Thème Statistique

Exercice 1 (Questions de cours (3 points)).

1. (0,5) fonction de répartition, empirique ou théorique, variable discrète. Les valeurs en ordonnées vont de 0 à 1 (0,5).

2. (0,5) histogramme, empirique, variable continue.

3. (0,5) fonction de répartition, théorique, variable continue. Les valeurs en ordonnées vont de 0 à 1 (0,5).

4. (0,5) densité, théorique, variable continue.

Exercice 2 (Table de distribution (4 points)).

1. (0,5) Les deux variables sont qualitatives.

2. (1,5)

Genre \ Classe	1	2	3	Total
Femme	133	103	152	388
Homme	151	158	349	658
Total	284	261	501	1046

3. (0,5) La taille d'échantillon est $n = 1046$. C'est la valeur en bas à droite du tableau (c'est la somme des effectifs du tableau de contingence).

4. (a) (0,5) Elle vaut $151/1046 \approx 14.4\%$.
   (b) (0,5) Elle vaut $151/658 \approx 22.9\%$.
   (c) (0,5) Elle vaut $151/284 \approx 53.2\%$.

Exercice 3 (Mention au BAC (5 points)).

1. (1) La taille de l'échantillon est $n = 292 + 608 = 900$.

2. (1) Un estimateur sans biais d'une probabilité est la fréquence empirique. On en déduit l'estimation $f = 292/900 \approx 32.4\%$.

3. (2) On estime une proportion. On utilise donc l'intervalle de confiance asymptotique :

$$\left[F - u_{1-\alpha/2}\frac{\sqrt{F(1-F)}}{\sqrt{n}}, F + u_{1-\alpha/2}\frac{\sqrt{F(1-F)}}{\sqrt{n}}\right],$$

   où F est la fréquence empirique et $u_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale $\mathcal{N}(0, 1)$. Ce choix est justifié car l'échantillon est grand : $n = 900 \geq 100$.

4. (0,5) L'intervalle est $[0.294, 0.356]$, donné par la commande `prop.test(292,900)`.

5. (0,5) L'intervalle contient la valeur $0.3 = 30\%$. On ne peut donc pas dire que la probabilité de mention pour les grenoblois est différente de la valeur nationale.

6. (0,5) Non, comme on l'a vu en TP.

Exercice 4 (Soda et adolescents (8 points)).

1. (1) Moyenne empirique : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Estimation : $\bar{x} = 10.4$ donné par la fonction `mean`.

2. (1) Variance corrigée : $S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Estimation : $s'^2 = 4.27$ donné par la fonction `var`.

3. (2) On estime la moyenne d'un échantillon non gaussien. On utilise l'intervalle de confiance asymptotique :
$$\left[\bar{X} - u_{1-\alpha/2}\frac{S'}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{S'}{\sqrt{n}}\right],$$
où $\bar{X}$ est la moyenne empirique, S' est l'écart-type corrigé et $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$. Ce choix est justifié car l'échantillon est grand : $n = 100 \geq 100$.

4. (1) Le calcul donne :
$$\left[10.4 - 1.96\frac{2.07}{10}, 10.4 + 1.96\frac{2.07}{10}\right] = [9.99, 10.81].$$

5. (0,5) Notons μ_1 la consommation moyenne des adolescents et μ_2 la consommation moyenne des adultes. Il faut poser $\mathcal{H}_0 : \mu_1 = \mu_2$ et $\mathcal{H}_1 : \mu_1 \neq \mu_2$.

6. (1) Comme les échantillons sont appariés (mêmes individus dix ans après), nous allons considérer l'échantillon des différences stockés dans `X1-X2`. Sur cet échantillon, nous allons utiliser un test de conformité de la moyenne de l'échantillon des différences $\mu_D = \mu_1 - \mu_2$ par rapport à la référence $\mu_0 = 0$. La statistique est
$$T = \sqrt{n}\frac{\bar{X}_D - 0}{S'_D},$$
et la règle de décision (au niveau de risque α) : rejet si et seulement si $|t_{obs}| > u_{1-\alpha/2}$.

7. (0,5) Comme nous n'avons pas fait d'hypothèse de loi normale, il faut un grand échantillon. Ce qui est vérifié ici.

8. (1) D'après les valeurs de `R`, nous avons $\bar{x}_D = 0.3$ et $s'_D = 0.965$ donc la statistique vaut
$$t = \sqrt{100}\frac{0.3}{0.965} \approx 3.11.$$
D'après l'énoncé, `pnorm(3.11)`$= 0.999$ donc on trouve la p-valeur $\alpha^* = 2(1 - 0.999) = 0.002$. Cette p-valeur est petite (plus petite que 5%) donc on rejette l'hypothèse nulle et on conclut que la consommation moyenne est différente entre les adolescents et les adultes.