Examen de statistiques — lundi 3 mars 2025

Exercice 1. ∼ On dispose de variables réelles à expliquer $y_1, \dots, y_n$, que l'on suppose être des fonctions de $x_1, \dots, x_n$ (des nombres réels distincts). On propose de faire une régression linéaire utilisant comme variables explicatives

$$X_i = (1, \cos(x_i), \sin(x_i), \cos(2x_i), \sin(2x_i), \dots, \cos(Nx_i), \sin(Nx_i))$$

où N est un nombre entier plus grand que 1. On supposera qu'il existe un $\theta \in \mathbb{R}^{1+2N}$ tel que pour chaque i, $y_i = X_i\theta + \varepsilon_i$, où les ε_i sont iid de loi $N(0, \sigma^2)$.

1. Rappeler la formule donnant l'estimateur des moindres carrés ordinaires de θ, ainsi que l'estimateur de σ^2, et rappeler leur loi jointe.
2. Proposer un test de l'hypothèse selon laquelle y_i est une fonction symétrique de x_i.
3. (Optionnel - bonus seulement) En pratique, comment choisiriez-vous N ?

Exercice 2. ∼ La loi de Benini de paramètre $\beta > 0$ a pour densité

$$\varrho(x) = \frac{2\beta \ln(x)}{x} e^{-\beta \ln(x)^2} \mathbf{1}_{x \geqslant 1}.$$

On dispose de n observations iid $X_1, X_2, \dots, X_n$ de loi de Benini de paramètre inconnu β, et on cherche à estimer β.

1. Montrer que l'estimateur du maximum de vraisemblance de β est
$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \ln(X_i)^2}.$$
2. (a) Montrer que si X est une variable aléatoire de Benini de paramètre β, alors $\ln(X)^2 \sim \mathrm{Exp}(\beta)$.
   (b) Calculer $\mathbb{E}[\hat{\beta}]$.
3. Calculer l'information de Fisher de la loi de Benini de paramètre β.
4. Expliquer pourquoi $\hat{\beta}$ est un estimateur asymptotiquement normal et donner sa variance asymptotique.

Exercice 3. ∼ On observe un vecteur gaussien aléatoire X de loi $N(0, \sigma^2 I_n)$, où I_n est la matrice identité de taille n. On cherche à tester l'hypothèse nulle $H_0 : \sigma = 1$ contre l'hypothèse alternative $H_1 : \sigma = \sigma_0$, où $\sigma_0 > 1$ est fixé.

1. Rappeler la densité de $N(0, \sigma^2 I_n)$.
2. Donner la région de rejet du test optimal (au sens de l'erreur totale) de H_0 contre H_1.
3. On note $\alpha(n)$ l'erreur de première espèce de ce test, et $\beta(n)$ la puissance. Montrer que $\alpha(n) \to 0$ et $\beta(n) \to 1$ lorsque $n \to \infty$ (indice : il faut faire une étude de fonction).
4. On fixe un niveau de risque, par exemple $\alpha = 1\%$. On autorise maintenant σ_0 à varier avec n : à quelle vitesse peut-on faire tendre σ_0 vers 1 pour que le test ci-dessus conserve un niveau de risque de α ?

Exercice 4. ∼ Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires iid, distribuées selon une même loi gaussienne $N(\mu, \sigma^2)$: on va chercher à estimer le *ratio de Sharpe*[1] de ces variables aléatoires, défini par $\zeta = \mu/\sigma$. Soient $\hat{\mu}$ et $\hat{\sigma}^2$ les estimateurs usuels de μ et σ^2 :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

On propose d'estimer ζ par $\hat{\zeta} = \hat{\mu}/\hat{\sigma}$.

1. Dans cette question, on calcule le biais de $\hat{\zeta}$.
   (a) Montrer que $\hat{\zeta} = \mu/\hat{\sigma} + C$, où C est une variable aléatoire centrée.
   (b) Rappeler la loi de $\hat{\sigma}^2$.
   (c) Montrer que
   $$\mathbb{E}[\hat{\zeta}] = \zeta \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sqrt{\frac{n-1}{2}}$$
   et vérifier que $\mathbb{E}[\hat{\zeta}]/\zeta = 1 + O(1/n)$ lorsque $n \to \infty$.
2. On va montrer que $\hat{\zeta}$ est asymptotiquement normal.
   (a) Montrer que $\sqrt{n}(\hat{\mu} - \mu)$ converge en loi vers $N(0, \sigma^2)$.
   (b) Montrer que $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$ converge en loi vers $N(0, 2\sigma^4)$.
   (c) En déduire que le vecteur aléatoire $\sqrt{n}\begin{bmatrix}\hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2\end{bmatrix}$ converge en loi vers
   $$N\left(0, \begin{bmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{bmatrix}\right).$$
   (d) Après avoir soigneusement énoncé le théorème de la delta-méthode, montrer la convergence en loi suivante :
   $$\sqrt{n}(\hat{\zeta} - \zeta) \xrightarrow[n\to\infty]{\text{loi}} N\left(0, 1 + \frac{\zeta^2}{2}\right). \tag{1}$$
3. À l'aide de (1), construire
   (a) Un intervalle de confiance asymptotique pour ζ.
   (b) Un test asymptotique de l'hypothèse $\zeta = 1$.

1. En finance, le ratio de Sharpe d'un portefeuille est un indicateur de la rentabilité de ce portefeuille, ajustée par le niveau de risque encouru.

Solution de l'exercice 1.

La première question est contenue dans le cours. Pour la seconde, il suffit de remarquer qu'une fonction de type

$$\theta_0 + \sum_{k=1}^{N} \theta_{2k-1}\cos(kx) + \theta_{2k}\sin(kx)$$

est symétrique si et seulement si la « partie sinus » est nulle, donc si $\theta_{2k} = 0$ pour tout $k > 0$. On peut donc tester l'hypothèse $H_0 : \theta_2 = \theta_4 = \cdots = \theta_{2N} = 0$. Il s'agit du test de N contraintes linéaires sur les coefficients : un test de Wald fait l'affaire. Formellement, on teste si $C\theta = 0$ où C est la matrice $N \times (1 + 2N)$ dont les lignes sont $(0, 0, 1, 0, 0, \ldots, 0)$, $(0, 0, 0, 0, 1, 0, \ldots, 0)$, etc. La statistique de test est alors

$$w = \frac{\langle C\hat{\theta}, [C(X^\top X)^{-1}C^\top]^{-1}C\hat{\theta}\rangle}{N\hat{\sigma}^2}$$

et sa loi est $\mathscr{F}_{N,n-d}$. L'hypothèse sera rejetée lorsque $w > f$, où f est le quantile $1-\alpha$ de la loi $\mathscr{F}_{N,n-d}$.

Solution de l'exercice 2.

La log-vraisemblance est donnée par

$$\ell(\beta) = \sum_{i=1}^{n} \ln(2\beta) + \ln(\ln(X_i)) - \ln(X_i) - \beta\ln(X_i)^2.$$

La dérivée vaut $\ell'(\beta) = \frac{n}{\beta} - \sum_{i=1}^{n}\ln(X_i/x_0)^2$ et ne s'annule donc que pour $\hat{\beta} = n/\sum_{i=1}^{n}\ln(X_i)^2$. Il faut ensuite vérifier qu'il s'agit bien d'un maximum. Clairement, $\ell(\beta)$ est continue, elle tend vers $-\infty$ lorsque $\beta \to 0$ et vers $-\infty$ lorsque $\beta \to \infty$. Elle possède donc un maximum global, et comme elle n'a qu'un seul point critique, elle n'a qu'un seul maximum global.

Pour la loi de $\ln(X)^2$, on se donne une fonction test f et on calcule $\mathbb{E}[f(X)]$: le changement de variable $u = \ln(x)^2$ donne $du = 2\ln(x)dx/x$, et donc

$$\int_1^\infty \frac{2\beta\ln(x)}{x} e^{-\beta\ln(x)^2} f(\ln(x)^2)dx = \int_0^\infty \beta u e^{-\beta u} f(u)du.$$

On reconnaît une loi exponentielle de paramètre β. Pour en déduire l'espérance de $\hat{\beta}$, on commence par remarquer que

$$\frac{n}{\hat{\beta}} = \sum_{i=1}^{n} \ln(X_i)^2$$

a la loi de la somme de n variables iid $\mathrm{Exp}(\beta)$, et donc c'est une $\Gamma(n, \beta)$, de densité

$$\frac{\beta^n}{\Gamma(n)} x^{n-1} e^{-\beta x} \mathbf{1}_{x>0}.$$

Ainsi,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= n\frac{\beta^n}{\Gamma(n)}\int_0^\infty \frac{1}{x}e^{-\beta x}x^{n-1}dx \\
&= n\frac{\beta^n}{\Gamma(n)}\int_0^\infty e^{-\beta x}x^{(n-1)-1}dx \\
&= n\frac{\beta^n}{\Gamma(n)}\frac{\Gamma(n-1)}{\beta^{n-1}} \\
&= \frac{n}{n-1}\beta.
\end{aligned}$$

Pour l'information de Fisher : le score est $(1/\beta - \ln(X)^2)$, donc l'information de Fisher est donnée par

$$I(\beta) = \mathrm{Var}(\ln(X)^2) = \frac{1}{\beta^2}.$$

L'EMV est asymptotiquement normal dans les modèles exponentiels (c'est du cours). Sa variance asymptotique est donc $1/I(\beta) = \beta^2$.

Solution de l'exercice 3.

La densité de $N(0,\sigma^2)$ est $e^{-|x|^2/2\sigma^2}/(2\pi\sigma^2)^{n/2}$. Le test optimal au sens de l'erreur totale de $\sigma = 1$ contre $\sigma = \sigma_0$ est alors

$$\frac{e^{-|X|^2/2\sigma_0^2+|X|^2/2}}{\sigma_0^n} > 1.$$

En réorganisant les termes, la région de rejet devient

$$\frac{|X|^2}{n} > \frac{\ln(1/\sigma_0^2)}{1/\sigma_0^2 - 1}.$$

Posons $h(t) = \ln(t)/(t-1)$. Il est bien connu que $\ln(t) < t-1$ (avec égalité si et seulement si $t = 1$). Si $\sigma_0 > 1$, alors $t = 1/\sigma_0^2$ est plus petit que 1, donc $t-1$ est négatif et donc $h(t) > 1$. Or, sous l'hypothèse nulle, $|X|^2$ suit une loi du χ^2 à n degrés de liberté (même chose que la somme de n carrés de gaussiennes iid). Il est donc clair que $|X|^2/n \to 1$ presque sûrement (par la LGN). Par conséquent, lorsque $n \to \infty$, sous l'hypothèse nulle, le test ne sera jamais rejeté, et $\alpha(n) \to 0$. Un raisonnement similaire montre le résultat pour la puissance.

Pour la dernière question, voici une façon de faire : $\sqrt{n}(|X|^2/n - 1) \to N(0,1)$ sous H0 d'après le TCL. On pose $t_n = 1/\sigma_0(n)^2$. La région de rejet du test est de la forme

$$\sqrt{n}(|X|^2/n - 1) > \sqrt{n}(h(t_n) - 1).$$

Pour qu'elle soit asymptotiquement de niveau de confiance $1-\alpha$, il faut s'assurer que $\sqrt{n}(h(t_n) - 1)$ converge vers $q_{1-\alpha}$ (quantile non symétrique). Posons $t_n = 1 - \delta_n$ (car $t_n < 1$). Alors, comme $\ln(1-\delta_n) = \delta_n + \delta_n^2/2 + o(\delta_n^2)$, on a

$$h(t_n) = \frac{\delta_n + \delta_n^2/2 + o(\delta_n^2)}{\delta_n} = 1 + \delta_n/2 + o(\delta_n).$$

Par conséquent, $\sqrt{n}(h(t_n) - 1) = \sqrt{n}\delta_n/2 + o(\sqrt{n}\delta_n)$. On doit donc choisir δ_n de sorte que $\sqrt{n}\delta_n/2$ converge vers $q_{1-\alpha}$ c'est-à-dire $\delta_n = 2q_{1-\alpha}/\sqrt{n}$ (qui tend bien vers 0). En remontant jusqu'à σ_0 qui est égal à $\sqrt{1/t_n}$, cela nous ramène à

$$\sigma_0 = \frac{1}{\sqrt{1 - 2q_{1-\alpha}/\sqrt{n}}} \approx 1 - \frac{q_{1-\alpha}}{\sqrt{n}}.$$

Par conséquent, il reste toujours possible de distinguer deux hypothèses très proches l'une de l'autre ($\sigma = 1$ contre $\sigma = 1 + O(1/\sqrt{n})$) avec un niveau de risque fixé $1 - \alpha$.

Solution de l'exercice 4.

On peut écrire $\hat{\zeta}$ comme $\mu/\hat{\sigma} + (\hat{\mu} - \mu)/\hat{\sigma}$. Comme les lois gaussiennes sont symétriques, le second terme l'est, et donc son espérance est nulle. On sait que $(n-1)\hat{\sigma}^2/\sigma^2$ suit une loi du χ^2 à $n-1$ degrés de liberté. On peut donc écrire

$$\mathbb{E}[1/\hat{\sigma}] = \frac{\sqrt{n-1}}{\sigma} \frac{(1/2)^{(n-1)/2}}{\Gamma((n-1)/2)} \int_0^\infty e^{-t/2} t^{(n-1)/2-1} t^{-1/2} dt \tag{2}$$

$$= \frac{\sqrt{n-1}}{\sigma} \frac{(1/2)^{(n-1)/2}}{\Gamma((n-1)/2)} \int_0^\infty e^{-t/2} t^{n/2-1-1} dt \tag{3}$$

$$= \frac{\sqrt{n-1}}{\sigma} \frac{(1/2)^{(n-1)/2}}{\Gamma((n-1)/2)} \frac{\Gamma(n/2-1)}{(1/2)^{n/2-1}} \tag{4}$$

$$= \frac{\sqrt{n-1}}{\sigma\sqrt{2}} \frac{\Gamma((n-1)/2)}{\Gamma(n/2-1)} \tag{5}$$

On en déduit le résultat demandé. Pour l'asymptotique, on utilise la formule de Stirling, $\Gamma(t+1) \sim t^t e^{-t}\sqrt{2\pi t}$, pour obtenir les l'équivalent suivant de $\mathbb{E}[\hat{\zeta}/\zeta]$:

$$\sqrt{(n-1)/2}\frac{((n-4)/2)^{(n-4)/2} e^{-(n-4)/2} \sqrt{(n-4)/2}}{((n-3)/2)^{(n-3)/2} e^{-(n-3)/2} \sqrt{n-3}}$$

Après réorganisation, tout ceci vaut

$$\frac{(n-1)\sqrt{(n-4)}}{\sqrt{2(n-3)}} e^{1/2} \left(1 - \frac{1}{n-3}\right)^{(n-4)/2} ((n-4)/2)^{-1/2}$$

Le terme entre parenthèses tend vers $e^{-1/2}$ et vient annuler le $e^{1/2}$. Il reste donc

$$\frac{\sqrt{(n-4)(n-1)}}{\sqrt{(n-3)(n-4)}} = \sqrt{\frac{n-1}{n-3}} = \sqrt{1 + \frac{2}{n-3}} = 1 + O(1/n).$$

On montre maintenant que $\hat{\zeta}$ est asymptotiquement normal. La convergence $\sqrt{n}(\hat{\mu} - \mu) \to N(0, \sigma^2)$ est une simple application du TCL. De même, comme $(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi_2(n-1)$ et que $\chi_2(n-1)$ est aussi la loi de la somme de $n-1$ carrés de variables gaussiennes standard iid, on en déduit immédiatement que $\hat{\sigma}^2$ a la même loi que la moyenne de $n-1$ carrés de gaussiennes iid $N(0, \sigma^2)$. Le TCL s'applique, et montre que $\sqrt{n-1}(\hat{\sigma}^2 - \sigma^2)$ converge vers $N(0, v)$ avec $v = \mathrm{Var}(N(0, \sigma^2)^2) = 3\sigma^4 - \sigma^4 = 2\sigma^4$. On peut remplacer le $n-1$ par n grâce au lemme de Slutsky. Enfin, il est bien connu que $\hat{\mu}$ et $\hat{\sigma}^2$ sont indépendants, donc les deux convergences ci-dessus ont lieu ensemble, et les deux limites sont indépendantes — d'où la forme de la matrice de covariance demandée. On peut maintenant appliquer la delta-méthode avec la fonction $f(x, y) = x/\sqrt{y}$ dont le gradient est donné par

$$\nabla f(x, y) = \begin{bmatrix} 1/\sqrt{y} \\ -x/2y^{3/2} \end{bmatrix}.$$

En particulier,

$$\nabla f(\mu, \sigma^2) = \begin{bmatrix} 1/\sigma \\ -\mu/2\sigma^3 \end{bmatrix}.$$

La delta-méthode dit que $\sqrt{n}(f(\hat{\mu}, \hat{\sigma}^2) - f(\mu, \sigma^2))$ converge en loi vers $N(0, \nabla f(\mu, \sigma^2)^T \Sigma \nabla f(\mu, \sigma^2))$. Le terme de variance est en fait très facile à calculer car Σ est diagonale : il vaut

$$\frac{\sigma^2}{\sigma^2} + \frac{\mu^2}{2\sigma^4} = 1 + \frac{\zeta^2}{2}.$$

Pour construire un intervalle de confiance asymptotique, il ne faut pas oublier que ζ est inconnu dans la variance limite. En utilisant le lemme de Slutsky, on voit que

$$\frac{\sqrt{n}(\hat{\zeta} - \zeta)}{\sqrt{1 + \hat{\zeta}^2/2}} \xrightarrow[n\to\infty]{\text{loi}} N(0, 1).$$

Si $z_{1-\alpha}$ est le quantile symétrique de la gaussienne ($\mathbb{P}(|N(0, 1)| \leqslant z_{1-\alpha}) = 1 - \alpha$), on a donc

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\zeta} - \zeta)}{\sqrt{1 + \hat{\zeta}^2/2}}\right| \leqslant z_{1-\alpha}\right) \to 1 - \alpha.$$

Par conséquent, l'intervalle de confiance obtenu est

$$I = \left[\hat{\zeta} \pm \frac{z_{1-\alpha}\sqrt{1 + \hat{\zeta}^2/2}}{\sqrt{n}}\right].$$

On peut utiliser cet intervalle de confiance pour tester si $\zeta = 1$: la région de rejet est l'événement « 1 n'est pas dans I ».