TP2: correctif
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

NB : tous les réponses ont été arrondis à deux chiffres après la virgule.

Réponses

1. On trouve les dimensions $(20\,000, 9)$. Le premier chiffre, 20 000, représente le nombre de lignes (le nombre d'individus dans l'échantillon). Le deuxième chiffre, 9, représente le nombre de colonnes (le nombre de variables).
2. Voir fichier `R`. `genhlth` est une variable qualitative (ordinale) avec valeurs `poor`, `fair`, `good`, `very good`, `excellent`.
3. Variable qualitative (ordinale) : `genhlth`. Variables qualitatives (nominales) : `exerany`, `hlthplan`, `smoke100`, `gender`. Variables quantitatives : `height`, `weight`, `wtdesire`, `age`.
4. Voir fichier `R`.
5. On voit apparaitre une densité en forme de cloche, même si elle ne ressemble pas exactement à une densité normale ou student t à cause d'une légère asymétrie (l'extremité droite est plus longue que l'extremité gauche). L'argument `breaks` donne le nombre de barres (c'est-à-dire le nombre de classes) à prendre lors du calcul de l'histogramme.
6. Grâce à la fonction `summary()`, on voit que l'IMC moyen est de 26.31 et qu'un répondant appartient au top 25% des répondants en terme de l'IMC à partir d'une valeur de 28.89 (troisième quartile). L'écart-type est de 5.21.
7. On est "confiants à 99.9%" que l'IMC moyen de la population des États-unis se trouve entre 26.19 et 26.43. Plus précisément, si on sélectionne des échantillons de taille $n = 20\,000$ de la population et que l'on calcule à chaque fois l'intervalle de confiance, l'IMC moyen de la population sera comprise en moyenne 99.9% des fois dans l'intervalle calculé. Le choix de la distribution t n'est pas du tout important ici car n est très grand, et les quantiles des deux distributions sont quasiment identiques.

   On peut aussi remarquer que l'intervalle de confiance est très étroit. Ceci est également du à la très grande valeur de n, qui nous permet de donner un résultat très précis concernant l'IMC moyen.
8. Voir fichier `R`.
9. On veut effectuer le test d'hypothèse

$$H_0 : \mu = 25, \qquad H_1 : \mu > 25$$

pour $\alpha = 0.0001$. La région de rejet est

$$\left\{\bar{x} \geq 25 + \frac{5.210655}{\sqrt{20000}} 3.719706\right\} = \{\bar{x} \geq 25.14\}$$

On a trouvé un IMC moyen de 26.31. Cette valeur se trouvant dans la région de rejet, on peut rejeter l'hypothèse nulle (ou accepter l'hypothèse alternative) et conclure que l'IMC moyen de cette population dépasse la valeur de 25 avec un niveau de confiance de 99.99%.

10. 4657 répondants indiquent être en excellente santé. 3.39% des répondants indiquent être en mauvaise santé.
11. 10 559 répondants ont fumé moins que 100 cigarettes au cours de leur vie, tandis que 9441 répondants ont fumé au moins 100 cigarettes au cours de leur vie. On ne peut pas faire un histogramme d'une variable qualitative
12. 6012 femmes et 4547 hommes ont fumé moins que 100 cigarettes au cours de leur vie, tandis que 4419 femmes et 5022 hommes ont fumé au moins 100 cigarettes au cours de leur vie.
13. Le premier graphique met en relation le genre et le fait de fumer plus de 100 cigarettes. Nous pouvons constater que la majorité des personnes ayant fumé moins de 100 cigarettes dans leur vie sont des femmes alors que les hommes représentent la majorité des personnes ayant fumé plus de 100 cigarettes.

    Le second graphique contenant les variables `smoke100` et `genhlth` nous montre que la majorité des personnes ayant un état de santé général `poor` ont fumé plus de 100 cigarettes dans leur vie. Ce pourcentage de personnes ayant fumé plus de 100 cigarettes dans leur vie diminue pour chaque amélioration de niveau de santé générale.

    Le troisième graphique nous donne la répartition de l'IMC des répondants par rapport à leur état de santé général sous forme de boites à moustache. On peut y observer une diminution au niveau de la médiane de l'IMC pour chaque niveau supérieur de santé générale (et pareil pour le premier et le troisième quartile). On voit également une diminution de variabilité de l'IMC au fur et à mesure que le niveau de santé général augmente.

    Enfin, le dernier graphique représente la relation entre le poids effectif et le poids désiré sous forme d'un nuage de points. Nous pouvons observer une relation positive entre les deux variables. Cela voudrait dire qu'au plus une personne a un poids effectif élevé, au plus son poids désiré sera élevé. A l'inverse, au plus le poids effectif d'une personne est faible, au plus son poids désiré sera faible. Pour la majorité des individus, le poids désiré est inférieur ou égal à leur poids effectif.