A rendre pour le mercredi 1er mars 2006

1 Estimateurs (inspirés de [1])

1.1 Estimateur de variance minimum *correction vue en TD*

Soit $X_1, \ldots, X_n$ un échantillon i.i.d d'une loi de moyenne μ et de variance σ^2.

1. Donner la condition sur les constantes réelles $a_1, \ldots, a_n$ pour que $\sum_{i=1}^n a_i X_i$ soit un estimateur sans biais de μ.
   Réponse : $\sum_{i=1}^n a_i = 1$.
2. Parmi les estimateurs sans biais de cette forme déterminer celui qui est de variance minimum. Calculer sa variance.
   Réponse : $a_1 = \ldots = a_n = \frac{1}{n}$.

1.2 Estimateurs sans biais

Sachant que X se distribue selon une loi normale de variance σ^2 et connaissant les variances et les tailles de trois échantillons, à savoir, $s_1^2 = 8$, $s_2^2 = 10$, $s_3^2 = 14$, et $n_1 = 10$, $n_2 = 7$ et $n_3 = 5$

1. montrer que, dans le cas de k échantillons, l'estimateur

$$T = \frac{n_1 s_1^2 + \ldots + n_k s_k^2}{n_1 + \ldots + n_k}$$

   est un estimateur sans biais de σ^2 ;
   Réponse. Il s'agit ici des variances empiriques, réalisations des variables aléatoires $S_1, \ldots, S_k$, *avec :*

$$S_p^2 = \frac{1}{n_p - 1} \sum_{p=1}^{n_p} (X_i^p - \bar{X}_p)^2$$

   où X_i^p *est la* i*-ème variable aléatoire (de même loi que* X*) du* p*-ième échantillon et* $\bar{X}_p$ *la moyenne des v.a. de ce même échantillon. Remarque 1 : par souci d'homogénéité avec le cours, on aurait dû utiliser des lettres majuscules pour désigner les variables aléatoires et des minuscules pour leurs réalisations. Remarque 2 : lorsqu'on parle d'échantillon reposant sur l'utilisation d'une variable aléatoire, cela veut dire que l'on considère un ensemble de v.a. qui suivent toutes la même loi que* X *et qui sont indépendantes (i.e., si* n *est la taille de l'échantillon, cela veut dire que l'on considère un ensemble* n *variables aléatoires i.i.d selon la loi de* X*).*
   Comme on l'a vu en TD, on a : $E(S_p) = \sigma^2$*, d'où le résultat immédiat :*

$$E(T) = \sigma^2.$$

2. déterminer une estimation sans biais de σ^2 en utilisant n_1, n_2 et n_3 ;
   Réponse. On demande une estimation, c'est-à-dire une réalisation de l'estimateur T *défini sur les 3 échantillons :*

$$t = \frac{10 \cdot s_1^2 + 7 \cdot s_2^2 + 5 \cdot s_3^2}{10 + 7 + 5} = 10$$

3. montrer que

$$\frac{1}{5}(2s_1^2 + 2s_2^2 + s_3^2)$$

   est aussi une estimation sans biais de σ^2 ;
   Réponse : trivial, linéarité de l'espérance.
4. quel est le meilleur estimateur de celui utilisé en 2 et celui utilisé en 3 ?
   Réponse. Si σ^2 *est la variance de* X *alors pour chaque* p*,* $\frac{n_p-1}{\sigma_p^2} S_p^2$ *est une variable aléatoire dont la loi est celle du* χ^2 *à* $n_p - 1$ *degrés de liberté. En utilisant le fait que* $\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2)$ *lorsque* X_1 *et* X_2 *sont indépendantes (vérifiez-le !) et le fait que* $\mathbb{V}(aX) = a^2\mathbb{V}(X)$*, on a le résultat désiré.*

1.3 Estimateur sans biais, loi uniforme

Une v.a. aléatoire X suit une loi uniforme sur $[0;\theta]$ a une densité de probabilité :

$$f(x) = \frac{1}{\theta}.$$

Montrer que $2\bar{X}$ est un estimateur sans biais de θ.
Réponse. Soit n>0. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *où tous les* X_i *sont i.i.d selon la loi uniforme sur* $[0;\theta]$.

$$\begin{aligned} E(2\bar{X}) &= 2E(X) \\ &= 2\int_0^\theta x f(x) dx \\ &= 2\int_0^\theta \frac{x}{\theta} \\ &= 2\left[\frac{x^2}{2\theta}\right]_0^\theta \\ &= \theta, \end{aligned}$$

d'où le résultat.

2 Estimateurs du maximum de vraisemblance (cf. [1])

2.1 Loi de Poisson

Déterminer l'estimateur de maximum de vraisemblance de m pour une loi de Poisson de paramètre m fondé sur n observations. Quel est le biais de cet estimateur ?
Réponse. Rappel sur la loi de Poisson :

$$P(K = k) = e^{-m}\frac{m^k}{k!}.$$

Quelques calculs nous donnent :

$$m^{MV} = \frac{sum_{i=1}^{n}}{n}.$$

C'est un estimateur sans biais.

2.2 Lois exponentielles

Cet exercice s'intéresse aux estimateurs de vraisemblance de quelques lois exponentielles.

1. Soit la densité de probabilité

$$p_\theta(x) = (1+\theta)x^\theta, \quad \theta > -1, 0 < x < 1.$$

   (a) Montrer que p_θ définit bien une densité de probabilité.
   Réponse. Il suffit de montrer que p_θ *est positive et que sont intégrale sur le domaine de définition est de 1.*

   (b) Donner l'estimateur de maximum de vraisemblance de θ à l'aide d'un échantillon de taille n.
   Réponse. On calcule la log-vraisemblance (on se rappelle que puisque $\log$ *est une fonction strictement croissante, le maximum de la log-vraisemblance correspond à celui de la vraisemblance) :*

$$\begin{aligned} \log \mathcal{L}(x_1,\ldots,x_n;\theta) &= \log\left[(1+\theta)^n (x_1\cdots x_n)^\theta\right] \\ &= n\log(1+\theta) + \theta\sum_{i=1}^{n}\log x_i, \end{aligned}$$

Une condition nécessaire pour l'atteinte d'un maximum est l'annulation de la dérivée de la log-vraisemblance par rapport à θ :

$$\frac{d\log\mathcal{L}(x_1,\ldots,x_n;\theta)}{d\theta}=0\Rightarrow\frac{n}{1+\theta}+\sum_{i=1}^{n}\log x_i=0$$
$$\Leftrightarrow\theta^{MV}=-1-\frac{n}{\sum_{i=1}^{n}\log x_i}$$

On vérifie aisément que $\theta^{MV}>-1$ et que la dérivée seconde de la log-vraisemblance en $theta^{MV}$ est bien négative (θ^{MV} correspond bien à un maximum).

2. Soit la densité de probabilité
$$p_\theta(x)=\theta e^{-x\theta},\quad x>0.$$
   (a) Montrer que p_θ définit bien une densité de probabilité.
   Réponse : idem que précédemment.
   (b) Donner l'estimateur de maximum de vraisemblance de θ à l'aide d'un échantillon de taille n.
   Réponse. En procédant de manière similaire à ce qui a été fait précédemment, on a :
$$\theta^{MV}=\frac{n}{\sum_{i=1}^{n}x_i}$$
   (c) Soit $X\sim p_\theta$. Donner un estimateur de $\mathbb{E}[X]$.
   Réponse. Un rapide calcul nous donne que
$$\mathbb{E}[X]=\frac{1}{\theta}$$
   .
   Afin de définir un estimateur de maximum de vraisemblance de $\mathbb{E}[X]$, il suffit de voir que $g:x\mapsto\frac{1}{x}$ est bijective sur $]0;+\infty[$ et que, par conséquent, $g(\theta^{MV})$ est un estimateur de maximum de vraisemblance de $g(\theta)$. En somme, on peut déduire un estimateur de maximum de vraisemblance de $\frac{1}{\theta}$ et donc de $\mathbb{E}[X]$, par
$$\mathbb{E}[X]^{MV}=\frac{\sum_{i=1}^{n}x_i}{n}.$$

3 Intervalles de confiance (cf. [2])

3.1 Cours *corrigé en TD*

Soit $X\sim\mathcal{B}(n,\pi)$. Dans cet exercice, nous nous intéressons à la démonstration du résultat de cours donnant un intervalle de confiance de niveau $1-\alpha$ pour π de la forme

$$IC_{1-\alpha}=\left[\frac{\frac{x}{n}+\frac{u_{1-\alpha/2}^2}{2n}\pm\frac{u_{1-\alpha/2}}{\sqrt{n}}\sqrt{\frac{u_{1-\alpha/2}^2}{4n}+\frac{x}{n}\left(1-\frac{x}{n}\right)}}{1+\frac{u_{1-\alpha/2}^2}{n}}\right]\tag{1}$$

où $u_{1-\alpha/2}$ est le quantile de niveau $1-\alpha/2$ de la loi normale centrée réduite.

1. Quand peut-on considérer que la loi de X peut être approchée par une loi normale ? Quels sont les paramètres de cette loi normale ? En supposant l'approximation normale valide, donner une statistique pivotale T^{pv} permettant de former un intervalle de confiance de niveau $1-\alpha$ pour π.
2. Suivant le cours, l'intervalle de confiance de niveau $1-\alpha$ pour π s'obtient à partir de la statistique pivotale T^{pv} et de la propriété que $P(t_{\alpha/2}<T^{pv}<t_{1-\alpha/2})=1-\alpha$, t_a désignant le quantile de niveau a de la loi de T^{pv}. Préciser pourquoi $t_{\alpha/2}=-u_{1-\alpha/2}$ et $t_{1-\alpha/2}=u_{1-\alpha/2}$.
3. Déduire des questions précédentes l'expression (1) de $IC_{1-\alpha}$ en précisant chaque étape du calcul.

3.2 Temps de réaction

La mesure du temps de réaction, supposé suivre une loi normale, de plusieurs sujets sur une expérience donnée fournit un écart-type de 0.05s. Quelle doit être la taille de l'échantillon de mesures pour que l'erreur de l'estimation du temps de réaction n'excède pas 0.01s à 95% ? à 99% ? *Réponse. Ici, on ne précise pas que 0.05 est l'écart-type réel. On suppose donc que c'est l'écart-type empirique, que l'on note s. De fait, afin de déterminer un intervalle de confiance dans cette situation, on utilisera les quantiles de la loi de Student.*

Pour α donné, on voudrait être sûr avec une probabilité $1-\alpha$ de ne pas faire une erreur de plus de 0.01s sur l'estimation du temps de réaction des sujets. Pour cela, il suffit de déterminer un intervalle $IC_{1-\alpha}$ de confiance $1-\alpha$ faisant une longueur $|IC_{1-\alpha}|$ de 0.01 au maximum. Donc, puisque :

$$IC_{1-\alpha} = [\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}],$$

on a

$$\begin{aligned}|IC_{1-\alpha}| &= \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} - \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \\ &= 2t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}},\end{aligned}$$

et on cherche n tel que

$$2t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq 0.01$$

Il suffit alors de chercher dans la table de Student le quantile (et le n) qui convient.

Pour $\alpha = 0.05$ et $\alpha = 0.01$, on se rend compte en cherchant dans la table de Studentles quantiles d'intérêt, on se rend compte que le nombre d'observations doit être au moins de ... On peut ainsi se reporter à la table de la loi normale centrée réduite pour chercher les quantiles qui nous intéressent. Cela nous donne les résultats suivants :

- *pour $\alpha = 0.05$: $n \geq \left(\frac{1.96 \cdot 0.05}{0.005}\right)^2$, soit $n \geq 385$;*
- *pour $\alpha = 0.01$: $n \geq \left(\frac{2.58 \cdot 0.05}{0.005}\right)^2$, soit $n \geq 666$.*

Si on utilise la demi-erreur (comme la plupart l'ont fait, ce qui n'est inexact), on a $n = 97$ et $n = 167$.

3.3 Loi de Student

Soit T un v.a de Student à ν degrés de liberté. Quelle sont les valeurs critiques de t pour lesquelles $P(T > t) = 0.05$ pour $\nu = 16$, $\nu = 27$, $\nu = 200$? *Réponse. ν est bien évidemment le nombre de degrés de liberté.*

- *$\nu = 16$: $t = 1.75$;*
- *$\nu = 27$: $t = 1.70$;*
- *$\nu = 200$: $t = 1.65$;*

3.4 Sphères

On procède à 10 mesures du diamètre d'une sphère. On trouve une moyenne de 4.38cm et un écart-type de 0.06cm. Donner des intervalles de confiance à 95% et à 99% pour le diamètre de la sphère. *Réponse. Facile, il suffit d'utiliser les quantiles de la loi de Student, avec :*

$$IC_{i-\alpha} = [\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}]$$

3.5 Roues

Les mesures des diamètres de n roues dentées issues d'un échantillon aléatoire, fabriquées pendant une journéee par une centaine de machines, donnent un diamètre moyen de 0.824cm pour un écart-type empirique de 0.042cm.

1. si $n = 2000$, déterminer les intervalles de confiance à 95% et à 99% du diamètre moyen des roues ;
2. faire ce calcul de deux manières différentes lorsque $n = 201$;
3. faire la même chose lorsque $n = 20$.

Idem que précédemment. Lorsqu'on parle de deux manière différentes, on signifie que l'on utilise d'une part les quantiles lus dans la table de Student et d'autre part ceux dans la table de la loi normale centrée réduite.

3.6 Poids d'un groupe de sprinteurs

On a pesé 10 athlètes courant le 100m capables de couvrir cette distance en moins de 10 secondes. Nous avons les poids suivant :

$$75.9, 75.0, 75.5, 75.6, 76.1, 76.5, 77.0, 75.2, 76.0, 76.7$$

En admettant que ces résultats sont issus d'une population infinie distribuée selon une loi normale de moyenne μ et de variance σ^2.

1. Construire un intervalle de confiance à 95% pour μ si σ^2 est supposé connu et $\sigma^2 = 0.25$.
2. Construire un intervalle de confiance à 95% pour μ lorsque σ^2 est inconnu.
3. Construire un intervalle de confiance à 95% pour σ.

Utilisation directe des formules du cours. Seule nouveauté par rapport aux autres exercices : intervalle de confiance pour un écart-type.

$$IC_{1_\alpha}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right],$$

et donc

$$IC_{1_\alpha}(\sigma) = \left[\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}}\right].$$

Dans le cas présent on a donc

$$IC_{1_\alpha}(\sigma) = \left[\sqrt{\frac{9 \cdot 0.42}{19.02}}; \sqrt{\frac{9 \cdot 0.42}{2.70}}\right].$$

3.7 Ampoules

On étudie la production d'une entreprise spécialisée dans la fabrication d'ampoules. On extrait de la production totale un échantillon de 40 ampoules et on trouve 3 ampoules défectueuses.

1. Construire un intervalle de confiance à 95% pour la proportion d'ampoules défectueuses.
2. Construire un intervalle de confiance à 95% dans le cas où l'échantillon est de taille 400 et qu'on y découvre 30 ampoules défectueuses.

Réponse. Utiliser les formules vues en cours. On constate que lorsque le nombre étudié d'ampoules devient grand, les deux approximations se rapprochent. L'approximation simple (qui ne fait pas intervenir de racine carrée) est de mauvaise qualité lorsque peu d'ampoules sont analysées et donnent un intervalle de confiance dont une des bornes est négative.

Références

[1] P. G. Hoel. *Introduction to Mathematical Statistics*. John Wiley and Sons, 1984.

[2] M. R. Spiegel. *Théorie et applications de la statistique*. Series Schaum, 1976.