Mini-Projet

Consignes

La date limite pour rendre le projet est le 15 Octobre 2019, 23h59, sur l'espace de rendu de l'onglet 'mini-projet 2019' du site pédagogique, où vous trouverez des consignes plus détaillées. https://sitepedago.telecom-paristech.fr/front/frontoffice.php?SP_ID=3207#R3803 Le langage à utiliser est `R` et le rendu est attendu sous la forme d'un notebook réalisé avec `R-studio` ou Jupyter notebook (avec le noyau `R`).

Rappels et compléments de cours :

— La p-valeur est la probabilité que, sous l'hypothèse nulle, la statistique de test prenne une valeur au moins aussi extrême que celle qui a été observée.

— La fonction puissance est la probabilité de rejeter sous l'hypothèse alternative H_1 : $h(\tilde{\lambda}) = \mathbb{P}[T \in \bar{A} \,|\lambda = \tilde{\lambda}\,]$ pour $\lambda \in \Lambda_{H_1}$ où T est la statistique de test, $\bar{A}$ et la région de rejet, λ est le paramètre à tester et Λ_{H_1} signifie l'ensemble des paramètres appartenant à la région de l'hypothèse alternative.

— Si X et Y sont deux variables indépendantes distribuées respectivement selon une loi de Poisson de paramètre λ et μ, alors $X+Y$ suit une loi de Poisson de paramètre $\lambda+\mu$.

On s'intéresse au nombre de 'grandes' inventions et découvertes scientifiques par an, entre 1860 et 1959. Le jeu de données `discoveries` disponible dans `R` est accessible par la commande `data(discoveries)` . Chargez ainsi le jeu de données et inspectez-le. Dans toute la suite on notera x_i la i^{eme} entrée du jeu de données et on considérera que les variables aléatoires associées X_i sont indépendantes et identiquement distribuées.

Exercice 1 (Analyse exploratoire)**:**

En présence de données positives ou nulles, à valeurs réelles, on est tenté de considérer le modèle des lois de Poisson ou celui des lois géométrique. On rappelle que les fonctions de densité de probabilité (par rapport à la mesure de comptage), de répartition, de quantiles, et le générateur de variables aléatoires correspondant aux lois de poisson et aux lois géométriques sont respectivement, en `R` , `dpois`, `ppois`, `qpois`, `rpois`, `dgeom`, `pgeom` , `qgeom`, `rgeom`.

1. Ajustez une loi géométrique (avec les conventions de la loi `dgeom` de `R`) sur le jeu de données en utilisant l'estimateur de maximum de vraisemblance. Comme le détaille l'aide, la densité de la loi géométrique de support $\mathbb{N}$ et de paramètre $\theta \in]0,1[$ est donnée par $p_\theta(x) = \theta(1-\theta)^x$, $x \in \mathbb{N}$. Quelle est la valeur de l'estimateur pour ce jeu de données ?
2. Même question avec un modèle de Poisson et son paramètre $\lambda > 0$.
3. Comparez les moyennes et variances empiriques du jeu de données avec les espérances et variances théoriques dans les deux modèles considérés. à première vue, quel modèle vous semble plus adapté ?

4. Tracez sur le même graphique la fonction de densité de la loi géométrique estimée, celle de la loi de poisson estimée, et l'histogramme des données (utilisez l'option `probability = TRUE`). Votre première impression est-elle confirmée ?
5. Inspection à l'aide d'un QQ-plot (diagramme quantile-quantile). On pourra consulter la page https://fr.wikipedia.org/wiki/Diagramme_Quantile-Quantile pour une explication détaillée sur les QQ-plots. En résumé, un QQ-plot d'un jeu de données à comparer à une loi de fonction de répartition F et de quantile F^{-1} est le tracé des points $(x_{(i)}, F^{-1}(i/(n+1))$, où $x_{(i)}$ est la i^{eme} plus grande valeur de l'échantillon $(x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$. Puisque $x_{(i)}$ est un quantile d'ordre i/n de la loi empirique, les points doivent être proches de la diagonale si la loi des X_i est proche de la loi candidate F. Tracez les deux qq-plots associés respectivement aux lois géométriques et de poisson calculées précédemment. Commentez.
6. Pour clore le débat, on va effectuer un test d'ajustement du chi-2 pour chacun des modèles. On admet le résultat suivant (voir Saporta (2006), chapitre 14.6.2.1) :

   — Soient $X_i, i \leq n$ des variables discrètes indépendantes de même loi, et soit $\mathcal{I} = \{I_1, \ldots, I_k\}$ un partitionnement du support des X_i. Pour $j \in \{1, \ldots, k\}$, soit $n_j = |\{i : X_i \in I_j\}|$ et soit $p_j = \mathbb{P}(X_i \in I_j)$. Alors la statistique

   $$S = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$$

   suit (lorsque n est grand) une loi du χ^2 (chi-2) à $k-1$ degrés de liberté.

   — Si les p_i sont estimés par maximum de vraisemblance à partir des données X_i, dans un modèle paramétrique de dimension p, alors la statistique S ci-dessus suit une loi du χ^2 (chi-2) à $k-1-p$ degrés de liberté.

   — Le test du χ^2 de niveau de confiance α consiste à rejeter l'hypothèse nulle " H_0 : $Xi \sim \mathrm{P}_\theta$" si la statistique S obtenue en prenant $p_j = \mathbb{P}_\theta(X \in I_j)$ dépasse le quantile d'ordre $1-\alpha$ de la loi du χ^2 associée. Utilisez l'aide de `R( ?qchisq` à propos des quantiles de la loi du χ^2.

   — Pour être valide, les n_j doivent être tous suffisamment grand, l'usage courant consiste à choisir les I_j de sorte qu'on ait $n_j > 5$ pour tout j.

   Effectuez un test d'ajustement du χ^2 pour le modèle de poisson d'une part et le modèle géométrique d'autre part, en prenant comme partitionnement $k = 6$,

   $$\mathcal{I} = \Big\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}, [5, +\infty)\Big\}.$$

   Quel est le degré de libertés de S dans chacun des modèles ? quelle est la p-valeur de votre statistique de test dans chaque cas ? Quel modèle pouvez-vous accepter avec niveau de confiance 5% ?

Exercice 2 (Analyse de l'incertitude dans le modèle de Poisson)**:**
On suppose désormais que les données de l'exercice précédent suivent effectivement une loi de Poisson de paramètre λ inconnu.

1. Montrer que pour tout $s \geq 1$ fixé, la fonction $\lambda \mapsto \mathbb{P}_\lambda(\sum_{i=1}^n X_i \geq s)$ est une fonction croissante de λ.
2. En déduire un test de $H_0 : \lambda \leq 3$ contre $H_1 : \lambda > 3$, au niveau inférieur ou égal à 5% aussi grand que possible sous cette contrainte, basé sur la statistique $T = \sum_{i=1}^n (X_i)$, de type $\delta(T) = \mathbb{1}\{T > s\}$. On précisera s en fonction d'un quantile d'une loi classique et on donnera une valeur numérique obtenue avec `R`.
3. Rejetez-vous ou acceptez- vous l'hypothèse nulle H_0 ?
4. À moyenne empirique $m = \frac{1}{n}\sum_{i=1}^n x_i$ fixée, égale à la moyenne empirique de `discoveries`, quelle est le nombre minimum n_0 de données nécessaires pour pouvoir rejeter H_0 à l'aide d'un test de niveau inférieur ou égal à 5% comme celui construit précédemment ?
5. Tracez la fonction puissance du test précédent $\beta(\lambda) = 1 - \mathbb{P}_\lambda[$ accepter à tort $H_0\,]$ en fonction de λ, pour $n = 100$ données et $\lambda \in (3, 4]$. Combien de données faut-il au minimum pour que $\beta(\lambda = 3.5) \geq 0.9$?

Exercice 3 (Analyse bayésienne)**:**

1. Choisir un prior conjugué pour λ dans le modèle de Poisson (c.f. https://en.wikipedia.org/wiki/Conjugate_prior), c'est à dire un modèle paramétrique pour λ tel que la loi a posteriori appartienne au même modèle paramétrique. Choisir les paramètres du prior tels que l'espérance et la variance de $\boldsymbol{\lambda}$ sous la loi a priori soient respectivement $\mathbb{E}_\pi(\boldsymbol{\lambda}) = 5, \mathbb{V}\mathrm{ar}_\pi(\boldsymbol{\lambda}) = 100$ (prior peu informatif).
2. Calculez l'expression des paramètres de la loi a posteriori en fonction de $x_{1:n}$ et donnez l'expression de l'estimateur de l'espérance a posteriori pour le paramètre λ. Comparez qualitativement avec l'estimateur du maximum de vraisemblance.
3. Application numérique : quelles valeurs numériques (des paramètres a posteriori et de l'espérance a posteriori) obtenez-vous avec les données `discoveries` ?
4. Un intervalle de crédibilité a posteriori de niveau α est un intervalle I de l'espace des paramètres tel que $\mathbb{P}[\boldsymbol{\lambda} \in I | x_{1:n}] \geq \alpha$ On peut construire un tel intervalle grâce aux quantiles d'ordre $(1-\alpha)/2$ et $(1+\alpha)/2$ de la loi a posteriori. Donnez l'expression d'un tel intervalle dans le modèle considéré en fonction de quantiles d'une loi adaptée. Donnez le résultat numérique (bornes supérieures et inférieures de l'intervalle) en utilisant la fonction quantile correspondante dans `R`, pour le niveau $\alpha = 0.95$.

Références

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.