Ecole Normale Supérieure
FIMFA – Statistique 2012-2013

Feuille d'exercices n°3.

Exercice 1. Cours : deux inégalités classiques.

1. *Inégalité de Hoeffding.* Soit $Y_1, ..., Y_n$ des variables réelles indépendantes centrées, vérifiant $a_i \leq Y_i \leq b_i$, $\forall i \in [[1, n]]$, les $a_i, b_i, i \in [[1, n]]$ étant des réels donnés.
   (a) Considérons une variable aléatoire $Z \in [0, 1]$ p.s. et posons
   $$\phi(x) = \ln\left(\mathbb{E}[\exp(xZ)]\right), \quad x \in \mathbb{R}.$$
   Montrer que $\phi''(x) \leq 1/4$, $\forall x \in \mathbb{R}$. En déduire que pour tout $x \geq 0$, $\phi(x) \leq x\mathbb{E}[Z] + x^2/8$, puis,
   $$\mathbb{E}\left[\exp(xY_i)\right] \leq \exp\left(\frac{x^2(b_i - a_i)^2}{8}\right), \quad \forall i \in [[1, n]].$$
   (b) En déduire que pour tous $x, t > 0$,
   $$\mathbb{P}\left(\sum_{i=1}^n Y_i \geq t\right) \leq \exp\left(-xt + \frac{x^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right).$$
   On obtient alors l'inégalité de Hoeffding en prenant $x = 4t/(\sum_{i=1}^n (b_i - a_i)^2)$:
   $$\mathbb{P}\left(\sum_{i=1}^n Y_i \geq t\right) \leq \exp\left(-\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$
   (c) En déduire que si $X_1, ..., X_n$ sont des variables réelles i.i.d. vérifiant $a \leq X_1 \leq b$, avec $a, b \in \mathbb{R}$, $a < b$, d'espérance $\mu \in \mathbb{R}$, alors
   $$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right).$$

2. *L'inégalité de Berry-Essen* permet d'évaluer la vitesse de convergence de la fonction de répartition d'une moyenne empirique centrée réduite vers la loi normale centrée réduite. Plus précisément, soit $(X_1, ...X_n)$ un échantillon de variables i.i.d. d'espérance $\mu \in \mathbb{R}$ et de variance $\sigma^2 < \infty$. Alors,
   $$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) - \Phi(x)\right| \leq 3\frac{\mathbb{E}\left[|X_1 - \mu|^3\right]}{\sigma^3\sqrt{n}}, \quad \forall n \geq 1,$$
   où Φ désigne la fonction de répartition d'une loi normale centrée réduite. Pour une preuve, voir par exemple le livre de Durrett, *Probability : Theory and Examples*.

Exercice 2. On observe des variables i.i.d. $X_1, ..., X_n$ de densité

$$f_\theta(x) = e^{-(x-\theta)} \mathbf{1}_{[\theta,\infty[}(x),$$

$\theta \in \mathbb{R}$. On veut estimer θ par diverses méthodes.

1. A partir de la méthode des moments, fabriquer un estimateur $\hat{\theta}_n$ de θ. Préciser son biais et son risque quadratique.
2. Donner, en se basant sur une inégalité classique, un intervalle de confiance non-asymptotique de niveau $1 - \alpha$ pour θ.
3. Donner la loi limite de $\hat{\theta}_n$ et en déduire un intervalle de confiance asymptotique de même niveau pour θ.
4. Déterminer un pivot simple de θ, qu'on exprimera en fonction de $\hat{\theta}_n$. En déduire un nouvel intervalle de confiance non-asymptotique pour θ.

Comparer les différents intervalles de confiance obtenus.

Exercice 3. On considère $X_i, i \geq 1$ des variables i.i.d. de loi de Poisson de paramètre $\lambda > 0$ inconnu.

1. Construire un intervalle de confiance asymptotique de λ de niveau $1-\alpha$ en utilisant le TCL et le lemme de Slutsky.
2. Construire un intervalle de confiance asymptotique de λ de niveau $1-\alpha$ en utilisant le TCL et la méthode de la stabilisation de la variance.

Comparer les deux intervalles de confiance obtenus.

Exercice 4. On considère le modèle statistique $(\mathbb{R}^{p\times n}, \{\mathcal{N}_p(\mu, \sigma^2 \mathrm{I}_p)^{\otimes n}\}_{\mu\in\mathbb{R}^p, \sigma^2>0})$. On note $(X_1, ..., X_n)$ un échantillon associé et pour chaque $i, 1 \leq i \leq n$, $X_i = (X_{i,1}, ..., X_{i,p})$. L'objectif est d'estimer par région de confiance au niveau $1 - \alpha$ la moyenne μ.

1. En distinguant les cas où la variance σ^2 est connue ou inconnue, donner un pivot de μ.
2. En déduire une région de confiance pour μ au niveau $1 - \alpha$.
3. Donner une région de confiance sous forme de boule euclidienne dans le cas où σ^2 est connue.

Exercice 5. Application pratique. Une radio généraliste s'inquiète de sa perte d'audience. Une enquête faite auprès de 1186 Français choisis au hasard dans l'annuaire et ayant accepté de répondre montre que 239 d'entre eux ont déclaré écouter au moins de temps en temps la station. Pour assurer son avenir, elle voudrait améliorer sa connaissance des habitudes de ses plus jeunes auditeurs : parmi les sondés, 102 étaient étudiants, et 30 l'écoutaient au moins de temps en temps. La direction de la prospective va diligenter un nouveau sondage destiné uniquement aux étudiants.

1. Modéliser le problème (pour chacun des deux sondages), en précisant la population, les données et le modèle ; indiquer en particulier les paramètres d'intérêt respectifs.

2. Déduire du premier sondage une précision et un intervalle de confiance pour les paramètres d'intérêt de chacun des deux sondages.
3. Combien de personnes faut-il interroger au cours de la seconde enquête, si le degré de confiance retenu est de 95 % et la précision désirée 3 % ?
4. A l'issue du second sondage, il a été constaté 38.3 % d'auditeurs. Donner une estimation et un intervalle de confiance du paramètre faisant l'objet de l'étude (avec un degré de confiance de 95 %).
5. Peut-on affirmer que l'audience du segment étudiant a augmenté d'une enquête à l'autre ? (Les deux enquêtes sont séparées de six mois et la station a revu entre-temps sa grille de programmes.)