Corrigé 12

Exercice 1. On a observé les réponses de 800 personnes choisies au hasard. Pour la $i$ème personne, considérons une variable aléatoire

$$X_i = \begin{cases} 1 & \text{si la personne choisie est en faveur de l'indépendance,} \\ 0 & \text{sinon.} \end{cases}$$

Les variables $X_1, \dots, X_{800}$ sont iid Bernoulli(p), où $p \in (0,1)$ est le vrai pourcentage des personnes en faveur de l'indépendance.

Il faut tester $H_0 : p \leq 0.5$ contre $H_1 : p > 0.5$. On peut baser le test sur la différence entre l'estimateur $\bar{X}_{800}$ du paramètre p et sa valeur hypothétique 0.5. Le test rejette H_0 si la variable $\bar{X}_{800} - 0.5$ est significativement grande. Plus précisément, on rejette H_0 quand $\bar{X}_{800} - 0.5 > c$, où c est une constante telle que la probabilité de rejet de H_0 quand H_0 est vérifiée est α (le niveau de signification choisi pour le test). Ainsi, c doit vérifier

$$\max_{p \in [0,0.5]} \Pr_p(\bar{X}_{800} - 0.5 > c) = \alpha. \tag{1}$$

Le théorème central limite nous donne que $\bar{X}_{800} - 0.5$ suit approximativement la loi normale $\mathcal{N}\left(p - 0.5, \frac{p(1-p)}{n}\right)$. Etant donné que les fonctions $p \mapsto p - 0.5$ et $p \mapsto \frac{p(1-p)}{n}$ sont strictement croissantes sur l'intervalle $[0, 0.5]$, on peut montrer que le maximum apparaissant en (1) est atteint pour $p = 0.5$.

On cherche donc une valeur d telle que

$$\Pr_{p=0.5}(T > d) = \alpha, \tag{2}$$

où

$$T = \frac{\bar{X}_{800} - 0.5}{\sqrt{0.5 \times (1 - 0.5)/800}}.$$

D'après le théorème central limite, on peut approximer la distribution de T par la loi normale $\mathcal{N}(0,1)$. Ainsi, (2) se réécrit

$$1 - \Phi(d) = \alpha,$$

ce qui donne $d = \Phi^{-1}(1-\alpha)$. Donc le test rejette H_0 en faveur de H_1 quand $T > \Phi^{-1}(1-\alpha)$. Notons que le niveau de ce test est approximativement α et non exactement α du fait de l'approximation résultant du théorème central limite.

Dans notre cas, $\bar{x}_{800} = 0.55$, donc

$$t_{obs} = \frac{0.55 - 0.5}{\sqrt{0.5 \cdot 0.5/800}} = 2.83.$$

Si l'on choisit $\alpha = 0.05$, on a $\Phi^{-1}(1-\alpha) = 1.64$. Comme $t_{obs} > 1.64$, on rejette l'hypothèse H_0 en faveur de H_1. On peut dire que l'on a montré, à un niveau de signification de 5 %, que la majorité de la population est favorable à l'indépendance du Québec.

Exercice 2. Dénotons p_I la probabilité de naître au premier trimestre, p_{II} la probabilité de naître au deuxième trimestre, p_{III} la probabilité de naître au troisième trimestre et p_{IV} la probabilité de naître au quatrième trimestre.

(i). On a

$$H_0 : p_I = 2 \times p_{II},\ p_{II} = p_{III} = p_{IV}.$$

Puisque $p_I + p_{II} + p_{III} + p_{IV} = 1$, on a que $H_0 :\ p_I = 0.4,\ p_{II} = p_{III} = p_{IV} = 0.2$. On teste l'adéquation de cette distribution à l'aide du test chi-deux. Le tableau des nombres observés ($np_i = 300p_i$) et attendus est :

Trimestre	Janv-Mars	Avr-Juin	Juil-Sept	Oct-Déc	Total
Nombre observé (o_j)	110	57	53	80	300
Nombre attendu (e_j)	120	60	60	60	300

La statistique à utiliser est

$$T = \sum_{j=1}^{4} \frac{(O_j - E_j)^2}{E_j}$$

qui, sous H_0, suit approximativement la loi χ^2_ν avec $\nu = 4 - 1 - 0 = 3$. On calcule la valeur observée de T, $t_{obs} \approx 8.47$ qui est une valeur plus petite que le quantile de χ^2_3 au niveau 0.99, $\chi^2_{3,0.99} \approx 11.34$ (le quantile peut être lu dans la table *quantiles de la loi Khi-deux* sur Moodle). Comme $t_{obs} < \chi^2_{3,0.99}$, on ne rejette pas l'hypothèse nulle.

(ii). Maintenant on doit tester

$$H_0 :\ \ p_I = p_{IV},\ p_{II} = p_{III}.$$

La différence avec la partie précédante est qu'ici nous n'avons pas de nombres concrets pour les proportions attendues sous H_0. Il faut donc les estimer. Avant de le faire, on réfléchit au nombre minimal de paramètres à estimer. En fait, il suffit d'estimer p_I car sous H_0 on a $p_{IV} = p_I$ et $p_{II} = p_{III} = (1 - 2\,p_I)/2$.
Sous H_0, on estime p_I par $\widehat{p}_I = (o_1/n + o_4/n)/2 = (o_1 + o_4)/(2\sum_{i=1}^4 o_i) = (110 + 80)/600 = 95/300$. Ceci est plus raisonnable que de prendre juste o_1/n car cet estimateur n'utilise pas l'information que $p_I = p_{IV}$. (Il est possible de montrer que $\widehat{p}_I$ est l'estimateur du maximum de vraisemblance.)
En utilisant $\widehat{p}_{IV} = \widehat{p}_I$ et $\widehat{p}_{II} = \widehat{p}_{III} = (1 - 2\,\widehat{p}_I)/2$, on obtient les nombres attendus estimés :

Trimestre	Janv-Mars	Avr-Juin	Juil-Sept	Oct-Déc	Total
Nombre observé (o_j)	110	57	53	80	300
Nombre attendu estimé (e_j)	95	55	55	95	300

On utilise la statistique de test

$$T = \sum_{j=1}^{4} \frac{(O_j - E_j)^2}{E_j},$$

qui, sous H_0, suit la loi χ^2_ν avec $\nu = 4 - 1 - 1 = 2$ (le degré de liberté change car on a estimé un paramètre). La réalisation de la statistique T est $t_{obs} \approx 4.88$. Celle-ci est plus petite que le quantile $\chi^2_{2,0.99} \approx 9.21$ donc on ne rejette pas H_0.

Remarque : Aucune des deux hypothèses nulles n'a été rejettée alors que les deux sont incompatibles, donc cela peut paraître étrange. Mais rappelons-nous que "ne pas rejeter" n'est pas "accepter".

Exercice 3. Cette situation peut au premier abord paraître très différente de ce que l'on a fait dans les exercices précédents. Mais en fait, elle est similaire à la partie (2) de l'exercice précédant. On a

$$\begin{aligned} H_0 :&\quad \text{les données viennent d'une loi normale,} \\ H_1 :&\quad H_0 \text{ n'est pas vraie.} \end{aligned}$$

On considère le nombre d'observations dans certains intervalles. Sous H_0, la probabilité que le taux d'oxygénation soit dans un intervalle (a, b) est $F(b) - F(a)$, où F est la fonction de répartition d'une loi normale. Une fois les paramètres de la loi normale connus, on peut calculer les nombres attendus estimés sous H_0 dans tous les intervalles. On estime les paramètres de la loi normale par $\hat{\mu} = \bar{x}$ et $\hat{\sigma}^2 = s_x^2$.

Pour calculer le nombre attendu estimé sous H_0 dans l'intervalle $(0.1, 0.15]$ par exemple, on procède comme suit. On considère une variable aléatoire $X \sim \mathcal{N}(0.173, 0.066^2)$, et on calcule

$$e_2 = 83 \times P(0.1 < X \leq 0.15) = 83 \times P\left(\frac{0.1 - 0.173}{0.066} < \frac{X - 0.173}{0.066} \leq \frac{0.15 - 0.173}{0.066}\right) =$$
$$= 83\times(\Phi(-0.348)-\Phi(-1.106)) = 83\times(1-\Phi(0.348)-1+\Phi(1.106)) = 83\times(\Phi(1.106)-\Phi(0.348)) = 19.039.$$

De cette manière, on obtient le tableau suivant pour les nombres observés et théoriques.

	o_j	e_j
≤ 0.1	12	11.151
$(0.1,\ 0.15]$	20	19.039
$(0.15,\ 0.20]$	23	24.487
$(0.20,\ 0.25]$	15	18.224
> 0.25	13	10.099

Maintenant on peut utiliser la statistique

$$T = \sum_{j=1}^{5} \frac{(O_j - E_j)^2}{E_j},$$

qui, sous H_0, suit la loi χ^2_ν avec $\nu = 5 - 1 - 2 = 2$ (il y a 2 paramètres estimés : $\hat{\mu}$ et $\hat{\sigma}^2$). On constate que $t_{obs} = 1.607 < \chi^2_2(0.95) = 5.99$, donc on ne rejette pas l'hypothèse nulle.

Exercice 4. Il s'agit d'un test d'indépendance de deux caractéristiques de données. On peut de nouveau baser la statistique de test sur les différences entre les nombres observés et attendus.

Si on a n_1 classes pour la première caractéristique et n_2 classes pour la deuxième caractéristique, et si on note N_{ij} et E_{ij} les nombres observés et attendus d'observations de la $i$ème classe de la première caractéristique et $j$ème classe de la deuxième caractéristique, la statistique de test à utiliser est

$$T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}.$$

On estime les nombres attendus par

$$e_{ij} = n \times \frac{\sum_{i=1}^{n_1} n_{ij} \times \sum_{j=1}^{n_2} n_{ij}}{\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} n_{ij}\right)^2}.$$

Cela vient du fait que sous H_0 on a $p_{ij} = p_i \times p_j$, où p_i est la probabilité d'être dans la $i$ème classe de la première caractéristique, p_j est la probabilité d'être dans la $j$ème classe de la deuxième caractéristique et p_{ij} est la probabilité d'être dans la $i$ème classe de la première caractéristique et dans la $j$ème classe de la deuxième caractéristique. On estime p_i par $(\sum_{j=1}^{n_2} n_{ij})/(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} n_{ij})$ et p_j par $(\sum_{i=1}^{n_1} n_{ij})/(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} n_{ij})$.

La distribution asymptotique de la statistique T sous H_0 est χ^2_ν avec $\nu = (n_1-1)\times(n_2-1)$. On peut obtenir ν, le nombre de degrés de liberté, comme suit. Le nombre total de classes est $n_1 \times n_2$. Le nombre de paramètres à estimer est $(n_1-1)+(n_2-1)$ (on a n_1-1 estimateurs pour p_i et n_2-1 estimateurs pour p_j). Enfin, $n_1 \times n_2 - 1 - n_1 - n_2 + 2 = (n_1-1)\times(n_2-1)$. Dans notre cas, nous reprenons le tableau des données dans lequel nous introduisons entre parenthèses les nombres attendus estimés sous H_0 :

	L1	L2	L3	Total
T1	50 (53.84)	16 (20.37)	31 (22.80)	97
T2	61 (57.17)	26 (21.63)	16 (24.21)	103
Total	111	42	47	200

La valeur observée de la statistique

$$T = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

est $t_{obs} \approx 8.08 > \chi^2_{2,0.95} \approx 5.99$, donc, au niveau de 5 %, on a montré qu'il y a une dépendance entre le type et la localisation du défaut. Notons que l'approximation par la loi χ^2 est possible, car le nombre d'observations est grand et tous les nombres attendus e_{ij} sont plus grands que 5.

Exercice 5. (i). $N(0,1)$.

(ii). On rejette quand T est grand, c'est-à-dire quand $T > c$ pour une constante c telle que

$$\alpha = \mathrm{Pr}_{H_0}(T > c) = 1 - \mathrm{Pr}_{H_0}(T \le c) = 1 - \Phi(c)$$

où Φ est la fonction de répartition d'une loi normale standard. La valeur critique est donc $c = \Phi^{-1}(1-\alpha) = z_{1-\alpha}$.

(iii). On sait que $W_n \sim \mathbb{N}(0,1)$. On va donc faire des manipulation algébriques pour insérer W_n dans les calculs de la probabilité (où $\mu > \mu_0$)

$$\beta(\alpha,\mu,n) = \mathrm{Pr}_\mu(T_n > z_{1-\alpha}) = \mathrm{Pr}_\mu\left(\sqrt{n}\frac{\overline{Y}_n}{\sigma} > z_{1-\alpha} + \frac{\mu_0}{\sigma}\right) = \mathrm{Pr}_\mu\left(\sqrt{n}\frac{(\overline{Y}_n - \mu)}{\sigma} > z_{1-\alpha} + \sqrt{n}\frac{\mu_0}{\sigma} - \sqrt{n}\right.$$

$$= \mathrm{Pr}_\mu(W_n > z_{1-\alpha} - \sqrt{n}(\mu-\mu_0)/\sigma) = 1 - \Phi\left(z_{1-\alpha} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\right).$$

Comme Φ est croissante, la puissance augmente lorsque μ augmente. Donc, plus la vérité est loin de H_0, plus facile il va être le détecter en rejettant H_0. Comme $\mu - \mu_0 > 0$, la puissance augmente également avec n. Donc, plus l'échantillon qu'on a à disposition est grand, plus on va pouvoir détecter des déviations de H_0. Quant à α, plus il est petit, plus $z_{1-\alpha}$ sera grand et la puissance sera petite. Donc la puissance augmente avec α : plus on prend de risque à rejeter H_0 à tort (erreur de type I), plus on aura de la chance à la rejeter à juste titre (puissance). Finalement, la puissance baisse avec σ, car une grande variance des données implique plus d'incertitude : si $\overline{X}_n > \mu_0$, cela pourrait être dû au hasard et non pas parce que la vraie espérance μ est plus grande que μ_0.