Université de Rennes 1 Année 2013–2014
PRSB (Probabilités et Statistiques, L1 Bio)

Feuille d'exercices 6

Exercice 1 Encore une application du TCL. Dans un test de parapsychologie, il y a un jeu avec 5 cartes. Le sujet voit le dos d'une carte, prise au hasard parmi les 5, et doit deviner de laquelle il s'agit. Cette expérience est répétée 100 fois. Un sujet sans pouvoirs inhabituels donnera donc en moyenne 20 bonnes réponses. Combien de bonnes réponses faut-il pour pouvoir affirmer qu'il n'y a que une chance sur 100.000 d'obtenir un si bon score par pure chance ? (La réponse est qu'il faut au moins 38 bonnes réponses.)

Exercice 2 Considérons des données $(x_1, \ldots, x_n)$ dont on supposera qu'elles sont des réalisations de n variables aléatoires *i.i.d.* $(X_1, \ldots, X_n)$, qui suivent toutes une loi géométrique : $X_i \sim \mathcal{G}(p)$. Donner un estimateur du paramètre inconnu p. Cet estimateur est-il biaisé ? Consistant ?

Exercice 3 On a rencontré en cours la *variance empirique*, en tant qu'estimateur de la variance d'une variable aléatoire. Pour rappel, si $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d., on note d'abord

$$\overline{X}(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n} \quad \text{la moyenne empirique.}$$

et on définit ensuite la *variance empirique* comme la variable aléatoire

$$S^2(X_1, \ldots, X_n) \;=\; \frac{(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n} \;=\; \frac{X_1^2 + \ldots + X_n^2}{n} - \overline{X}^2$$

Le but de cet exercice est de montrer que cet estimateur a tendance à sous-estimer la variance. Pour simplifier, on va supposer que $\mathbb{E}(X_1) = \ldots = \mathbb{E}(X_n) = 0$. On notera σ^2 la "vraie" variance (inconnue) : $\mathbb{V}(X_1) = \ldots = \mathbb{V}(X_n) = \sigma^2$.

(a) Montrer que $\mathbb{E}(X_i^2) = \sigma^2$ pour $i = 1, \ldots, n$

(b) Montrer que $\mathbb{E}(X_i \cdot X_j) = 0$ (si $i \neq j$)

(c) Montrer que $\mathbb{E}(S^2) = \frac{n-1}{n} \cdot \sigma^2$. Autrement dit : en moyenne, la variance empirique sous-estime la variance par un facteur $\frac{n-1}{n}$.

(d) Inventer un meilleur estimateur pour la variance (un estimateur non-biaisé).

Exercice 4 Supposons que $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d., distribués selon une loi uniforme $X_i \sim \mathcal{U}([0, T])$, où le paramètre T est inconnu, et on veut l'estimer. En cours on a rencontré l'estimateur

$$\widehat{T}_n(X_1, \ldots, X_n) = \max(X_1, \ldots, X_n)$$

(a) Soit $k > 1$. Calculer la probabilité $\mathbb{P}(T \in [\widehat{T}_n, k \cdot \widehat{T}_n])$. (Indication : la réponse est $1 - \frac{1}{k^n}$)

(b) En déduire que $[\widehat{T}_n, \frac{1}{\sqrt[n]{\alpha}} \cdot \widehat{T}_n]$ est un intervalle de confiance de niveau $1 - \alpha$.

(c) Un exemple concret : si parmi 100 nombres tirés, le plus grand est 1,85, donner un intervalle de confiance de niveau 95% pour le paramètre T.

Exercice 5 Supposons une variable aléatoire X suit une loi normale centrée réduite : $X \sim \mathcal{N}(0, 1)$.

(a) Trouver le nombre z tel que $\mathbb{P}(-z \leqslant X \leqslant z) = 90\%$

(b) Trouver le nombre z tel que $\mathbb{P}(-z \leqslant X \leqslant z) = 95\%$

Indication : vous pouvez utiliser la table avec les quantiles de la loi $\mathcal{N}(0, 1)$.

Exercice 6 Let but de cet exercice est d'établir un intervalle de confiance pour l'estimation de la moyenne p d'une variable aléatoire de Bernoulli $\mathcal{B}(1, p)$. Plus précisement, supposons que $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d. distribuées selon une loi de Bernoulli $X_i \sim \mathcal{B}(1, p)$. Soit $\overline{X} = \frac{X_1 + \ldots + X_n}{n}$ la moyenne empirique. On va montrer que, pour des grandes valeurs de n, un intervalle de confiance de niveau au moins $1 - \alpha$ pour p est l'intervalle

$$\left[\overline{X} - \frac{z}{2\sqrt{n}} \ , \ \overline{X} + \frac{z}{2\sqrt{n}}\right] \quad \text{où } z \text{ est le quantile d'ordre } \ 1 - \tfrac{\alpha}{2} \ \text{ de la loi } \mathcal{N}(0, 1)$$

(a) Démontrer que, pour des grandes valeurs de n, on a une approximation

$$\mathbb{P}\left(-z \leqslant \frac{\sqrt{n}}{\sqrt{p(1-p)}}(\overline{X} - p) \leqslant z\right) \ \cong \ 1 - \alpha$$

(b) Il est facile à vérifier (faites-le !) que pour $p \in [0, 1]$ on a $p(1-p) \in [0, \frac{1}{4}]$. Utiliser cette information pour démontrer que, pour des grandes valeurs de n,

$$\mathbb{P}\left(-z \leqslant 2\sqrt{n} \cdot (\overline{X} - p) \leqslant z\right) \ \geqslant \ 1 - \alpha$$

(c) Déduire que

$$\left(\overline{X} - \tfrac{z}{2\sqrt{n}} \ \leqslant p \ \leqslant \ \overline{X} + \tfrac{z}{2\sqrt{n}}\right) \geqslant 1 - \alpha$$

Exercice 7 (Applications de la formule dans l'exercice précédent)

(a) Parmi 900 poissons pêchés dans un lac, on a observé 180 porteurs de parasites. Entre quelles limites situez-vous la proportion des individus parasités dans la population des poissons du lac. (Niveau de confiance exigé : 95%.)

(b) On sait que, à chaque naissance, la probabilité p d'observer un garçon est très proche de $\frac{1}{2}$ (à peu près 105 garcons pour 100 filles en moyenne en France). Pour estimer précisément cette probabilité, on recherche son intervalle de confiance pour un coefficient de sécurité de 99.99% à partir de la proportion de garçons observée sur n naissances. Quelle valeur donner à n pour avoir une estimation à 0.001 près ?