Analyse de données - TP 3

ISEP – 7 Mars 2022

Instructions : Préparez un rapport incluant le code source et vos résultats, et déposez-le sur Moodle. Pas plus de 2 personnes par groupe. N'oubliez pas de mettre les 2 noms sur le rapport, ou de faire 2 rendus.

Bibliothèques

Ce TP nécessite les bibliothèques suivantes : Numpy, Matplotlib, Seaborn, et Scipy :

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns; sns.set()
```

A Questions sur le cours

1. Rappelez la définition de ce qu'est une variable catégorielle.
2. Pourquoi ne peut-on pas utiliser les measures classiques d'analyse bivariées telles que la corrélation ou la détermination sur les données catégorielles ?
3. Expliquez avec vos propres mots ce qu'est un test d'hypothèses.
4. Rappelez quelles hypothèses sont testées par le test de χ^2, et expliquez comment interpréter le résultat.

B Test d'indépendance et variables catégorielles

Dans cet exercice, nous voulons étudier le résultat d'un sondage effectué auprès de 2440 étudiants afin d'essayer d'établir le lien entre les études qu'ils suivent et le milieu sociaux-culturel de leurs parents.

1. Ouvrez le jeu de données "stats_socio.csv" et décrivez les différentes variables et leurs valeurs à partir d'histogrammes :
   — Affichez les histogrammes des différentes catégories.
   — Affichez les histogrammes croisés des différentes catégories.
   — Commentez.
2. Créez le tableau de contingence avec la commande **crosstab()** de pandas pour avoir le tableau croisé correspondant à ces données. Affichez et commentez la répartition des données dans le tableau résultant et déduisez le nombre de degrés de liberté de ce problème.

3. Utilisez la commande **chi2_contingency(·)** de la bibliothèque scipy.stats sur votre tableau. A partir des résultats et éventuellement en calculant d'autres indices, que pouvez vous conclure sur la dépendance entre les 2 variables ?
4. En sachant que la population Française compte en moyenne 70 enfants de cadres pour 100 enfants d'ouvriers, cette étude est-elle représentative ? Si oui justifiez. Si non, expliquez pourquoi et d'où peuvent venir les biais.

C Étude de l'influence de l'haplogroupe I-M170 du chromosome Y sur la taille des hommes

Dans cet exercise, nous reprenons une partie des données d'une étude menée par des chercheurs Croate sur l'influence de certaines mutations du chromosome Y sur la taille des individus de sexe masculin. Ils se sont en particulier intéressé à l'haplogroupe I-M170, une variante du chromosome Y assez répandue dans les Balkans ainsi qu'en Scandinavie. Une partie de leur étude consistait à chercher un influence éventuelle de cette variante du chromosome sur la croissance et la taille finale des individus en étant porteurs.

Pour celà, ils disposent des données suivantes de 2005 sur différents pays Européens :

Pays	Taille moyenne (cm)	Fréquence de l'haplogroupe I-M170 (%)
Azerbaïdjan	173	3.5
Arménie	173	5
Malte	173	9
Portugal	173.8	8
Albanie	174	13.5
Chypre	174.7	7.5
Moldavie	174.8	28
Roumanie	174.9	27
Turquie	175	7
Bulgarie	175.25	26
Géorgie	175.75	4
Grèce	176	15
Italie	176.4	7.5
Ukraine	176.5	22.4
Espagne	177	10
Russie	177.2	17.5
Macédoine	177.25	24
Biélorussie	177.33	23.8
Royaume Uni	177.4	17
France	177.75	13
Suisse	178.33	17.6
Irlande	178.5	12.5
Pologne	178.5	17.5
Finlande	178.5	29
USA (blancs)	178.8	19
Slovaquie	179.2	28
Lituanie	179.25	12
Belgique	179.25	19.5
Kosovo	179.4	8
Autriche	179.5	24
Slovénie	179.8	28
Hongrie	179.9	27.5
Norvège	180	42
Lettonie	180.25	7
Allemagne	180.25	24
Danemark	180.33	38.5
Croatie	180.4	45
République Tchèque	180.9	18.5
Serbie	181.2	47.5
Suède	181.3	41.5
Estonie	181.4	16.5
Islande	181.8	34
Monténégro	182.9	37.5
Pays-Bas	183.75	33

A ces données, ils ont ajouté les informations suivantes sur quelques zones spécifiques de Croatie dans une région appelée "la côte des géants" :

1. Proposez une visualisation de ces données en utilisant les outils de votre choix (Python, R ou Excel par exemple). Commentez.

Ville/Région	Taille moyenne (cm)	Fréquence de l'haplogroupe I-M170 (%)
Zenica	181.35	54
Dubrovnik	183.75	63
Zadar	182.8	72.5
Herzegovina	183.33	71.5

2. A partir des outils de statistiques bivariés étudiés lors des 2 derniers cours, que pouvez-vous dire sur le l'influence éventuelle de la mutation I-M170 sur la taille des individus de sexe masculin ? Vous détaillerez votre raisonnement et vos calculs.