TP Statistiques 4

Juhyun Park, Phuong Thuy Vo, Lassaad Mchri, Nicolas Brunel

28 avril 2023

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. Consideron le cas o'u σ est inconnu et répétez les questions précédentes. Y a-t-il une différence dans votre conclusion?

Application: Air quality monitoring

Airparif exploite un système de surveillance de la qualité de l'air avec un réseau de sites dans la région de la capitale (Ile de France) sur lesquels les mesures de la qualité de l'air sont effectuées automatiquement. Ces mesures sont utilisées pour résumer les niveaux actuels de pollution atmosphérique, pour prévoir les niveaux futurs et pour fournir des données pour la recherche scientifique, contribuant à l'évaluation des risques pour la santé et des impacts environnementaux des polluants atmosphériques.

Nous examinerons *l'ozone troposphérique* (O_3). Ce polluant n'est pas émis directement dans l'atmosphère, mais est produit par des réactions chimiques entre le dioxyde d'azote (NO_2), les hydrocarbures et la lumière du soleil.

Nous nous concentrerons sur les données de deux sites de surveillance: un site urbain à Neuilly-sur-seine (**NEUIL**) et un site rural (**RUR.SE**) près de la forêt de Fontainbleu.

Les données de chaque site sont des mesures quotidiennes de la concentration moyenne horaire maximale de O_3 enregistrée en microgrammes par mètre cube ($\mu g/m^3$), de 2014 à 2019 inclusivement. Pour nous concentrer sur la question de la saison, nous comparons les données de *hiver* (novembre-février inclus) (`Ozone_hiver.csv`) et *été* (mai - août inclus) (`Ozone_ete.csv`).

Nous souhaitons savoir comment la distribution des mesures de l'ozone varie-t-elle 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 et définissons la variable $D_i = X_i - Y_i$ pour la difference.

6. Appliquer l'analyse exploratoire des données (TPs 1-2) et suggérer un modèle approprié pour D_i.
7. En supposant que les différences D_i forment un échantillon iid suivant une loi normale $N(\mu, \sigma^2)$, quelle est l'hypothèse sous-jacente que nous voulons tester ? Définir H_0 et H_1 et effectuez le lest pour les données en été et en hiver séparément. Quelle est la conclusion?

Méthodes de simulation pour les tests d'hypothèse

Un test d'hypothèse valide exige que nous rejetions incorrectement l'hypothèse nulle une proportion appropriée du temps (par exemple, au plus 5% de fois).

Ainsi pour une statistique de test donnée, $T(\mathbf{X})$ 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 aider à choisir un k_α approprié. Ce que nous devons faire est de simuler des ensembles de données répliqués sous l'hypothèse Nulle.

Monte Carlo approximation to the sampling distribution of the test statistic

(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.

Test d'ajustement de Kolmogorov

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:

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

8. 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 $P_{\hat{\boldsymbol{\theta}}}$.

9. 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.