ENSIMAG 1ère année Avril 2022

Principes et Méthodes Statistiques

Durée : 3 heures.

Tous documents manuscrits autorisés.

Le répertoire HOME/exam/ExamenPMS2022 contient l'énoncé de l'examen, le jeu de données `Failures.txt` et un répertoire ChamiloPMS qui vous donnera accès à une version restreinte de la page Chamilo du cours.

Les deux parties sont indépendantes.

Les résultats vus en cours ou en TD peuvent être utilisés sans être redémontrés.

Il sera grandement tenu compte de la qualité de la rédaction (présentation et justification des réponses) dans la notation.

Le rendu de l'examen comprend d'une part une copie manuscrite et d'autre part un script au format `Rmd` nommé Prénom.Nom.Rmd, à sauvegarder dans votre répertoire. Toutes les questions posées doivent faire l'objet d'une réponse manuscrite. Pour certaines d'entre elles, vous pourrez vous appuyer sur les résultats d'une analyse en `R`. Les calculs mathématiques et les commentaires des résultats ne doivent être donnés que sur la copie manuscrite. Le fichier `Rmd` ne contient que le code `R` et les figures demandées.

Il faut impérativement mettre votre nom au début du fichier `Rmd`.

Important : Pour sauvegarder votre travail, commencez par faire une sauvegarde du script sous `RStudio`. Puis il faudra sauvegarder votre répertoire, ce que vous pouvez faire à tout moment via les icônes sur le bureau :

- soit pour sauvegarder l'état actuel de votre répertoire et continuer l'examen (icône "Envoyer") ;
- soit pour sauvegarder le rendu final et arrêter le PC (icône "Finir"). **Cette opération doit impérativement être réalisée avant de quitter la salle.** Vérifiez, à l'aide de l'heure de sauvegarde, que c'est bien la bonne version du fichier `Rmd` qui est présentée à l'écran.

Si le PC crashe, on peut récupérer la dernière sauvegarde en redémarrant la machine.

Vous ne devez pas modifier ni déplacer tout fichier nommé `whoami.txt` : c'est le fichier qui vous identifie. Tout résultat placé en dehors du répertoire examen ou un de ses sous-répertoires sera ignoré.

Barème indicatif : Partie 1 : 8 pts, Partie 2 : 12 pts.

Première partie

On veut étudier l'influence de la prise d'alcool sur le temps de réaction à un certain stimulus. Une première expérience a montré que le temps de réaction (en millisecondes) d'individus sobres à ce stimulus était distribué selon une loi normale de moyenne 70 et d'écart-type 3.6. Une seconde expérience a été menée sur 20 individus ayant tous ingurgité la même quantité d'alcool. Les temps de réaction de ces 20 individus sont les suivants :

73.1	72.4	76.3	77.1	83.6	75.5	78.3	80.7	77.2	77.8
71.3	73.7	82.1	80.3	69.8	74.1	77.6	80.9	68.7	75.2

1. A l'aide de graphiques de statistique descriptive, montrer qu'on peut admettre l'hypothèse que le temps de réaction d'individus alcoolisés est une variable aléatoire de loi normale. Donner des estimations de la moyenne et de l'écart-type de cette loi.

2. On se demande si la prise d'alcool augmente la variabilité du temps de réaction.

   (a) Montrer que l'on peut répondre à cette question à l'aide d'un test d'hypothèses. Donner les hypothèses H_0 et H_1 et la région critique W.

   (b) Faites le test au seuil 5% et conclure.

   (c) Calculer la p-valeur et conclure.

3. On se demande si la prise d'alcool augmente le temps de réaction moyen.

   (a) Montrer que l'on peut répondre à cette question à l'aide d'un test d'hypothèses. Donner les hypothèses H_0 et H_1 et la région critique W.

   (b) Calculer la p-valeur et conclure.

4. Donner un intervalle de confiance bilatéral et deux intervalles de confiance unilatéraux de seuil 5% pour le temps de réaction moyen d'individus alcoolisés. En quoi est-ce que ces résultats confirment celui de la question précédente ?

Deuxième partie

Une variable aléatoire X positive est dite de loi de Pareto $\mathcal{P}a(\theta)$, avec $\theta > 2$ si et seulement si sa fonction de répartition est :

$$F(x) = 1 - \frac{1}{(1+x)^{\theta}}, \forall\, x \geq 0.$$

et sa densité est :

$$f(x) = \frac{\theta}{(1+x)^{\theta+1}}, \forall\, x \geq 0.$$

1. Soient $x_1, \ldots, x_n$ un échantillon de données. Expliquer comment construire un graphe de probabilités pour la loi $\mathcal{P}a(\theta)$ à partir de cet échantillon et à quelles conditions sur ce graphe l'hypothèse que les observations proviennent de cette loi pourra être admise. Expliquer comment calculer une estimation graphique θ_g de θ.

2. Montrer que $E[X] = \dfrac{1}{\theta - 1}$.

3. On suppose que $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{P}a(\theta)$. Calculer l'estimateur des moments $\tilde{\theta}_n$ de θ.

4. Calculer l'estimateur de maximum de vraisemblance $\hat{\theta}_n$ de θ.

5. Donner la loi de probabilité de $Y = \ln(1 + X)$.

6. Montrer que $\hat{\theta}_n$ est biaisé. En déduire un estimateur sans biais $\hat{\theta}'_n$. Montrer que cet estimateur est convergent.

7. On a mesuré les durées en milliers d'heures entre les défaillances successives d'un système informatique, supposées indépendantes et de même loi. Ces données sont dans le fichier `Failures.txt`. Ce jeu de données peut être chargés en `R` grâce à la commande `Failures <-scan("Failures.txt")`, en se positionnant dans le bon répertoire à l'aide de `Session/Set Working Directory`.

   (a) Justifier le fait que l'on admette que cet échantillon est issu de la loi $\mathcal{P}a(\theta)$.

   (b) Donner les valeurs des estimations θ_g, $\tilde{\theta}_n$, $\hat{\theta}_n$ et $\hat{\theta}'_n$.

   (c) Au final, quelle valeur allez-vous retenir pour estimer θ ?

8. Donner une fonction pivotale pour θ. En déduire l'expression d'un intervalle de confiance bilatéral de seuil α pour θ. Pour les données de l'exemple, donner cet intervalle au seuil 5%.

9. Au seuil 5%, peut-on conclure que la durée moyenne entre deux défaillances successives est supérieure à 0.5 milliers heures ?