TP Statistiques 6

Juhyun Park, Angie Pineda, Nicolas Brunel

13 mai 2022

Tests d'Hypothèse

Tests paramétriques

Pour les échantillons $X_i \sim \text{LogNormal}(\mu, \sigma^2), i = 1, \ldots, n$, c'est-à-dire, $\log X_i \sim N(\mu, \sigma^2)$ avec

$$EX_i = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad Var(X_i) = \exp(\sigma^2 - 1)\exp(2\mu + \sigma^2),$$

considérez un test d'hypothèse simple

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu = \mu_1$$

où $\mu_1 > \mu_0$ et $\sigma = \sigma_0$ est connu. Pour $\mathbf{X} = (X_1, \ldots, X_n)$ donné, nous savons que le test de Neyman-Pearson (NP) rejeté H_0 si

$$T(\mathbf{X}) > k_\alpha$$

pour une valeur seuil appropriée k_α. On rappelle que

$$\alpha = P_{H_0}(T(\mathbf{X}) > k_\alpha) \quad \beta = P_{H_1}(T(\mathbf{X}) > k_\alpha)$$

Celles-ci donnent des garanties théoriques pour contrôler les erreurs de décision, α et $1 - \beta$.

1. Construire le NP test. Quelle est la statistique du test, $T(\mathbf{X})$? Étant donné $n = 10, \sigma_0 = 1, \mu_0 = 0, \mu_1 = 0.1$, évaluer les valeurs théoriques pour k_α et β. Quelle est l'interprétation de ces valeurs α et β?

2. Simulez les données avec le paramètre ci-dessus et effectuez le test de niveau $\alpha = 0.1$ $M = 100$ fois. Donnez un approxmation de α et β. Le test contrôle-t-il l'erreurs comme promis ?

3. Au lieu de déterminer k_α pour les tests, nous pouvons calculer la valeur p, définie comme

   $$p - val = P_{H_0}(T(\mathbf{X}) > T(\mathbf{x}))$$

   où $T(\mathbf{x})$ est la statistique du test observée. Expliquez comment utiliser la valeur p pour établir une règle de décision pour le test.

4. Répéter pour les tailles d'échantillon croissantes $n = 20, 50, 100$. Quelle est l'influence de la taille de l'échantillon n sur le test?

5. Considérer le cas où σ est inconnu et répéter les questions précédentes. Y a-t-il une différence dans votre conclusion?

6. Rappelez-vous nos données sur l'ozone. Nous souhaitons savoir comment la distribution des mesures de l'ozone varie entre les sites urbains et ruraux. Nous désignons les données sur l'ozone du site urbain par X_i et le site rural par Y_i, $i = 1, \ldots, n$, l'indice indiquant les n jours différents pour lesquels nous avons des mesures. En supposant que les différences $D_i = X_i - Y_i$ forment un échantillon iid suivant (i) une loi normale $N(\mu, \sigma^2)$ et (ii) une loi log-normale, $\text{LogNormal}(\mu, \sigma^2)$, nous consideron tester

   $$H_0 : \mu = 0 \quad vs \quad H_1 : \mu \neq 0\,.$$

   Quelle est la différence qualitative du comportement de l'échantillon sous l'hypothèse nulle avec le modèle normal et log-normal ? Effectuer le test pour les données en été et en hiver séparément. Quelle est la conclusion?

Méthodes de simulation pour les tests d'hypothèses

Un test d'hypothèse valide exige que l'hypothèse nulle soit rejetée à tort avec une proportion appropriée (par exemple, au plus 5% de fois).

Ainsi, pour une statistique de test donnée, $T(\mathbf{X})$ est telle que nous rejetterons l'hypothèse nulle si $T(\mathbf{X})$ est plus grande (ou plus petite) qu'un certain seuil. Si nous voulons un test de taille α nous devons pouvoir calculer k_α tel que

$$\Pr(T(\mathbf{X}) > k_\alpha | \text{Hypothèse nulle vraie}) = \alpha.$$

Si nous ne pouvons pas calculer k_α analytiquement, nous pouvons utiliser la simulation pour choisir un k_α approprié. Ce que nous devons faire est de simuler des ensembles de données répliqués sous l'hypothèse Nulle.

Approximation Monte Carlo de la distribution de la statistique de test

(1) Simuler des M ensembles de données indépendants, $\mathbf{x}_1, \ldots, \mathbf{x}_M$ sous l'hypothèse nulle.

(2) Pour chaque $\mathbf{x}_i$, calculez la statistique de test $T(\mathbf{x}_i)$.

(3) Les valeurs $\{T(\mathbf{x}_1, \ldots, T(\mathbf{x}_M)\}$ sont les échatillion de la loi de $T(\mathbf{X})$, desquelles on obtient la fontion de répartition empirique comme l'approximation de Monte Carlo.

(4) Estimez le seuil k_α comme le centile empirique $100(1-\alpha)$ de cette approximation.

Notez que pour implémenter cela, nous avons seulement besoin de pouvoir simuler des données sous l'hypothèse nulle. Dans de nombreuses applications scientifiques, les scientifiques effectueront un test d'hypothèse en utilisant cette approche. Ils/Elles décideront d'une statistique du test, en fonction de leur compréhension du problème, et utiliseront la simulation pour calculer la valeur seuil.

Tests non-paramétriques: test de permutation et approximation Monte Carlo.

Pour les données sur l'ozone, si l'on suppose que les lois sont les mêmes sur les deux sites, l'espérance de la différence est égale à zero.

Ainsi, une statistique du test appropriée est la moyenne de l'échantillon des différences, $T(\mathbf{X}, \mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \bar{D}$. La distribution des différences observées est raisonnablement proche d'une distribution normale.

Cependant, pour l'instant, nous préférons faire le moins d'hypothèses possible sur la distribution sous-jacente et nous supposons donc que la distribution des différences est seulement symétrique autour de 0.

$H_0 : D_1, \ldots, D_n$ suivent la loi symmetric autour 0. $\quad H_1 : D_1, \ldots, D_n$ ne suivent pas la loi symmetric autour 0.

Cela signifie que la probabilité d'obtenir une différence d est identique à celle d'obtenir une différence $-d$ pour toute valeur de d. Nous avons obtenu l'échantillon $(d_1, \ldots, d_n)$. Si notre hypothèse selon laquelle la loi des différences est symétrique par rappport à zéro est vraie, nous aurions également pu obtenir l'échantillon

$$\begin{aligned}
(d_1^*, \ldots, d_n^*) &= (-d_1, \ldots, -d_n) \quad \text{ou} \\
(d_1^*, \ldots, d_n^*) &= (-d_1, d_2, \ldots, d_{n-1}, -d_n) \text{ou} \\
(d_1^*, \ldots, d_n^*) &= \ldots \quad \text{etc.}
\end{aligned}$$

Pour couvrir tous ces vecteurs, il faudrairt 2^n possibilité de ce type ! Ceci définit **un test de permutation**.

Pour gagner du temps, nous changeons simplement de façon aléatoire les signes des valeurs de données indépendamment pour chaque observation d_i est nous générons ensuite de façon répétée un nombre raisonnable de tels vecteurs de données. Nous voulons voir à quel point notre échantillon est inhabituel parmi tous ces échqntillons alternatifs possibles.

Puisque nous utilisons la moyenne de l'échantillon comme la statistique du test, nous évaluons la moyenne d'échantillon

$$\bar{d}^* = \frac{1}{n}\sum_{i=1}^{n} d_i^*$$

pour chaque vecteur. Ce sont les échantillons de la statistique du test sous l'hypothèse nulle que nous avons générés sans supposer aucun modèle paramétrique. Nous pouvons utiliser la loi empirique de ces échantillons pour déterminer la région de rejet.

7. Compléter le test non-paramétrique pour les données dur l'ozone, en été et en hiver. Quelles sont les hyptohèses de ce test? La conclusion est-elle cohérente avec celle du test paramétrique ?

Hypothèses avec plusieurs paramètres: Test du rapport de vraisemblance

La méthode basée sur la simulation n'est pas limitée aux modèles non-paramétriques comme nous l'avons démontré ci-dessus. Pour les modèles paramétriques complexes, il est délicat de dériver la distribution d'échantillonnage de la statitistique du test sous l'hypothèse nulle, mais souvent beaucoup plus facile de simuler les échantillons à partir de celui-ci.

Considérons des variables aléatoires indépendantes, $X_1, \ldots, X_n$ issues d'une loi de P_X de paramètre $\boldsymbol{\theta}_1$, et $Y_1, \ldots, Y_n$ issues d'une loi de P_Y de paramètre $\boldsymbol{\theta}_2$, indépendantes des X's. Nous souhaitons tester

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \quad vs \quad H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$$

Notre paramètre est noté par $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Un choix naturel de statistique du test pour de telles hypothèses est la statistique du test du rapport de vraisemblance definie par

$$T(\mathbf{X}) = 2 \log \frac{\max_{\boldsymbol{\theta} \in H_1} L(\boldsymbol{\theta}; \mathbf{X})}{\max_{\boldsymbol{\theta} \in H_0} L(\boldsymbol{\theta}; \mathbf{X})}$$

La loi asymptotique de cette statistique est connue pour suivre une distribution du χ^2 avec un nombre de degrés de liberté donné par $df = dim(H_1) - dim(H_0)$. Nous utilisons ici l'approche basée sur la simulation, valable pour un échantillon fini n.

8. Pour les données sur l'ozone, on suppose que les valeurs d'ozone pour chaque site (en été) suivent une loi log normale. Effectuez le test du rapport de vraisemblance pour les hypothèses

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2 \,.$$

Quelle est la conclusion?

!! Bonus !! : Test d'ajustement de Kolmogorov (partie optionnelle - à faire si les questions 1 à 8 sont réalisées)

Nous pouvons appliquer la stratégie de simulation pour évaluer la pertinence des modèles statistiques pour les données.

Soit $X_1, \ldots, X_n$ un échantillon de loi inconnue $P_{\boldsymbol{\theta}}$ de fonction de répartition F supposée continue. L'objectif du test de Kolmogorov est l'ajustement de la loi inconnue P à une loi connue P_0 de fonction de répartition continue F_0:

$$H_0 : F = F_0 \qquad H_1 : F \neq F_0$$

9. Constuire le test de Kolmogorov (Kolmogorov-Smirnov) de niveau α (sur la base de l'approximation asymmtotique). Suggérer une méthode alternative basée sur la simulation.

Supposons que le modèle le mieux adjusté ait la valeur de paramètre $\hat{\boldsymbol{\theta}}$. Soit F_0 la fonction de répartition du modèle ajusté $P_{\hat{\boldsymbol{\theta}}}$.

10. Pour les données sur l'ozone, nous voulons tester si l'thypothèse de gaussianité était appropriée. Nous envisageons deux scénarios. Le premier est que les données originales de l'ozone suivent une loi gaussienne ($H_0^{(1)}$). Le second suppose que seules les différences suivent la loi gaussienne ($H_0^{(2)}$). Effectuez les tests utilisant la méthode asymptotique et la méthode de simulation. Résumez vos conclusions.

11. Répétez les tests pour l'hypothèse de modèles log-gaussiens. Résumez vos conclusions.

12. Sur la base de vos différentes analyses des ensembles de données sur l'ozone effectuées tout au long du cours, résumez vos conclusions.