Corrigé de l'examen du 11/01/2016

Exercice 1

1. Notons $p \in [0,1]$ la proportion d'électeurs qui voteront pour C. L'hypothèse nulle est $H_0 =$ "$p \leqslant \frac{1}{2}$", et l'hypothèse alternative est $H_1 =$ "$p > \frac{1}{2}$". En effet, avec ce choix, l'erreur de première espèce est $\mathbb{P}_{H_0}(\text{rejeter } H_0)$, c'est donc la probabilité de déclarer C vainqueur alors qu'il ne l'est pas. D'après la consigne, c'est bien l'erreur que l'on veut contrôler. Au contraire, l'erreur de seconde espèce, $\mathbb{P}_{H_1}(\text{garder } H_0)$, est la probabilité de ne pas déclarer C vainqueur alors qu'il l'est. On ne demande pas de contrôle sur cette erreur.
2. Notons $X_i = 1$ si la $i^{\text{ième}}$ personne a déclaré voter pour C, et $X_i = 0$ sinon, pour $i \in [\![1, N]\!]$, $N = 1000$. Avec cette notation,
$$\hat{p} \ := \ \frac{1}{N}\sum_{i=1}^{N} X_i$$
est un estimateur sans biais de p. Plus précisément, d'après le Théorème Central Limite, $\hat{p}$ suit approximativement une loi normale $\mathcal{N}\left(p, \frac{p(1-p)}{N}\right)$. On veut tester le fait que la proportion p est petite, autrement dit la zone de rejet R_α de niveau $\alpha = 5\%$ va être de la forme $\{\hat{p} > s\}$, pour un seuil s à déterminer. En effet, on rejettera l'hypothèse selon laquelle la proportion d'électeurs pour C est plus petite que 1/2 si on observe, sur un échantillon, une proportion suffisamment grande. Le seuil s doit être déterminé de sorte à ce que
$$\mathbb{P}_{H_0}\left(\hat{p} > s\right) \ \leqslant \ \alpha.$$
Or,
$$\mathbb{P}_{H_0}\left(\hat{p} > s\right) \ = \ \mathbb{P}_{H_0}\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/N}} > \frac{s-p}{\sqrt{p(1-p)/N}}\right) \ = \ \mathbb{P}_{H_0}\left(G > \frac{s-p}{\sqrt{p(1-p)/N}}\right),$$
où G est une variable normale centrée réduite. Sous H_0, $p \leqslant \frac{1}{2}$, de sorte que
$$\mathbb{P}_{H_0}\left(G > \frac{s-p}{\sqrt{p(1-p)/N}}\right) \ \leqslant \ \mathbb{P}\left(G > \frac{s-\frac{1}{2}}{\sqrt{1/(4N)}}\right).$$
On cherche donc s tel que
$$\mathbb{P}\left(G > \frac{s-\frac{1}{2}}{\sqrt{1/(4N)}}\right) \ \leqslant \ \alpha = 5\%,$$
ce qui d'après la table de la loi gaussienne donne
$$\frac{s-\frac{1}{2}}{\sqrt{1/(4N)}} \ \simeq \ 1.65$$
et donc $s = \frac{1}{2} + \frac{1.65}{2\sqrt{N}} \simeq 52.6\%$. Autrement dit, si la vraie proportion p de partisans de C dans la population totale est inférieure à 1/2, alors la probabilité d'observer une proportion supérieure à $52,6\%$ sur un échantillon de 1000 personnes est inférieure à 5%. Le test est donc le suivant : si la proportion $\hat{p}$ observée sur l'échantillon est supérieure à 52.6%, on rejette H_0, c'est-à-dire qu'on prévoit C vainqueur. Sinon, on ne déclare pas C vainqueur. Or, dans l'exemple présent, la proportion observée sur l'échantillon est $\hat{p} = 51.6\%$. On ne rejette donc pas H_0, et on ne déclare pas C vainqueur.

3. La puissance est $\mathbb{P}_{H_1}(\text{rejette } H_0)$. Ici, c'est donc la probabilité, C étant vainqueur, que le sondage permette de conclure que C sera vainqueur.
4. On l'a vu dans la question 2, le sondage permet de déclarer C vainqueur (avec 5% de chances d'erreur) si la proportion $\hat{p}$ observée est supérieure à 52.6%. Or, $\hat{p}$ est une variable aléatoire de loi normale $\mathcal{N}\left(p, \frac{p(1-p)}{N}\right)$, où dans cette question on suppose connaître $p =$ 51.5%. Dans ce cas,

$$\begin{aligned}\mathbb{P}\,(\text{C déclaré vainqueur}) &= \mathbb{P}\,(\hat{p} > 0.526)\\ &= \mathbb{P}\left(G > \frac{0.526 - 0.515}{\sqrt{0.515(1-0.515)/1000}}\right)\\ &\simeq \mathbb{P}\,(G > 0.7)\\ &\simeq 24\%.\end{aligned}$$

Si la proportion réelle est $p = 51.5\%$, alors on a moins d'une chance sur 4 d'avoir la bonne conclusion. L'échantillon est trop petit.
5. Notons à nouveau

$$\hat{p} \;:=\; \frac{1}{N}\sum_{i=1}^{N} X_i,$$

mais avec cette fois-ci N à determiner. En reprenant la question 2, on obtient une zone de rejet à 5% de la forme $\{\hat{p} > s\}$ avec un seuil s tel que

$$\frac{s - \frac{1}{2}}{\sqrt{1/(4N)}} \;\simeq\; 1.65,$$

et donc $s = \frac{1}{2} + \frac{1.65}{2\sqrt{N}}$. En supposant que la vraie proportion est $p = 51.5$, la probabilité de ne pas se tromper est donc

$$\begin{aligned}\mathbb{P}\,(\text{C déclaré vainqueur}) &= \mathbb{P}\left(\hat{p} > \frac{1}{2} + \frac{1.65}{2\sqrt{N}}\right)\\ &= \mathbb{P}\left(G > \frac{\frac{1}{2} + \frac{1.65}{2\sqrt{N}} - 0.515}{\sqrt{0.515(1-0.515)/N}}\right).\end{aligned}$$

On veut que cette probabilité soit supérieure à 95%, et donc

$$\begin{aligned}95\% &\leqslant \mathbb{P}\left(G > \frac{\frac{1.65}{2\sqrt{N}} - 0.015}{\sqrt{0.515(1-0.515)/N}}\right)\\ &= \mathbb{P}\left(G < -\frac{\frac{1.65}{2\sqrt{N}} - 0.015}{\sqrt{0.515(1-0.515)/N}}\right),\end{aligned}$$

qui donne

$$-\frac{\frac{1.65}{2\sqrt{N}} - 0.015}{\sqrt{0.515(1-0.515)/N}} \simeq 1,65$$

et donc

$$N = \left(\frac{1.65}{0.015}\left(\frac{1}{2} + \sqrt{0.515(1-0.515)}\right)\right)^2 \simeq 12\ 100.$$

Exercice 2

1. Notons respectivement μ et σ la moyenne et l'écart-type de la masse de raisin par souche. Dans ce cas, $\bar{x}$ et

$$\bar{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

sont respectivement des estimateurs sans biais de μ et σ^2. Plus précisément, d'après le TCL, $\bar{x}$ suit approximativement la loi $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Ainsi, n'importe quel seuil $h > 0$,

$$\begin{aligned}\mathbb{P}\left(|\bar{x} - \mu| \leqslant h\right) &= \mathbb{P}\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \leqslant \frac{h}{\sigma/\sqrt{n}}\right) \\ &\simeq \mathbb{P}\left(|G| \leqslant \frac{h}{\sigma/\sqrt{n}}\right) \\ &\simeq \mathbb{P}\left(|G| \leqslant \frac{h}{\bar{\sigma}/\sqrt{n}}\right),\end{aligned}$$

où G est une variable gaussienne centrée réduite.

On cherche un seuil h tel que $\mathbb{P}\left(|\bar{x} - \mu| \leqslant h\right) \geqslant 95\%$, autrement dit

$$95\% \simeq \mathbb{P}\left(|G| \leqslant \frac{h}{\bar{\sigma}/\sqrt{n}}\right) = 2\left(\mathbb{P}\left(G \leqslant \frac{h}{\bar{\sigma}/\sqrt{n}}\right) - \frac{1}{2}\right),$$

c'est-à-dire $\mathbb{P}\left(G \leqslant \frac{h}{\bar{\sigma}/\sqrt{n}}\right) \simeq \frac{195}{200}$, qu'on résout par $\frac{h}{\bar{\sigma}/\sqrt{n}} \simeq 1.96$ et $h = 1.96\sqrt{\frac{2.14}{40}} \simeq 0.45$. Conclusion, un intervalle de confiance à 95% pour μ est $[\bar{x} - 0.45, \bar{x} + 0.45] = [4.25, 5.15]$.

2. D'après la question précédente, on estime que $\bar{x} \simeq \mu \pm 0.45$ avec une probabilité 95%. Autrement dit, la marge d'erreur à 95% est 0.45.

3. D'après la question 1, la marge d'erreur h est telle que $\frac{h}{\bar{\sigma}/\sqrt{n}} \simeq 1.96$. Si l'on veut $h = 0.3$, cela donne $n = 2.14\,(1.96/0.3)^2 \simeq 91$.

4. L'hypothèse nulle est $H_0 =$ "$\mu = 5$" et l'hypothèse alternative $H_1 =$ "$\mu \neq 5$". On rejettera H_0 si le poids moyen observé sur un échantillon est trop loin de 5kg, autrement dit la zone de rejet est de la forme $\{|\bar{x} - 5| > s\}$ pour un seuil s à déterminer. Sous l'hypothèse H_0, $\bar{x}$ suit approximativement la loi $\mathcal{N}(5, \bar{\sigma})$, de sorte que

$$\mathbb{P}_{H_0}\left(|\bar{x} - 5| > s\right) \simeq \mathbb{P}\left(|G| > \frac{s}{\bar{\sigma}\sqrt{n}}\right)$$

où G est une variable gaussienne centrée réduite. On cherche s tel que cette probabilité soit plus petite que 5%, autrement dit

$$\mathbb{P}\left(G < \frac{s}{\bar{\sigma}\sqrt{n}}\right) = 0.975,$$

de sorte que $\frac{s}{\bar{\sigma}\sqrt{n}} = 1.96$ et $s = \frac{1.96 \times 2.14}{\sqrt{40}} \simeq 0.66$. Autrement dit, on rejette H_0 si l'on observe $|\bar{x} - 5| > 0.66$. Or, on observe $\bar{x} = 4.7$, donc on ne rejette pas H_0.