Université Grenoble Alpes
L2/STA351

TP3 : Données Simulées et Estimation

Objectifs : Savoir calculer une probabilité de façon exacte et savoir l'estimer à l'aide de simulations. Comprendre la propriété d'estimateur sans biais et celle d'estimateur convergent (sans biais et de variabilité décroissante avec la taille de l'échantillon).

1 Calculs de probabilités théoriques ou empiriques

Exercice 1

On sait par expérience qu'une certaine opération chirurgicale a 60% de chance de succès. On réalise cette opération chez 4 patients. Soit X la variable aléatoire qui modélise le nombre de succès parmi ces 4 patients.

1. Quelle loi (quel modèle) proposez vous pour X ?
2. On veut calculer la probabilité que l'opération échoue les 4 fois.
   (a) Ecrire mathématiquement cette probabilité.
   (b) La calculer directement avec RStudio, en ayant spécifié au préalable les valeurs de n et p et en utilisant la densité de la loi binômiale `dbinom()`.
   (c) On va à présent évaluer cette probabilité de manière empirique c'est à dire en utilisant les résultats observés de la même expérience réalisée un nombre N de fois (on choisira N grand). Pour cela :
      i. Définir les valeurs pour n, p et le nombre N de tirages (taille de l'échantillon) :
      `n <- 4; p <- 0.6 # parametres  N <- 100 # taille de l'échantillon`
      ii. Simuler N tirages d'une binômiale de paramètres (n, p) avec `rbinom()` et les affecter à `x`.
      iii. Calculer les fréquences observées de chaque modalité dans votre échantillon `x`. Combien de fois la modalité 0 est-elle apparue ?
      iv. Visualiser la répartition de l'échantillon à l'aide d'un diagramme en barres. Y ajouter en rouge la droite horizontale qui passe par la valeur calculée dans la question 2 b). Que se passe-t-il lorsque l'on augmente N ?
   (d) On s'intéresse maintenant à la probabilité que l'opération échoue exactement 2 fois parmi les 4 patients.
      i. Calculer exactement cette probabilité avec `dbinom()`.
      ii. A quelle modalité correspond l'évènement d'intérêt ? Calculer la fréquence empirique de cet événement sur un échantillon de taille $N = 10\,000$.

Exercice 2

La taille X des hommes en France est modélisée par une loi normale $\mathcal{N}(172, 196)$ (unité : cm).

1. On défini les paramètres connus du modèle avec avec : `mfm <- 172; sdfm <- sqrt(196)`
2. Quelle proportion d'hommes français mesurent moins de 160cm ? Affecter le résultat à `p`.

3. Visualiser cette probabilité sur les deux graphiques suivants :

```
par(mfrow=c(1,2)) # partage la fenetre graphique en deux
curve(dnorm(x,mean=mfm,sd=sdfm),from=mfm-3*sdfm, to=mfm+3*sdfm) # fonction densité
abline(v=mfm,col="red") # valeur moyenne
abline(v=160,col="blue") # valeur consideree
curve(pnorm(x,mean=mfm,sd=sdfm), from=mfm-3*sdfm, to=mfm+3*sdfm) # fonction de repartition
théorique
abline(v=160,col="blue") # valeur consideree
abline(h=p,col="blue") # probabilite
```

4. Quelle proportion d'hommes français mesurent plus de 2 mètres ?
5. Quelle proportion d'hommes français mesurent entre 165 et 185 cm ?
6. Si 10 000 hommes français étaient choisis au hasard, et rangés par ordre croissant de leur taille, quelle serait la taille du 9000-ième (la fonction quantile de la normale est `qnorm()`) ?

2 Etude de propriétes d'estimateurs par simulations

Exercice 3 – Intervalle de fluctuation pour la moyenne empirique

Soit X une variable aléatoire $\mathcal{N}(0,1)$. On va simuler N échantillons de taille n et évaluer les fréquences de réalisation d'un évènement sur N n-échantillons (chacun de de taille n).

1. Définir n et N. Calculer l'intervalle de fluctuation de niveau 90% de la variable $\bar{X}_n$.
2. Simuler N échantillons de taille n avec $N * n$ tirages rangés dans une matrice à N lignes et n colonnes avec (on affecte le résultats à `Gdata`) :
   `Gdata<-matrix(rnorm(N*n,0,1),ncol=n)`.
3. Calculer la moyenne empirique de chaque n-échantillon (ligne de la matrice de données `Gdata`) avec la fonction `rowMeans()` et l'affecter à `xbar`. Quelle est sa dimension ?
4. Calculer également la moyenne des carrés pour chaque échantillon puis la variance empirique (moyenne des carrés moins carré de la moyenne) et l'affecter à `s2`. Calculer ensuite les variances estimées sans biais et les affecter à `s'2`.
5. Combien de fois parmi N la moyenne empirique est-elle tombée dans l'intervalle précédent ? Que se passe-t-il lorsque N augmente ?

Exercice 4

On reprend les données précédemment simulées et on se propose d'étudier empiriquement les comportements de $\bar{X}$, S^2 et S'^2.

1. $\bar{X}_n$ **Estimateur de** μ
   (a) Le biais de l'estimateur $\bar{X}$ est $E(\bar{X}) - 0 = 0$ (on le sait d'après le calcul des probabilités). Pour le vérifier de façon numérique calculer la différence entre la moyenne des estimations obtenues pour les N n-échantillons et la valeur que l'on cherche à estimer (ici $\mu = 0$). Observer l'évolution de cette différence lorsque l'on augmente N puis lorsque l'on augmente n.
   (b) Représenter la répartition des N réalisations de $\bar{X}_n$ avec un histogramme et y ajouter une verticale en vert qui passe par la moyenne empirique des N valeurs de $\bar{X}_n$ et une verticale rouge qui passe par la vraie valeur $\mu = 0$. Ré-exécuter plusieurs fois le script depuis la simulation des données jusqu'au instructions permettant de réaliser le graphique en changeant N et (ou) n. A quoi ressemble la répartition observée pour n petit, ou n grand ? Quelle densité peut-on proposer pour "fitter" au mieux l'histogramme ? On pourra choisir à priori les valeurs 0 et $1/\sqrt{n}$ comme paramètres de la loi ajustée (la tracer en rouge) ou les valeur estimées (en vert) $\hat{\mu}$ et $\hat{\sigma}/\sqrt{n}$ (calculées par `mean()` et `sd()/sqrt(n)`). Que permet de faire N ?

(c) Variabilité de $\bar{X}_n$: la calculer de façon exacte et l'évaluer sur les données précédemment simulées. Que se passe-t-il lorsque n augmente (on laissera $N = 10000$) ?

2. **S_n^2 et $S_n'^2$ estimateurs de σ^2**

(a) Tracer la répartition observée des N réalisations de S_n^2 qui ont été affectées à `s2` (calculées dans la question 4 de l'exercice précédent). Y ajouter la verticale rouge qui passe par la vraie valeur du paramètre (soit $\sigma^2 = 1/n$) et celle qui passe par la moyenne de `s2` en vert. Faire tourner le script pour $n = 10$ et $N = 10000$. L'estimateur S^2 est-il sans biais ?

(b) Refaire la question précédente avec $S_n'^2$. Est-ce un estimateur sans biais de σ^2 (qui ici vaut 1) ?