Examen de statistiques (STF8)

Simon Coste

Vendredi 1er mars 2024 — de 13h30 à 16h30.

Exercice 1. ∼ Questions de cours.

1. Énoncer et démontrer la décomposition biais-variance d'un estimateur.
2. Dans un modèle exponentiel en dimension 1 dont la densité par rapport à la mesure de Lebesgue est $p_\theta(x) = e^{\theta T(x)}/Z(\theta)$, calculer la dérivée de $\ln Z(\theta)$.

Exercice 2. ∼ On dispose de deux échantillons : d'une part, $x_1, \dots, x_n$ sont des réalisations de variables aléatoires iid X_i, et d'autre part $y_1, \dots, y_m$ sont des réalisations de variables iid Y_i. Les tailles n et m des deux échantillons ne sont pas forcément identiques et les lois des X_i, Y_j ne sont pas forcément les mêmes ; cependant, on suppose que les X_i et les Y_i sont toutes à valeurs dans $[-1, 1]$, et on note μ, ν leurs moyennes respectives.

1. Proposer un test T non asymptotique avec niveau de confiance $1 - \alpha$ de l'hypothèse nulle $\mu = \nu$.
2. L'hypothèse alternative est $\mu \neq \nu$. Rappeler la définition de la puissance d'un test, et donner une borne sur la puissance du test ci-dessus.

Exercice 3. ∼ On dispose de n variables explicatives $x_1, \dots, x_n \in \mathbb{R}^d$ (vecteurs lignes, la constante est incluse), et n variables à expliquer $y_1, \dots, y_n$. La particularité du problème vient du fait que les y_i sont supposées à valeurs dans $\mathbb{N}$: par exemple, on peut considérer que les x_i représentent des indicateurs économiques concernant un individu, et y_i est le nombre de voitures que possède cet individu.

1. Est-il pertinent de faire une régression linéaire des x_i sur les y_i ?
2. On suppose que les y_i sont des réalisations de variables aléatoires Y_i indépendantes, de lois respectives Poisson$(e^{x_i \lambda})$ avec $\lambda \in \mathbb{R}^d$ (vecteur colonne) un même paramètre pour tous les y_i. Écrire la vraisemblance du modèle.
3. Justifier l'existence et l'unicité de l'estimateur du maximum de vraisemblance de λ dans ce modèle.
4. Écrire l'équation vérifiée par cet estimateur.

Exercice 4. ∼ Un naturaliste capture des papillons dans une forêt lointaine. Il effectue deux collectes de même durée. Soit S le nombre total d'espèces distinctes dans cette forêt ; on note X_k^1 (respectivement X_k^2) le nombre de spécimens de l'espèce k qui ont été capturées pendant la première collecte (respectivement, la seconde). On fait l'hypothèse selon laquelle les X_k^i sont des variables aléatoires indépendantes avec $X_k^i \sim$ Poisson(θ_k) où θ_k est un paramètre positif qui représente la rareté de l'espèce k.

1. On pose N_r le nombre d'espèces dont exactement r spécimens ont été capturés **pendant la première collecte**. Par exemple, N_3 est le nombre d'espèces dont on a capturé 3 spécimens et N_0 est le nombre d'espèces qu'on n'a pas capturées.
   (a) Calculer l'espérance et la variance des N_r.
   (b) Quelle est la probabilité qu'aucun spécimen de l'espèce k n'ait été capturé pendant la première collecte, mais qu'au moins un ait été capturé pendant la seconde ?
2. En utilisant le développement $e^x - 1 = \sum_{t=1}^{\infty} x^t/t!$, montrer que le nombre d'espèces N qui n'ont pas été capturées pendant la première collecte mais qui ont été capturées pendant la seconde collecte vérifie
$$\mathbb{E}[N] = \sum_{t=0}^{\infty}(-1)^t \sum_{k=1}^{S} e^{-\theta_k} \frac{\theta_k^{t+1}}{(t+1)!}.$$
3. En déduire que
$$\hat{N} := N_1 - N_2 + N_3 - N_4 + N_5 - \cdots$$
est un estimateur sans biais de N *même lorsque* S *n'est pas connu.*
4. Expliquer pourquoi ce résultat est étonnant. Quelles pourraient être les limitations de cet estimateur ?

EXERCICE 5. $\sim$ Soient $x_1, \ldots, x_n$ des variables aléatoires indépendantes gaussiennes dans $\mathbb{R}^d$, de lois respectives $N(\mu_i, \sigma_i^2 I_d)$, où $\mu_i \in \mathbb{R}^d$ et $\sigma_i^2 > 0$ sont inconnus. On note $H_{0,i}$ l'hypothèse « $\mu_i = 0$ ».

1. Construire un test de $H_{0,i}$ de niveau de confiance $1 - \alpha$. On le notera $T_\alpha^{(i)}$.
2. On effectue chacun des tests $T_\alpha^{(i)}$, indépendamment les uns des autres. Si tous les μ_i sont simultanément nuls, quelle est la probabilité qu'aucune des hypothèses $H_{0,i}$ ne soit rejetée ? Quelle est la limite de cette probabilité lorsque $n \to \infty$?
3. Dorénavant, on notera H_0 l'hypothèse selon laquelle tous les μ_i sont nuls. Quel niveau de confiance $1 - \delta$ faudrait-il choisir pour que, sous H_0, la probabilité de l'événement « aucun des tests $T_\delta^{(i)}$ ne rejette $H_{0,i}$ » soit égale à $1 - \alpha$?
4. On s'intéresse maintenant à la propriété suivante : « sous H_0, la probabilité de l'événement "plus de $k\%$ des tests $T_\delta^{(i)}$ ont rejetés $H_{0,i}$" est inférieure à α ».
   (a) Sous H_0, quelle est la loi du nombre d'hypothèses $H_{0,i}$ rejetées par $T_\delta^{(i)}$?
   (b) Trouver un δ (dépendant de k, α, n) tel que la famille de tests $T_\delta^{(i)}$ vérifie la propriété ci-dessus. (indice : Wassily).

Solution de l'exercice 2

1. Si $\mu = \nu$ alors $S := \bar{X}_n - \bar{Y}_n$ est une variable aléatoire centrée. De plus,

$$\bar{X}_n - \bar{Y}_n = \sum_{i=1} \frac{X_i}{n} + \sum_{i=1}^{m} \frac{-Y_i}{m}.$$

Si on note $Z_i = X_i/n$ pour $i = 1, \dots, n$ et $Z_i = -Y_i/m$ si $i = n+1, \dots, n+m$, alors on voit que $S = Z_1 + \cdots + Z_{n+m}$ est une somme de $n+m$ variables indépendantes bornées, donc l'inégalité de Hoeffding dit que

$$\mathbb{P}(|S - \mathbb{E}S| > t) \leqslant 2e^{-\frac{2t^2}{n\times(1/n)^2 + m\times(1/m)^2}} = 2e^{-\frac{2t^2}{1/n+1/m}}.$$

En prenant $t = \sqrt{(m^{-1}+n^{-1})\ln(2/\alpha)/2}$ on obtient donc $\mathbb{P}(|S| > t) \leqslant \alpha$. Comme n et m sont connus, le test dont la région de rejet est

$$\{|S| > t\}$$

est un test de niveau de confiance supérieur à $1-\alpha$.

On pouvait aussi utiliser l'inégalité de Bienaymé-Chebychev, auquel cas le t à choisir était plutôt

$$t = \sqrt{\frac{m^{-1}+n^{-1}}{\alpha}}$$

puisqu'on pouvait voir que, par indépendance, la variance de $\bar{X}_n - \bar{Y}_m$ est la somme des variances des Z_i, lesquelles sont chacune plus petites [1] que $1/n^2$ ou $1/m^2$ en fonction de i.

2. On rappelle que l'erreur de seconde espèce est

$$\beta = \sup_{P \in H_1} P(\text{ne pas rejeter})$$

et donc que la puissance est

$$1 - \beta = \inf_{P \in H_1} P(\text{rejeter})$$

En prenant par exemple des lois de Bernoulli de paramètres très proches, disons $\mu = 0.5$ et $\nu = 0.5 + \varepsilon$, on voit que la puissance est égale à l'erreur de première espèce du test. En effet, dans ce cas,

$$\text{puissance} \leqslant \mathbb{P}_{\mu=0.5,\nu=0.5+\varepsilon}(\text{rejeter}) \to \mathbb{P}_{\mu=0.5,\nu=0.5}(\text{rejeter}) = \text{erreur de 1ère espèce} \leqslant \alpha$$

lorsque $\varepsilon \to 0$ (par continuité, ou par convergence en loi).

Solution de l'exercice 3

Ce problème est simplement une variante de la régression logistique que nous avont faite en TD ; la différence est qu'à la place d'avoir des lois de Bernoulli, on a des lois de Poisson.

1. Non : les variables expliquées sont à valeurs dans $\mathbb{N}$. Ce qu'on veut prédire, c'est donc un nombre entier positif ; la prédiction donnée par une régression linéaire sera de la forme $x\theta$, c'est-à-dire un nombre réel peut-être même négatif.

1. Si X est une variable entre $-a$ et a, sa variance est plus petite que a^2 puisque $\mathrm{Var}(X) \leqslant \mathbb{E}[X^2] \leqslant a^2$.

2. Comme les variables Y_i sont indépendantes, la vraisemblance s'écrit

$$\prod_{i=1}^{n} e^{-e^{x_i\lambda}} \frac{(e^{x_i\lambda})^{y_i}}{y_i!} = \frac{e^{\sum_{i=1}^n y_i x_i \lambda}}{Z(\lambda)\prod_{i=1}^n y_i!}$$

où $Z(\lambda) = \exp(\sum_{i=1}^n e^{x_i\lambda})$ est la fonction de partition. En écrivant le modèle comme ceci, on constante clairement qu'il s'agit d'un modèle exponentiel. Pour anticiper les questions suivantes, on pourrait se demander s'il est identifiable. On sait déjà que les lois de Poisson sont identifiables ; il faudrait donc vérifier que si $e^{x_i\lambda} = e^{x_i\mu}$ pour tout i, alors $\lambda = \mu$. Si $e^{x_i\lambda} = e^{x_i\mu}$ pour tout i, alors $x_i(\lambda - \mu) = 0$ pour tout i, autrement dit $\lambda - \mu$ est dans l'intersection des sous-espaces $E_i = \mathrm{Vect}(x_i)^\perp$. Si cette intersection n'est pas réduite à $\{0\}$, alors le modèle n'est pas identifiable ! Cependant, soyons raisonnables : en règle générale, on peut supposer que la matrice des observations X est de rang d, comme dans le modèle linéaire classique ; dans ce cas, l'intersection des E_i serait exactement $\ker(X) = \{0\}$ et le modèle serait bien identifiable.

3. Nous venons de nous ramener à la recherche de l'EMV d'un modèle exponentiel, que l'on supposera identifiable. L'EMV, s'il existe, est donc unique ; comme la vraisemblance est strictement positive, l'EMV, s'il existe, maximise aussi la log-vraisemblance

$$\ell(\lambda) = -\sum_{i=1}^{n} e^{x_i\lambda} + y_i x_i \lambda - \ln \prod_{i=1}^{n} y_i!.$$

Il suffit donc de s'assurer que cette fonction possède au moins un maximum sur $\mathbb{R}^d$. Pour cela, il est suffisant de montrer que $\lim_{|\lambda|\to\infty} \ell(\lambda) = -\infty$ (coercivité). C'est assez clair, puisque ℓ a le même comportement que $x - e^x$, qui tend vers $-\infty$ dès que $x \to \pm\infty$; cependant, pour faire les choses proprement, il faut s'assurer que si $|\lambda| \to \infty$, alors au moins un des termes $x_i\lambda$ n'est pas borné. Si ce n'était pas le cas, il existerait une même constante K telle que $|x_i\lambda| \leqslant K$ pour tout i, autrement dit $\cap\{z : |x_i z| \leqslant K\}$ n'est pas une partie bornée puisqu'elle contient des éléments dont la norme tend vers l'infini. Montrons que si X est de rang d, c'est absurde.

Soit $|z_n| \to \infty$ tel que $|x_i z_n|$ est bornée pour tout i. Décomposons $z_n = \alpha_n + w_n$ avec α_n dans l'espace vectoriel généré par les x_i et w_n dans son orthogonal, qui n'est autre que le noyau de X. Il est facile de montrer que si $x_i z_n$ est bornée pour tout i, alors α_n est elle-même bornée ; on en déduit que si z_n n'est pas bornée, c'est que w_n n'est pas bornée. Mais cela n'est possible que si $\ker X$ n'est pas réduit à $\{0\}$, un cas que nous avons écarté ci-dessus.

3. La log-vraisemblance ℓ est $\mathscr{C}^\infty$. Ses points critiques sont les λ qui vérifient $\nabla\ell(\lambda) = 0$, c'est-à-dire

$$\sum_{i=1}^{n} x_i^\top e^{x_i\lambda} = y_i x_i^\top.$$

Solution de l'exercice 4

1.

a) On a

$$N_r = \sum_{k=1}^{n} \mathbf{1}_{X_i = r}$$

donc par linéarité et en utilisant la loi des X_i,

$$\mathbb{E}[N_r] = \sum_{k=1}^{n} \mathbb{P}(X_i = r) = \sum_{k=1}^{n} e^{-\theta_k} \frac{\theta_k^r}{r!}.$$

De même, par indépendance des X_i,

$$\mathrm{Var}(N_r) = \sum_{i=1}^{n} \mathrm{Var}(\mathbf{1}_{X_i=r}) = \sum_{i=1}^{n} e^{-\theta_i} \frac{\theta_i^r}{r!} \left(1 - e^{-\theta_i} \frac{\theta_i^r}{r!}\right)$$

b) La probabilité qu'aucun spécimen de l'espèce i n'ait été capturée en première collecte est $e^{-\theta_i}$. La probabilité qu'au moins un spécimen ait été capturé en seconde collecte est $1 - e^{-\theta_i}$. Les deux sont indépendants, donc la réponse est $e^{-\theta_i}(1 - e^{-\theta_i})$.

2. On vient de voir que $\mathbb{E}[N] = \sum_i e^{-\theta_i}(1 - e^{-\theta_i})$. En développant le terme $1 - e^{-\theta_i}$ on obtient le résultat. Il n'y a même pas besoin de justifier l'interversion des séries, car l'une d'elle est finie.

3. Il suffit de mettre le résultat de la question 1)a) dans la question 2.

4. Ce résultat permet d'estimer le nombre d'espèces que l'on n'a pas capturées *sans avoir à connaître le nombre total d'espèces* : en effet, l'estimateur $\hat{N}$ n'utilise pas du tout la connaissance de S. Le problème de cet estimateur est dans sa variance, qui peut être très élevée ; et dans l'hypothèse selon laquelle le nombre d'espèces capturées suit une loi de Poisson, qui peut être très discutable.

Solution de l'exercice 5

1. C'est un test de Student tout simple. En effet, sous l'hypothèse nulle, les d coordonnées de x_i (que l'on notera $x_i^1, \ldots, x_i^d$ pour clarifier les choses) sont des variables aléatoires gaussiennes iid de loi $N(0, \sigma_i^2)$. Sous cette hypothèse nulle, on a donc

$$\frac{\bar{x}_i}{\hat{\sigma}_i} \sim \mathscr{T}(d-1)$$

où $\hat{\sigma_i}^2 = (d-1)^{-1} \sum_{k=1}^{d} (x_i^k - \bar{x}_i)^2$ et $\bar{x}_i = (x_i^1 + \ldots + x_i^d)/d$. Le test de niveau de confiance $1 - \alpha$ sera donc $|\bar{x}_i/\hat{\sigma}_i| > t_{d-1,1-\alpha}$ où $t_{d-1,1-\alpha}$ est le quantile symétrique de $\mathscr{T}(d-1)$ d'ordre $1 - \alpha$.

2. Les n tests sont indépendants et la probabilité sous $H_{0,i}$ de rejeter le i-ème test est α. Donc la probabilité de ne pas rejeter est $1 - \alpha$, la probabilité de ne rejeter aucune hypothèse est $(1-\alpha)^n$, et cela tend vers 0 lorsque $n \to \infty$. Autrement dit, même si toutes les hypothèses nulles sont vraies, il y a au moins une hypothèse nulle qui sera rejetée à tort !

3. On vient de voir que la probabilité de ne rejeter aucune hypothèse nulle si elles sont toutes vraies est $(1 - \delta)^n$. On cherche donc δ de sorte que $(1 - \delta)^n = 1 - \alpha$ soit $\delta = 1 - (1 - \alpha)^{1/n} \approx \alpha/n$.

4. a) Le nombre N d'hypothèses rejetées par $T_\delta^{(i)}$ est une binomiale de paramètres (n, δ), puisque c'est la somme des variables de Bernoulli indépendantes $\mathbf{1}_{T_\delta^{(i)}=1}$.

b) Par l'inégalité de Hoeffding version non-symétrique,

$$\mathbb{P}(N > t) = \mathbb{P}(N - n\delta > t - n\delta) \leqslant e^{-2(t-n\delta)^2/n}$$

pourvu que $t - n\delta > 0$. Avec $t = kn/100$, on calcule bien la probabilité pour que plus de $k\%$ des hypothèses soient rejetées à tort. On résout donc

$$e^{-2n(k/100-\delta)^2} = \alpha$$

c'est-à-dire $\sqrt{\ln(1/\alpha)/2n} = |k/100 - \delta| = k/100 - \delta$. Le bon choix est donc

$$\delta = \frac{k}{100} - \sqrt{\frac{\ln(1/\alpha)}{2n}}.$$

Comme k est fixé et que n a vocation à être grand, un choix de δ positif existe bel et bien. Lorsque $n \to \infty$, ce niveau tend vers $k\%$. C'est normal : comme tous les tests sont indépendants et que chacun a une probabilité $\delta \approx k\%$ de rejeter, si l'on fait une infinité de tests on aura bien environ $k\%$ de rejets.

À titre d'illustration, si l'on choisit $n = 1000$, $\alpha = 5\%$ et $k\% = 10\%$ on obtient $\delta \approx 0.06$. Avec ces choix, on est sûr que si *toutes* les hypothèses nulles $H_{0,i}$ étaient vraies, on aurait 95% de chances d'en rejeter à tort moins de 100 d'entre elles. Autrement dit, avec grande probabilité, le nombre de faux positifs ne serait pas « trop » grand.

Ce genre de problèmes s'appelle *problèmes de tests multiples*. Il existe des algorithmes bien plus intelligents que celui exposé plus haut pour contrôler le nombre de faux positifs quand on fait beaucoup de tests indépendants !