modèle Normal

Le but ici est l'étude et la comparaison de deux estimateurs du paramètre μ : le premier T_1 est l'estimateur usuel $\bar{X}_n$ et le second que l'on notera T_2 est defini par $T_2 = (\min(X_i) + \max(X_i))/2$.

1. Définir $\mu = 3$, $\sigma = 1$, $N = 1000$ et $n = 10$. Générer N tirages d'échantillons de taille n et les affecter dans une matrice nommée `dataG` à N lignes et n colonnes.
2. Calculer les N réalisations de T_1 et de T_2 obtenues pour les N échantillons et les affecter à `est1` et `est2`. Pour cela, utiliser la fonction `apply` pour appliquer une fonction sur l'ensemble des lignes, par exemple `apply(dataG, MARGIN=1, sum)` pour avoir N réalisations de $\sum_{i=1}^{n} X_i$.
3. Calculer la moyenne empirique de `est1` (resp. de `est2`) et l'affecter à `moyest1` (resp. de `moyest2`). Calculer les écart-type empiriques corrigés (avec `sd()`) de chacune des deux séries `est1` et `est2` et les affecter à `etest1` et `etest2`.
4. Au vu des résultats obtenus peut-on dire que T_1 (resp. T_2) est sans biais (re-générez plusieurs autres échantillons aléatoire pour vous en convaincre)? Lequel de ces deux estimateurs a la plus faible variabilité ? En déduire le meilleur de ces deux estimateurs pour μ (meilleur au sens de "celui qui est sans biais et de variance minimale").
5. Représenter pour finir dans une même fenêtre et l'un au dessus de l'autre, les histogrammes des vecteurs `est1` et `est2` et y ajouter en rouge la verticale passant par le points d'abscisse μ et en vert la verticale qui passe par la moyenne empirique de l'échantillon. Les limites des axes des abscisses et des ordonnées devront être les mêmes pour les deux histogrammes, afin qu'ils soient visuellement facilement comparables. Commenter.
6. Refaire ces deux graphiques avec $n = 50$. La différence entre les deux estimateurs est-elle plus nette que lorsque $n = 10$?

Exercice 2 : Estimer la variance dans le Modèle Normal

On s'intéresse ici aux deux estimateurs usuels de σ^2 que l'on notera comme précédemment $T_1 = S_n^2 = (\sum_{i=1}^n X_i^2)/n - \bar{X}_n^2$ (variance empirique de l'échantillon aléatoire) et $T_2 = S_n'^2 = T_1 n/(n-1)$ (variance empirique corrigée).

1. On garde $\mu = 3$, $\sigma = 1$, $N = 1000$ et $n = 10$. Générer N tirages d'échantillons de taille n et les affecter dans une matrice nommée `dataG` à N lignes et n colonnes.
2. Calculer les N réalisations de T_1 et de T_2 obtenues pour les N échantillons et les affecter à `est1` et `est2` (rappel : utiliser la fonction `var()` pour calculer T_2 et en déduire ensuite le calcul de T_1).
3. Représenter sur une même page les histogrammes des vecteurs `est1` et `est2` et y ajouter en rouge la verticale passant par le points d'abscisse σ^2 et en pointillés vert la verticale qui passe par la moyenne empirique des N réalisations de l'estimateur étudié. Pour le graphique décrivant T_1 on ajoutera aussi en trait plein vert la verticale passant par le point d'abscisse $(n-1)\sigma^2/n$. Quelles propriétés des estimateurs ces graphiques mettent-ils en évidence (biais, variance et lois des estimateurs) ? Superposer la densité g de la loi théorique de T_1 à l'histogramme précédent qui porte sur T_1 et la densité h sur l'histogramme de T_2. Rappelons que la densité de la loi théorique de T_1 est la fonction $g(t) = n/\sigma^2 f(tn/\sigma^2)$ (ex. à faire en TD) où f désigne ici la densité d'un chi-deux à $n-1$ degrés de liberté puisque nT_1/σ^2 suit une loi $\mathcal{X}_{n-1}$ (la commande R pour calculer sa densité en x est `dchisq(x,df=n-1)`). Celle de T_2 est $h(t) = (n-1)/\sigma^2 f(t(n-1)/\sigma^2)$.
4. (facultatif) Afficher dans une fenêtre coupées en 6 morceaux (2 lignes et 3 colonnes) les histogrammes de T_1 obtenus pour les valeurs $10, 20$ et 100 de n en première ligne et ceux obtenus pour T_2 en deuxième ligne. Comparer et commenter.

2 Autres modèles

Exercice 3 : Modèle de Bernoulli

Dans cet exercice on considère un échantillon i.i.d. de X pour X de loi de Bernoulli $\mathcal{B}(p)$ avec $p \in]0,1[$ et de taille n. Avec R la loi de Bernoulli de paramètre p s'obtient comme un cas particulier de la loi binomiale : $\mathcal{B}(1,p)$. Par exemple, sa probabilité en k ($k = 0$ ou $k = 1$) sera calculée avec `dbinom(k,1,p)`.

1. Définir $p = 0.5$, $N = 1000$ et $n = 4$. Générer N tirages d'échantillons de taille n et les affecter dans une matrice nommée `dataB` à N lignes et n colonnes.
2. Calculer les N réalisations de l'estimateur usuel de p : $F_n = \bar{X}_n$ obtenues pour les N échantillons de taille n tirés. Affecter le résultat à `est`.
3. Représenter l'histogramme de `est` et y superposer la densité d'une $\mathcal{N}(p, p(1-p)/n)$.
4. Essayer plusieurs valeurs de n (entre 4 et 100) et apprécier graphiquement à partir de quelle valeur de n on peut considérer que la loi de l'estimateur F_n est normale. L'estimateur F_n est-il sans biais quelque soit n ?

Exercice 4 : Modèle Uniforme (facultatif)

Dans cet exercice on considère un échantillon i.i.d. de X pour X de loi de Uniforme $\mathcal{U}[0,a]$ avec $a \in]0,\infty[$ et de taille n. Avec R la loi de uniforme est décrite avec la fonctions `dunif`.

1. Définir $a = 2$, $N = 1000$ et $n = 10$. Générer N tirages d'échantillons de taille n et les affecter dans une matrice nommée `dataU` à N lignes et n colonnes.
2. Calculer les N réalisations de l'estimateur usuel de a : $T_1 = 2\bar{X}_n$ et $T_2 = \max(X_i)$.
3. Etude du biais, de la variance et de la loi des deux estimateurs proposés.