Examen partiel de statistiques (STF8)

Vendredi 26 janvier 2024 — de 13h30 à 15h30.

EXERCICE 1. ∼ Questions de cours (6 points).

1. Énoncer le théorème de Neyman-Pearson. Donner un exemple de test de deux hypothèses simples pour lequel il n'existe pas forcément de test de rapport de vraisemblance de niveau $1-\alpha$.
2. Dans un modèle où l'on observe n réalisations de gaussiennes iid $N(\mu, \sigma^2)$, donner un intervalle de confiance non-asymptotique de niveau $1-\alpha$ pour μ (la variance σ^2 n'est pas connue).

EXERCICE 2. ∼ (7 points) Soit X une variable aléatoire de loi $\mathrm{Bin}(n, p)$. On connaît p, et on cherche à estimer n à partir d'une seule réalisation de X.

1. Proposer un estimateur sans biais de n.
2. Trouver le test optimal au sens de l'erreur totale pour l'hypothèse nulle $n = 20$ contre l'hypothèse alternative $n = 21$.
3. Écrire l'inégalité de Hoeffding appliquée à X.
4. En déduire un intervalle de confiance non-asymptotique pour n (indice : avoir du second degré).

EXERCICE 3. ∼ (8 points) Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes de loi de Pareto d'indices $M > 0, \beta > 0$: leur densité commune est

$$\varrho(x) = \mathbf{1}_{x>M} \frac{\beta M^\beta}{x^{1+\beta}}.$$

L'objectif de l'exercice est d'estimer β.

1. On suppose dans un premier temps que M est connu.
   (a) Calculer $\mathbb{P}(X_i > t)$. Quelle est la loi de $\ln(X_i/M)$?
   (b) En déduire un estimateur de β qui est convergent et asymptotiquement normal (on précisera la limite).
   (c) Trouver un intervalle de confiance asymptotique pour β de niveau $1-\alpha$, et calculer l'ordre de grandeur de sa taille.
2. On suppose maintenant que M n'est pas connu, et qu'il faut donc l'estimer aussi.
   (a) Calculer la fonction de répartition de
   $$m_n = \min\{X_1, \dots, X_n\}.$$
   Quelle est sa loi ?
   (b) Montrer que m_n est un estimateur convergent de M.
   (c) En déduire un estimateur de β qui est convergent et asymptotiquement normal.
   (d) Trouver un intervalle de confiance asymptotique pour β de niveau $1-\alpha$.

Solution du premier exercice. Comme $\mathbb{E}X = np$, X/p est un estimateur sans biais de n. La vraisemblance du modèle sous les paramètres n, p s'écrit $\binom{n}{X}p^X(1-p)^{n-X}$. Le rapport de vraisemblance vaut

$$\frac{\binom{21}{X}p^X(1-p)^{21-X}}{\binom{20}{X}p^X(1-p)^{20-X}} = (1-p)\frac{21!(20-X)!}{20!(21-X)!} = (1-p)\frac{21}{21-X}$$

Cette quantité est plus grande que 1 si et seulement si $21 - 21p > 21 - X$ soit $X > 21p$, ce qui donne le test optimal de $n = 20$ contre $n = 21$ au sens de l'erreur totale.

Construisons maintenant un intervalle de confiance à l'aide de l'inégalité de Hoeffding. La variable aléatoire X/p est somme de n variables indépendantes qui valent soit 0 soit $1/p$, donc l'inégalité s'écrit $P(|X/p - n| > t) < 2e^{-2t^2/np^{-2}}$. En prenant $t = \sqrt{n\ln(2/\alpha)/2p^2} = c\sqrt{n}$, la probabilité de l'événement $|X/p - n| > t$ est donc plus petite que α, ce qui fournit dans un premier temps un intervalle de fluctuation :

$$X/p - c\sqrt{n} \leqslant n \leqslant X/p + c\sqrt{n}$$

Les deux bornes dépendent du paramètre n : il faut transformer cela en intervalle de confiance, et pour cela, effectuer un pivot. On pose $x = \sqrt{n}$. L'inégalité de droite équivaut à $x^2 - cx \leqslant X/p$ soit $(x - c/2)^2 \leqslant X/p + c^2/4$, donc $|\sqrt{n} - c/2| \leqslant \sqrt{X/p + c^2/4}$ et in fine

$$n \leqslant \left(\sqrt{\frac{X}{p} + \frac{\ln(2/\alpha)}{8p^2}} + \frac{\sqrt{\ln(2/\alpha)/2}}{2p}\right)^2$$

Par le même raisonnement à gauche, on obtient aussi

$$n \leqslant \left(\sqrt{\frac{X}{p} + \frac{\ln(2/\alpha)}{8p^2}} - \frac{\sqrt{\ln(2/\alpha)/2}}{2p}\right)^2$$

et donc on aboutit à l'IC non-asymptotique

$$\left[\left(\sqrt{\frac{X}{p} - \frac{\ln(2/\alpha)}{8p^2}} + \frac{\sqrt{\ln(2/\alpha)/2}}{2p}\right)^2 \;;\; \left(\sqrt{\frac{X}{p} + \frac{\ln(2/\alpha)}{8p^2}} + \frac{\sqrt{\ln(2/\alpha)/2}}{2p}\right)^2\right]$$

Solution du second exercice. Le problème est complètement routinier dès qu'on a compris que les $\ln(X/M)$ sont des lois exponentielles de paramètre β. Les questions visaient seulement à vous guider.

En notant $Y = \ln(X/M)$ on voit que $P(Y > t) = P(X > e^t M)$. Comme $X > M$, si $t \leq 0$ cette quantité est nulle, et sinon, elle vaut

$$M^\beta \int_{Me^t} \beta x^{-\beta-1} dx = -M^\beta [x^{-\beta}]_{Me^t}^\infty = e^{-\beta t}. \tag{1}$$

Les Y sont donc bien des exponentielles de paramètre β, dont on rappelle que la moyenne est $1/\beta$ et la variance est $1/\beta$ aussi. Par la LGN, la moyenne empirique $\bar{Y_n}$ des $Y_i = \ln(X_i/M)$ converge donc en probabilité et ps vers $1/\beta$. Par continuité, son inverse

$$\hat{\beta_n} = \frac{n}{\sum_{i=1}^n \ln(X_i/M)} \tag{2}$$

converge vers β : voilà notre estimateur convergent. De plus, le TCL appliqué aux exponentielles iid Y_i dit que $\sqrt{n}(\bar{Y}_n - \beta^{-1}) \to N(0, \beta^{-1})$. Le théorème de la delta-méthode appliquée à $x \mapsto 1/x$ (qui est bien continûment dérivable en $1/\beta$, de dérivée $-\beta^2$ non nulle) entraîne que

$$\sqrt{n}(\hat{\beta}_n - \beta) \to N(0, \beta^2).$$

Construisons maintenant un intervalle de confiance. Le résultat précédent peut s'écrire $\sqrt{n}(\hat{\beta}_n/\beta - 1) \to N(0,1)$. En notant z le quantile symétrique d'ordre $1-\alpha$ de $N(0,1)$, on a donc

$$P(-z \leqslant \sqrt{n}(\hat{\beta}_n/\beta - 1) \leqslant z) \to 1 - \alpha.$$

On pivote, et l'on obtient

$$P\left(\beta \in \left[\frac{\hat{\beta}_n}{1+\frac{z}{\sqrt{n}}} \;\; ; \;\; \frac{\hat{\beta}_n}{1-\frac{z}{\sqrt{n}}}\right]\right) \to 1-\alpha. \tag{3}$$

Lorsque $n \to \infty$, la longueur de cet IC est d'ordre $\hat{\beta}_n 2z/\sqrt{n}$ (faire un développement limité).

Si maintenant M n'est pas connu, on doit l'estimer, et comme dans les problèmes similaires déjà vus en cours [1], le minimum m_n est naturellement le bon candidat. Comme les X_i sont iid, $P(m_n > t) = P(X > t)^n = (M^\beta \int_t^\infty \beta x^{-\beta-1} dx)^n = (M^\beta [x^{-\beta}]_t^\infty)^n = (M/t)^{\beta n}$: la densité de m_n est donc $M^{\beta n}(\beta n) x^{-\beta n - 1}$ si $t > M$, zéro sinon : on reconnaît, (directement dans l'énoncé du problème) une loi de Pareto de paramètres $M, \beta n$. C'est une propriété de stabilité intéressante : le minimum de lois de Pareto est encore une loi de Pareto. Cela n'a rien d'étonnant, car les lois de Pareto sont des exponentielles de lois exponentielles (cf la première partie), et que les lois exponentielles sont elles-mêmes stables par le minimum.

Le fait que m_n converge vers M est facile : pour tout $\varepsilon > 0$ on a $P(m_n > M + \varepsilon) = (M/(M+\varepsilon))^{\beta n} \to 0$, donc $m_n \to M$ en probabilité. Dès lors, pour estimer β quand M n'est pas connu, je ne vois pas d'idée plus simple que de remplacer M par son estimateur m_n dans l'équation (2) : on obtient

$$\tilde{\beta}_n = \frac{\sum \ln(X_i/m_n)}{n} = \hat{\beta}_n - \ln(m_n/M).$$

Le second terme tend vers 0 en probabilité d'après la question précédente et la continuité de ln en 1, donc $\tilde{\beta}_n$ est bien convergent. Pour montrer qu'il est asymptotiquement normal, le plus simple est d'écrire

$$\sqrt{n}\left(\frac{\sum \ln(X_i/m_n)}{n} - \beta^{-1}\right) = \sqrt{n}\left(\frac{\sum \ln(X_i/M)}{n} - \beta^{-1}\right) - \sqrt{n}\ln(m_n/M).$$

Le premier terme est asymptotiquement normal (cf première partie) : il faut donc montrer que le second terme ci-dessus tend vers 0 en probabilité et appliquer le lemme de Slutsky !

Attention, ce n'est pas immédiat : ce n'est pas parce que $\ln(m_n/M)$ tend vers zéro que c'est le cas de $\sqrt{n}\ln(m_n/M)$. Il faut faire quelque chose de plus fin. Allons-y : nous avons vu précédemment que $m_n \sim \text{Pareto}(M, \beta n)$, donc par la partie I, $\ln(m_n/M) \sim \mathscr{E}(\beta n)$. En particulier, pour tout $\varepsilon > 0$,

$$P(\sqrt{n}\ln(m_n/M) > \varepsilon) = e^{-\beta n(\varepsilon/\sqrt{n})} = e^{-\beta\varepsilon\sqrt{n}} \to 0.$$

L'application du théorème de Slutsky est donc possible et conclut toute cette aventure. L'intervalle de confiance demandé est le même que (3), mais avec $\tilde{\beta}_n$ à la place de $\hat{\beta}_n$.

1. Estimation de θ dans un modèle uniforme sur $[0, \theta]$ par exemple.