Examen partiel de statistiques — jeudi 30 janvier 2024

Exercice 1. ∼ La loi de Zipf de paramètres $N \in \mathbb{N}^*, s > 0$ est la loi sur $\{1, \dots, N\}$ définie par

$$\mathbb{P}(X = k) = \frac{k^{-s}}{H_{N,s}},$$

où $H_{N,s} = 1 + 2^{-s} + 3^{-s} + \cdots + N^{-s}$. Son espérance est notée $\mu = \mu_{N,s}$. On observe un échantillon $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et identiquement distribuées de loi de Zipf de paramètres N, s.

1. On suppose d'abord N connu. En utilisant l'inégalité de Hoeffding, donner un intervalle de confiance de niveau $1-\alpha$ pour μ. Si l'on se donne un niveau de risque $\alpha = 1\%$, quelle est la taille d'échantillon n nécessaire pour obtenir un intervalle de confiance de longueur inférieure à 0.1 ?
2. On suppose maintenant N inconnu. Trouver un estimateur de N qui est convergent lorsque $n \to \infty$ (il y a plusieurs réponses possibles).
3. En déduire, toujours en utilisant l'inégalité de Hoeffding, un intervalle de confiance asymptotique de niveau $1-\alpha$ pour μ lorsque N est inconnu.

Exercice 2. ∼ Soit $(X_1, \dots, X_n)$ un échantillon indépendant et identiquement distribué de loi de Rayleigh $\mathscr{R}(\sigma)$ de paramètre $\sigma > 0$. Cette loi a pour densité

$$\varphi_\sigma(x) = \frac{x}{\sigma} e^{-\frac{x^2}{2\sigma^2}} \mathbf{1}_{x \geqslant 0}.$$

On cherche à tester l'hypothèse nulle $H_0 : \ \sigma^2 = 1$ contre l'hypothèse alternative $H_1 : \sigma^2 = 2$.

1. Écrire le rapport des vraisemblances de l'échantillon pour ces deux hypothèses — on le notera ϱ.
2. Montrer que la zone de rejet d'un test de rapport de vraisemblance a la forme $\{S > z\}$, où $S = \sum_{i=1}^n X_i^2$ et $z \in \mathbb{R}$.
3. On va calculer la distribution de S sous l'hypothèse nulle.
   (a) Quelle est la loi de X_i^2 sous l'hypothèse nulle ?
   (b) En déduire que S suit une loi $\Gamma(n, 1/2)$.
   (c) Construire le test de H_0 contre H_1 le plus puissant possible parmi les tests de niveau de confiance $1-\alpha$.

Exercice 3. ∼ La loi exponentielle tronquée $\mathscr{E}_M$ (avec $M > 0$) est la loi d'une variable exponentielle de paramètre 1 conditionnée à être plus petite que M.

1. (a) Montrer que la densité de $\mathscr{E}_M$ est donnée par

$$\varrho_M(x) = \frac{e^{-x}}{1 - e^{-M}} \mathbf{1}_{0 \leqslant x \leqslant M}.$$

   (b) Montrer que l'espérance de $\mathscr{E}_M$ est donnée par

$$1 - \frac{M e^{-M}}{1 - e^{-M}}.$$

   (c) Montrer que la fonction

$$\gamma(t) = 1 - \frac{t e^{-t}}{1 - e^{-t}}$$

   est un difféomorphisme [1] de $]0, \infty[$ dans $]0, 1[$.

2. On dispose d'un échantillon iid $(X_1, \dots, X_n)$ de loi $\mathscr{E}_M$, et on cherche à estimer M.
   (a) Proposer un estimateur de M par la méthode des moments.
   (b) Proposer un estimateur de M en utilisant $X_{(n)} = \max\{X_1, \dots, X_n\}$.
   (c) Lequel de ces estimateurs préféreriez-vous utiliser ?

1. Une bijection $\mathscr{C}^1$ d'inverse $\mathscr{C}^1$.

Solution du premier exercice.

Une application immédiate de l'inégalité de Hoeffding aux variables X_i, qui vivent dans l'intervalle $[1, N]$ qui est de longueur $N-1$, donne

$$\mathbb{P}(|S_n - n\mu| > t) \leqslant 2e^{-2t^2/n(N-1)^2}.$$

En choisissant

$$t = (N-1)\sqrt{\frac{n\ln(2/\alpha)}{2}},$$

on voit immédiatement que l'intervalle

$$I = \left[\bar{X}_n \pm (N-1)\sqrt{\frac{\ln(2/\alpha)}{2n}}\right] \tag{1}$$

contient μ avec probabilité au moins $1-\alpha$. Pour obtenir une longueur inférieure à 0.1 avec un niveau de risque $\alpha = .01$, il suffit de demander à ce que $2(N-1)\sqrt{\ln(200)/2n}$ soit plus petit que 0.1. Cela se produit dès que n est plus grand que

$$200(N-1)^2\ln(200) \approx 1060N^2.$$

Lorsque N n'est pas connu, il y a de nombreuses façons de l'estimer. Comme N est le maximum du support de la loi, on peut par exemple utiliser le maximum $\hat{N} = \max\{X_1, \ldots, X_n\}$ comme estimateur. Il est évident que

$$\mathbb{P}(\hat{N} \neq N) = \mathbb{P}(X_i < N)^n$$

et comme $\mathbb{P}(X_i < N) < 1$, cela tend vers zéro lorsque $n \to \infty$. L'estimateur est donc bien convergent. Maintenant, si l'on note $E(N, n)$ l'événement $\{\hat{N} \neq N\}$ et que l'on remplace N par $\hat{N}$ dans l'intervalle (1), on obtient un autre intervalle I' qui coïncide avec I lorsque $\hat{N} = N$. Ainsi,

$$\mathbb{P}(\mu \notin I') = \mathbb{P}(\hat{N} = N, \mu \notin I) + \mathbb{P}(\hat{N} \neq N, \mu \notin I') \leqslant \mathbb{P}(\mu \in I) + \mathbb{P}(\hat{N} \neq N).$$

Comme ce dernier terme tend vers zéro, on obtient $\limsup \mathbb{P}(\mu \notin I') \leqslant \alpha$ et l'intervalle

$$I = \left[\bar{X}_n \pm (\hat{N}-1)\sqrt{\frac{\ln(2/\alpha)}{2n}}\right] \tag{2}$$

est donc asymptotiquement de niveau $1-\alpha$. Ce raisonnement est valable pour n'importe quel estimateur convergent de N ; celui qui nous avons pris n'est probablement pas le meilleur.

Solution du second exercice.

Le rapport de vraisemblance est égal à

$$\varrho(x_1, \ldots, x_n) = \frac{1}{2^n}\exp\left\{\frac{1}{4}\sum_{i=1}^{n} x_i^2\right\}.$$

Un test de rapport de vraisemblance est de la forme $\{\varrho > t\}$ pour un certain t. Ici, c'est équivalent à $S > z$ avec $z = 4\ln(t2^n)$.

Calculons maintenant les lois. Un simple changement de variable montre que la densité de X_i^2 vaut $e^{-x/2}/2$ sur $[0, +\infty[$: c'est donc une loi exponentielle de paramètre $1/2$. La somme de n variables exponentielles de paramètre $1/2$ est une loi $\Gamma(n, 1/2)$ d'après ce que nous avons vu en cours. En notant $q_{n,1-\alpha}$ le quantile d'ordre $1-\alpha$ de cette loi [2], on voit donc que le test dont la région de rejet est

$$\{S > q_{n,1-\alpha}\}$$

2. Il s'agit du seul et unique q tel que $\mathbb{P}(\Gamma(n, 1/2) > q) = \alpha$.

est un test de rapport de vraisemblance de niveau de confiance $1-\alpha$. D'après le théorème de Neyman-Pearson, c'est aussi le test le plus puissant possible parmi ces tests.

Solution du troisième exercice.

Soit F la fonction de répartition. Par la formule du conditionnement,

$$\mathbb{P}(E<t|E<M)=\frac{\mathbb{P}(E<t)}{\mathbb{P}(E<M)}=\frac{1-e^{-t}}{1-e^{-M}}$$

lorsque $t<M$, et 0 sinon. La densité étant la dérivée de cette fonction, le résultat est immédiat. On en déduit tout de suite que $\mathbb{P}(X>t)=(e^{-t}-e^{-M})/(1-e^{-M})$ pour $t<M$ et 0 sinon. L'espérance est donnée par la formule

$$\mathbb{E}X=\int_0^M \mathbb{P}(X>t)dt=\int_0^M \frac{e^{-t}-e^{-M}}{1-e^{-M}}dt=1-\frac{Me^{-M}}{1-e^{-M}}.$$

Le maximum a pour loi

$$\mathbb{P}(X_{(n)}<t)=\mathbb{P}(X_1<t)^n=\left(\frac{e^{-t}-e^{-M}}{1-e^{-M}}\right)^n.$$

Pour estimer M, on veut utiliser la delta-méthode, mais il faut montrer que γ est un difféomorphisme. La dérivée de γ vaut

$$\gamma'(t)=\frac{-e^{-t}(1-e^{-t}-t)}{(1-e^{-t})^2}.$$

Il faut donc étudier le signe de $\delta(t)=1-e^{-t}-t$. En fait, exp est strictement convexe, donc elle est toujours au-dessus de ses tangentes, et $e^x>1+x$ pour tout $t\neq 0$, donc $\delta(t)<0$ pour tout $t>0$. On en déduit que $\gamma'(t)>0$ pour tout $t>0$, c'est-à-dire que γ est une fonction strictement croissante et clairement $\mathscr{C}^\infty$: c'est bien un difféormorphisme (croissant), allant de $\gamma(0)=0$ (calculer la limite de cette forme indéterminée) à $\gamma(\infty)=1$. On peut donc utiliser la méthode des moments pour estimer M : l'estimateur

$$\gamma^{-1}\left(\bar{X}_n\right)$$

est convergent, et même asymptotiquement normal. Seulement, il nécessite d'inverser γ.

A contrario, l'estimateur du maximum est également convergent. Pour les comparer, il faudrait calculer leur risque quadratique, ou bien construire des intervalles de confiance fondés sur leurs lois respectives et calculer leur longueur. Cependant, avec un peu d'intuition, on peut voir que le premier estimateur est asymptotiquement normal : la longueur de l'intervalle de confiance sera donc de l'ordre de $1/\sqrt{n}$, alors que le second estimateur est basé sur le maximum, et il est très facile de montrer que $n(M-X_{(n)})$ converge en loi vers une exponentielle. La longueur de l'intervalle de confiance sera donc de l'ordre de $1/n$, ce qui est bien meilleur.