MATH0487-2 : Devoir

L'équipe de MATH0487-2

Octobre 2021

Instructions

Objectifs. Les objectifs de ce devoir sont les suivants : extraire différentes statistiques descriptives, calculer et comparer des estimateurs ponctuels pour les paramètres d'un modèle statistique, calculer et comparer des estimateurs par intervalle pour les paramètres d'un modèle statistique, réaliser des tests d'hypothèse sur les moyennes de population.

Délivrables. Ce devoir doit être réalisé par groupe de 2 étudiants maximum. Chaque groupe doit rendre :

- Son rapport nommé `devoir-rapport.pdf`.

  Le rapport doit présenter les réponses aux questions posées, ainsi que les raisonnements, calculs symboliques et valeurs numériques qui vous ont permis de les obtenir. Le rapport doit être généré en LaTeX en remplissant le modèle `devoir-rapport.tex` (sans modifier la mise en page dans le préambule du modèle).

  Le rapport doit être écrit de manière *claire* et *concise*. Nous évaluerons évidemment la justesse de vos réponses mais aussi la qualité de votre rapport. Nous estimons que votre rapport comportera une dizaine de page. Le rapport doit respecter les indications du document `devoirs-indications.pdf`.

- Son code source avec un fichier principal nommé `main.x` (où `x` est l'extension de votre langage).

  Nous recommandons d'utiliser Python comme langage de programmation. Cependant, vous êtes libres de choisir le langage de programmation de votre choix (*e.g.*, R, Julia, Matlab, etc.). Un tutoriel d'installation et d'introduction aux concepts de base est disponible via le cours INFO8006 (tutoriel, video Linux, video Windows).

  Le fichier principal `main.x` doit être exécutable et générer l'ensemble des figures et résultats numériques demandés dans le devoirs. En dehors de cette contrainte, vous pouvez ensuite organiser votre code comme vous le souhaitez. N'oubliez pas de soumettre toutes les fonctions nécessaires à la bonne exécution de votre code, y compris les fonctions fournies avec l'énoncé.

  Le code source doit être écrit de manière *claire* et *commentée*. Le code doit respecter les indications du document `devoirs-indications.pdf`.

Le rapport et le code source de votre devoir doivent être soumis séparément sur la plateforme Gradescope (`https://www.gradescope.com/`). Utilisez votre adresse email `@student.uliege.be` pour créer un compte sur Gradescope et avoir accès au cours. N'attendez pas la dernière minute afin de vérifier que vous avez accès au cours sur Gradescope !

La date limite de soumission est fixée au 18 décembre 2021 à 23h59. Jusqu'à cette date, vous avez la possibilité de (re)soumettre votre rapport ou votre code autant de fois que vous le souhaitez. Au-delà de cette date, il ne sera plus possible de soumettre le devoir. N'attendez pas la dernière minute pour soumettre une première version de votre travail !

Si vous êtes dans l'incapacité de rendre votre travail pour la date de soumission pour une raison sérieuse, contactez l'équipe enseignante dès que possible pour trouver un arrangement.

Questions. Toutes vos questions sur le devoir doivent être postées dans le forum de *Ed Discussion* du cours sous la catégorie *Assignments/Homework* (une question par fil de discussion). Vos questions sur l'utilisation de LaTeX peuvent être postées sous la catégorie *LaTeX*.

Politique de collaboration. Vous pouvez discuter du devoir avec d'autres groupes, mais *vous devez écrire vous-même vos propres solutions, et écrire et exécuter vous-même votre propre code*. Copier la solution de quelqu'un d'autre, ou simplement apporter des modifications triviales pour ne pas copier textuellement, n'est pas acceptable.

Présentation du problème

Pour illustrer l'utilité des méthodes statistiques présentées dans ce cours, nous allons discuter dans ce devoir de deux sujets d'actualité : le réchauffement climatique et les inégalités salariales.

Pour ce faire, vous allez analyser au niveau mondial trois variables et leurs relations.

- `PIB_habitant` : Le *Produit Intérieur Brut (PIB) par habitant* est la valeur totale de tous les biens et services produits dans un pays donné au cours d'une année donnée.
- `CO2_habitant` : L'*empreinte* CO_2 *par habitant*, exprimée en tCO_2/an, comptabilise l'ensemble des émissions de CO_2 dues aux activités directement faites dans un pays, mais aussi les "émissions grises", liées à l'importation de biens et services à partir d'autres pays sur un an. Il ne faut pas la confondre avec le bilan CO_2 national qui ne prend pas ces importations en compte et qui est donc moins représentatif des impacts d'un pays sur le climat.
- `Top10` : Le *pourcentage de revenu national détenu par les* 10 % *des habitants les plus aisés de chaque pays* est un bon indicateur du taux d'inégalités à l'intérieur d'un pays.

L'ensemble de ces données est tiré de la base de données World Inequality Database, initiée en partie par Thomas Piketty, expert international des inégalités mondiales. Il s'agit à ce jour d'une des bases de données la plus complète sur les inégalités.

Dans le fichier `data.csv`, vous avez accès à ces trois variables pour 167 pays différents. Afin d'individualiser les résultats que vous allez obtenir, chaque groupe disposera d'une *population personnalisée* de 150 pays. Pour ce faire, appelez la fonction `population` qui vous est fournie (et qu'il faut laisser intacte) avec en arguments le jeu de données et une liste des identifiants ULiège (*e.g.*, `20123456`) des membres de votre groupe.

1 Analyse descriptive

Toute étude statistique commence par une analyse descriptive des données. Dans cette partie, vous allez vous familiariser avec les variables dans *votre population*.

(a) Pour vous donner une idée des ordres de grandeur de chaque variable, vous allez d'abord extraire les valeurs de ces variables pour quatre pays qui sont dans votre population : les États-Unis (nommé `USA` dans la base de données), la Belgique (`Belgium`), la Chine (`China`) et le Togo (`Togo`). Présentez dans un tableau, les valeurs des trois variables pour chacun de ces pays et analysez.

(b) Ensuite, vous allez analyser plus en détails la distribution de ces variables. Pour chacune d'entre elles, vous allez calculer différentes valeurs. En particulier :

  i. Calculez la moyenne et l'écart-type.

  ii. Calculez la médiane et les quartiles. Y a-t-il des données aberrantes ? Tracez un graphique par variable illustrant ces valeurs. Comment s'appelle ce type de graphique ? Que peut-on dire en comparant ces trois graphiques ?

  iii. Tracez l'histogramme et la fonction de répartition empirique pour chaque variable. Comparez.

(c) Enfin, vous allez analyser les relations qui existent entre les différentes variables. Comparez, numériquement et graphiquement, les trois couples de variables. Pour les graphiques, organisez vos nuages de points sous la forme d'une "matrice" (comme proposé dans la vidéo de Jean-luc Doumont).

2 Estimation ponctuelle

Dans cette partie, vous allez construire et comparer des estimateurs ponctuels pour les paramètres d'un modèle statistique de la variable `Top10`. Cette variable prend des valeurs dans l'intervalle $[0, 1]$. Étant

donné la grande variété de formes qu'elle peut prendre, vous allez utiliser comme modèle statistique une distribution Beta(a, b) de paramètres a et b inconnu.

Vous êtes curieux de découvrir les différences entre deux méthodes d'estimation ponctuelle de paramètres : la méthode des moments (MOM) et la méthode du maximum de vraisemblance (MLE).

Vous commencez par tester ces méthodes sur un échantillon de 50 pays.

(a) Démontrez mathématiquement les formules des estimateurs $\hat{a}_{\text{MOM}}$ et $\hat{b}_{\text{MOM}}$ des paramètres a et b obtenus en utilisant la méthode des moments.

(b) Calculez les valeurs de ces paramètres sur votre échantillon de 50 pays.

(c) Vous vous apprêtez à faire de même pour déterminer les estimateurs du maximum de vraisemblance quand un de vos collègues vous explique qu'il n'existe pas de formule analytique du maximum de vraisemblance pour une distribution Beta(a, b) avec deux paramètres inconnus. Par contre, il vous fait remarquer qu'on peut le déterminer numériquement en maximisant la fonction de log-vraisemblance suivante :

$$\log L(a, b; \mathbf{x}) = (a-1)\sum_{i=1}^{n} \log(x_i) + (b-1)\sum_{i=1}^{n} \log(1-x_i) - n \log \beta(a, b),$$

où $\mathbf{x}$ correspond à vos données, n le nombre de ces données et $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ est la constante multiplicative de la distribution Beta (voir cours de probabilité).
Démontrez que la formulation mathématique de la log-vraisemblance d'une distribution Beta donnée ci-dessus est bien valide.

(d) Votre collègue a même déjà codé cette fonction qu'il a appelé `beta_loglikelihood`. Utilisez cette fonction et la fonction `scipy.optimize.minimize` (Python) ou `fminsearch` (Matlab) pour calculer les estimateurs du maximum de vraisemblance $\hat{a}_{\text{MLE}}$ et $\hat{b}_{\text{MLE}}$ de a et b sur le même échantillon qu'au point 2(b).
Astuce : Utilisez $(1, 1)$ comme valeur initiale de (a, b) dans l'algorithme de minimisation.

(e) Superposez l'histogramme de votre population avec la distribution Beta(a, b) associée aux paramètres calculés avec chacune des méthodes. Que concluez-vous ?

Une étude scientifique poussée a démontré que les valeurs des paramètres a et b étaient respectivement égales à 13.35 et 16.31. Vous considérerez ces valeurs comme les *vraies valeurs* des paramètres a et b. Connaissant ces vraies valeurs, vous vous intéressez à la qualité de vos estimateurs. Pour ce faire, vous tirez 500 échantillons i.i.d. de 50 pays dans votre population et appliquez la procédure suivante.

(f) Pour chaque échantillon, calculez les estimateurs $\hat{a}$ et $\hat{b}$ en utilisant la méthode des moments. Calculez le biais, la variance et l'erreur quadratique moyenne des estimateurs MOM.

(g) Faites de même pour les estimateurs MLE obtenus par la méthode du maximum de vraisemblance.

(h) Comparez les résultats obtenus pour les estimateurs MOM et MLE. Quelle méthode donne les meilleurs estimateurs ? Discutez.

Bonus

(i) Votre boss vous dit qu'il serait intéressant de réitérer l'expérience précédente pour différentes tailles d'échantillons (20, 40, 60, 80 et 100 pays) afin d'étudier comment la qualité des estimateurs évolue avec la quantité de données disponibles. Il vous promet même un bonus si vous faites cette analyse.

3 Estimation par intervalle

Dans cette partie, vous vous intéresserez plus particulièrement à la variable `PIB_habitant`. Vous faites l'hypothèse que cette variable suit une distribution Exponentielle de paramètre λ. On vous demande de construire des intervalles de confiance pour ce paramètre λ.

Vous décidez de calculer cet intervalle en utilisant deux méthodes : la méthode du pivot et la méthode du bootstrap. Vous commencez par tester ces méthodes sur un échantillon de 50 pays.

(a) Définissez mathématiquement un intervalle de confiance à 95 % pour la variable d'intérêt en utilisant la méthode du pivot. Précisez quel est votre pivot et détaillez vos calculs.

(b) Calculez cet intervalle pour votre échantillon de 50 pays.

(c) Expliquez comment en utilisant la méthode du bootstrap, vous pouvez déterminer un intervalle de confiance à 95 %.

(d) Utilisez cette méthode pour calculer un intervalle à 95 % sur le même échantillon qu'au point 3(b) avec 100 échantillons bootstrap.

Pour comparer vos deux méthodes, vous décidez de calculer des intervalles de confiance pour des tailles d'échantillons de 5 à 50 (avec un incrément de 5). Pour chaque taille d'intervalle, tirez 500 échantillons. Pour chaque échantillon, construisez deux intervalles de confiance à 95 % en utilisant d'une part la méthode du pivot et d'autre part la méthode du bootstrap avec 100 échantillons bootstrap.

(e) Analysez l'évolution de la largeur moyenne de ces intervalles en fonction de la taille d'échantillon. Comparez les deux méthodes.

(f) Analysez l'évolution de la proportion d'intervalles contenant la *vraie valeur* du paramètre λ en fonction de la taille d'échantillon. Une étude scientifique a montré que la vraie valeur du paramètre λ pour la population est égale 5.247×10^{-5}. Comparez les deux méthodes.

(g) Était-il raisonnable de supposer que la variable suit une distribution Exponentielle ?

4 Test d'hypothèse

De nombreuses associations pour le climat affirment que la justice sociale mondiale est au coeur du débat climatique car "les pays riches émettent en moyenne plus de CO_2 par habitant que les pays pauvres".

Une équipe de scientifiques a voulu vérifier cette affirmation et est arrivée à la conclusion que la moyenne d'émission des pays riches est égale à la moyenne d'émission des pays pauvres plus un Δ que vous pouvez obtenir en utilisant la fonction `scientific_delta`. Vous pensez qu'en réalité la différence entre ces deux moyennes est supérieure à Δ et décidez de faire un test d'hypothèse pour vérifier laquelle de ces hypothèses est valide.

On précise que les pays riches sont ceux ayant un PIB par habitant supérieur ou égal au PIB par habitant médian dans votre population de pays.

(a) Formulez une hypothèse nulle et l'hypothèse alternative correspondante qui vous permettront de réaliser un test statistique pour établir votre comparaison. Vérifiez laquelle de ces hypothèses est vraie dans votre population.

(b) Décrivez comment réaliser un test d'hypothèse au seuil de signification $\alpha = 5\,\%$ pour l'hypothèse formulée précédemment sur un échantillon de taille $n = n_{\mathrm{r}} + n_{\mathrm{p}}$ où n_{r} est le nombre de pays riches et n_{p} est le nombre de pays pauvres. Considérez que les variances de population sont égales et inconnues.

(c) Tirez 100 échantillons i.i.d. de 75 pays. Pour chaque échantillon, testez l'hypothèse au seuil de signification $\alpha = 5\,\%$. Parmi les 100 échantillons, dans quelle proportion l'hypothèse nulle est-elle rejetée ? Comparez cette valeur à α. Interprétez.

(d) Sans modifier votre méthode de construction du test, répétez la même expérience avec 100 échantillons i.i.d. de **25** pays. Comparez au point précédent et interprétez.

Bon travail !