POLYTECH E2I4 année 2021-2022

Statistiques - Examen

Durée : 4h

Documents : Récupérer les documents sur Chamilo : l'énoncé, et les données des 4 exercices. Vous avez le droit aux documents de cours et aux TPs réalisés durant le cours. Durant l'épreuve il est interdit de communiquer et d'envoyer des mails.

Vous rédigerez vos réponses en utilisant un notebook (Jupyter). Faites attention à bien indiquer les numéros d'exercices, à répondre à toutes les questions, à détailler ce que vous faites, mais également à commenter les résultats. Veillez à préciser la signification de toutes les variables.

Exercice 1 :

Un fabricant de piles électriques affirme que la durée de vie moyenne des piles qu'il produit est de 170 heures. Un organisme de défense des consommateurs prélève au hasard un échantillon de $n = 100$ piles ; ces données se trouvent dans le fichier DureeVie_Piles.xlsx.

1/ Décrire les données : le nombre d'observation dans l'échantillon, la durée de vie moyenne obtenue sur l'échantillon et visualiser l'ensemble des valeurs sous la forme d'un histogramme.

2/ Donner une estimation ponctuelle de la durée de vie moyenne des piles produites par ce fabricant (justifier cette estimation ponctuelle) et donner l'intervalle de confiance à 98% de la durée de vie moyenne des piles.

3/ Proposer une méthode à l'organisme de défense des consommateurs pour tester l'affirmation du fabriquant, réaliser le test et formulez une conclusion.

Exercice 2 :

Un fabricant de vaccin réalise un essai clinique, en double aveugle, sur un échantillon de personnes âgées entre 25 et 30 ans. Cet échantillon comprend des hommes et des femmes (variable Genre). Les participants et les soignants ne connaissent pas le traitement administré (le traitement est soit un placebo soit la substance active). Les participants reçoivent alors le placebo ou la substance active et sont suivis sur plusieurs semaines, durant lesquelles les participants déclarent la maladie (Maladie = 1) ou non (Maladie = 0).

1/ Décriver les données en indiquant le nombre de participants à l'essai clinique, la répartition hommes/femmes, la répartition dans chaque bras de l'essai (placebo ou actif)

2/ Indiquer la proportion de participants en fonction de leur genre et du bras d'étude qui ont déclaré la maladie.

3/ Tester l'efficacité de la substance active sans tenir compte du Genre.

4/ Quelle conclusion faites-vous ?

Exercice 3 :

Les données de cet exercice se trouvent dans le fichier ProgType.xlsx. Chaque colonne, ProgType1, ProgType2, ProgType3, contient les temps d'exécution en millisecondes de 3 programmes d'analyse d'images. Ces 3 programmes réalisent les mêmes analyses avec des implémentations différentes. Nous souhaitons ici tester si l'une des implémentations est meilleure car plus rapide (en temps d'exécution).

1/ Tracer pour les comparer les histogrammes des mesures (temps d'exécution) pour les 3 programmes. Donner également les moyennes des temps d'exécution pour les 3 programmes et les variances.

2/ Quel test proposez-vous dans un premier temps pour comparer les moyennes des temps d'exécution des 3 programmes ? Réaliser ce test et donner une conclusion.

3/ Effectuer ensuite les tests de comparaison des moyennes des différents programmes 2 à 2.

Exercice 4:

Les données originales proviennent du site https://www.kaggle.com/uciml/glass. « This is a Glass Identification Data Set from UCI. It contains 10 attributes including the id of the sample. The response is glass type (Type : discrete 7 values). Les données ont été modifiées pour répondre au besoin de l'examen. Ici seules 7 variables seront analysées. Elles se trouvent dans le fichier GlassAnalysis.xlsx. La liste des variables :

- Id number of the sample
- RI: refractive index
- Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
- Al: Aluminum
- Si: Silicon
- Type : le type de glass

1/ Décrire les données en indiquant le nombre de variables et leur type (quantitatives, qualitatives, continues, discrètes etc.). En fonction du type de variable proposer des visualisations des données.

2/ Pour la phase d'analyse vous devez créer une nouvelle variable Type2 qui contiendra les valeurs 1, 2 ou 3. Ces valeurs seront définies à partir de la variable Type.

Type 2 = 1 quand Type = 1, Type 2 = 2 quand Type = 2 et Type 2 = 3 pour toutes les autres valeurs de Type.

3/ Analyser les différentes variables RI, Na, Al et Si en fonction de la variable Type 2, en testant si ces variables sont en moyenne différentes en fonction du Type de verre. Quelles conclusions faites-vous ?

Remarque : les commandes suivantes peuvent être utiles : X est la DataFrame qui contient les données.

```
X1 = X[ X['Type2'] == 1 ]
X2 = X[ X['Type2'] == 2 ]
X3 = X[ X['Type2'] == 3 ]
```