Statistique 3
2019-2020. Exercices d'entraînement : Approche non asymptotique (1)

Exercice 1 Distance entre la médiane et la moyenne

Soit X une variable aléatoire réelle de carré intégrable. Soit m une médiane de X.

1. En utilisant l'inégalité de Bienaymé-Chebychev, montrer que $|\mathbb{E}X - m| \leq \sqrt{2\mathsf{Var}X}$.

2. Dans cette question, on va démontrer l'inégalité plus précise : $|\mathbb{E}X - m| \leq \sqrt{\mathsf{Var}X}$.
   a) Montrer que $|\mathbb{E}X - m| \leq \mathbb{E}|X - m|$.
   b) Montrer que $\mathbb{E}|X - \mathbb{E}X| \leq \sqrt{\mathsf{Var}X}$.
   c) Conclure, en utilisant le fait que m minimise $\mathbb{E}|X - t|, t \in \mathbb{R}$.

Exercice 2 Inégalité de Chebychev-Cantelli

Soit X une variable aléatoire réelle de carré intégrable et $t > 0$.

1. Montrer que pour tout $\lambda \geq 0$,
$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq \frac{\mathsf{Var}X + \lambda^2}{(t+\lambda)^2}.$$

2. En déduire que
$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq \frac{\mathsf{Var}X}{\mathsf{Var}X + t^2}.$$

3. Quel est l'avantage de cette inégalité par-rapport à l'inégalité de Bienaymé-Chebychev ?

4. A l'aide de l'inégalité de Chebychev-Cantelli démontrée plus haut, retrouver le résultat de l'exercice précédent.

Exercice 3 Un raffinement de l'inégalité de Hoeffding: une inégalité de Bernstein (exercice donné lors de l'examen 2019)

1. Soit $X_1, \ldots, X_n$ des variables aléatoires iid de loi de Bernoulli de paramètre $p \in]0, 1[$.
   a) Réécrire l'inégalité de Hoeffding pour ces variables de Bernoulli.

b) En déduire un intervalle de confiance de niveau non asymptotique $\alpha \in]0,1[$ pour p.

c) Cet intervalle de confiance vous paraît-il raisonnable lorsque p est très proche de 0 ou de 1 ? Expliquez votre réponse.

2. Dans cette question, on démontre une inégalité plus fine que l'inégalité de Hoeffding lorsque les variables X_i ont une très petite variance.
Soit $X_1, \ldots, X_n$ des variables aléatoires iid bornées: il existe des réels $a \le b$ tels que $a \le X_1 \le b$ presque sûrement. On note μ l'espérance de X_1, V sa variance (supposée non nulle) et on définit $Y_i = X_i - \mu$ pour tout $i = 1, \ldots, n$.

a) Trouver un nombre $c > 0$ tel que $|Y_1| \le c$ presque sûrement.

b) Soit $\Psi(\lambda) = \ln \mathbb{E}[e^{\lambda Y_1}]$, pour tout $\lambda \ge 0$. Pourquoi la fonction Ψ est-elle bien définie?

c) Montrer que pour tout $\lambda \ge 0$,

$$\Psi(\lambda) \le \mathbb{E}[e^{\lambda Y_1} - 1 - \lambda Y_1].$$

d) Montrer que la fonction $u \mapsto \dfrac{e^u - 1 - u}{u^2}$ est bien définie sur $\mathbb{R}$, et qu'elle est croissante.

e) En déduire que

$$\Psi(\lambda) \le (e^{\lambda c} - 1 - \lambda c)\frac{V}{c^2}.$$

f) A l'aide de la borne de Chernoff, conclure que pour tout $t > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu) \ge t\right] \le \exp\left(-\frac{nV}{c^2} h\left(\frac{ct}{nV}\right)\right),$$

où $h(u) = (1+u)\ln(1+u) - u$, $\forall u \ge 0$.

g) On admet que pour tout $u \in \mathbb{R}$,

$$h(u) \ge \frac{u^2}{2(1+u/3)}. \tag{1}$$

En déduire que, pour tout $t > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu) \ge t\right] \le \exp\left(-\frac{t^2}{2\,(nV + ct/3)}\right).$$

h) Conclure que, pour tout $t > 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(X_i - \mu)\right| \ge t\right] \le 2\exp\left(-\frac{t^2}{2\,(nV + ct/3)}\right).$$

i) Supposons que les X_i sont des variables de Bernoulli, de paramètre $p \in]0,1[$. Déduire de la question précédente un nouvel intervalle de confiance pour p de niveau non asymptotique α, où $\alpha \in]0,1[$. Lorsque p est très proche de 0 ou de 1, expliquez pourquoi ce nouvel intervalle de confiance améliore celui de la première question.