# Corrigé 12

**Exercice 1.** On va tester

$H_0$ : l'affirmation du responsable de la communication est vraie contre
$H_1$ : l'affirmation du responsable de la communication est fausse.

En fait, si on dénote par $p_B$ la proportion de brun, par $p_J$ la proportion de jaune, par $p_R$ la proportion de rouge, par $p_O$ la proportion d'orange, par $p_V$ la proportion de vert, et par $p_D$ la proportion de doré, on a

$$H_0 : \quad p_B = 0.3,\ p_J = 0.2,\ p_R = 0.2,\ p_O = 0.1,\ p_V = 0.1,\ p_D = 0.1,$$
$$H_1 : \quad H_0 \text{ n'est pas vraie.}$$

Il s'agit d'un test d'adéquation. On peut baser la statistique de test sur les différences entre les nombres de bonbons des différentes couleurs observés ($O_i$) et les nombres qu'on attend si $H_0$ est vraie ($E_i$) :

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

où $n$ est le nombre de classes. Si $H_0$ est vraie et si le nombre d'observations est assez grand, la statistique $T$ suit approximativement une loi $\chi^2$ dont le nombre de degrés de liberté est égal à

$$(\text{nombre de classes}) - 1 - (\text{nombre de paramètres estimés sous } H_0).$$

Dans notre cas, les nombres observés et attendus sont :

| | Nombre observé ($o_i$) | Nombre attendu ($e_i$) |
|---|---|---|
| Bleu | 84 | $0.3 \times 370 = 111$ |
| Jaune | 79 | $0.2 \times 370 = 74$ |
| Rouge | 75 | $0.2 \times 370 = 74$ |
| Orange | 49 | $0.1 \times 370 = 37$ |
| Vert | 36 | $0.1 \times 370 = 37$ |
| Doré | 47 | $0.1 \times 370 = 37$ |

La taille de l'échantillon est grand et tous les nombres attendus $e_i$ sont plus grands que 5. On peut donc approximer la loi de la statistique $T$ sous $H_0$ par la loi asymptotique. Nous avons 6 classes et aucun paramètre à estimer sous $H_0$ (les proportions sous $H_0$ sont données). Cela donne 5 degrés de liberté pour la loi asymptotique.

Il reste à calculer la valeur observée de la statistique de test et la comparer avec une valeur critique. Testons à un niveau de 5 % (ce qui est le niveau standard). La valeur critique est $\chi^2_5(0.95) = 11.1$ (on peut la trouver dans le tableau de la loi $\chi^2$). La valeur observée de la statistique de test est

$$t_{obs} = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 13.54.$$

Puisque cette valeur est plus grande que la valeur critique, on rejette $H_0$ en faveur de $H_1$, et on peut dire que, à un niveau de signification de 5 %, on a montré que l'affirmation du responsable de communication est fausse.

**Exercice 2.** (a) On va tester

$H_0$ : la natalité du 1er trimestre est deux fois plus élevé que celle des autres trimestres contre
$H_1$ : $H_0$ est fausse.

En fait, si on dénote par $p_I$ la probabilité d'être né dans le premier trimestre, par $p_{II}$ la probabilité d'être né dans le deuxième trimestre, par $p_{III}$ la probabilité d'être né dans le troisième trimestre, et par $p_{IV}$ la probabilité d'être né dans le quatrième trimestre, on a

$$\begin{aligned} H_0 &: \ p_I = 2 \times p_{II},\ p_{II} = p_{III} = p_{IV}, \\ H_1 &: \ H_0 \text{ n'est pas vraie.} \end{aligned}$$

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 est donc dans la même situation que dans l'exercice précédent et on peut procéder de la même façon.
Le tableau des fréquences observées et attendues 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

$$\sum_{j=1}^{4} \frac{(o_j - e_j)^2}{e_j}$$

qui, sous $H_0$, suit la loi $\chi^2_\nu$ avec $\nu = 4 - 1 - 0 = 3$. On trouve $t_{obs} = 8.47$ qui est une valeur plus petite que $\chi^2_3(0.99) = 11.34$. On ne rejette donc pas l'hypothèse nulle.

(b) On va tester

$$\begin{aligned} H_0 &: \ p_I = p_{IV},\ p_{II} = p_{III} \text{ contre} \\ H_1 &: \ H_0 \text{ n'est pas vraie.} \end{aligned}$$

La différence entre cette situation et celle de la partie (a) est que nous n'avons pas de nombres concrets pour les proportions sous $H_0$, il faut donc les estimer. Avant de le faire on réfléchit sur le nombre minimal de paramètres à estimer. En fait, ici il suffit d'estimer $p_I$, parce que sous $H_0$ on a $p_{IV} = p_I$ et $p_{II} = p_{III} = (1 - 2\,p_I)/2$.
On estime $p_I$ par $\hat{p}_I = (o_1/\sum_{i=1}^4 o_i + o_4/\sum_{i=1}^4 o_i)/2 = (o_1 + o_4)/(2\sum_{i=1}^4 o_i) = (110 + 80)/600 = 190/600$. Pour être rigoureux, il faudrait vérifier que l'estimateur $\hat{p}_I$ est l'estimateur du maximum de vraisemblance ; ceci est laissé en exercice. En utilisant $\hat{p}_{IV} = \hat{p}_I$ et $\hat{p}_{II} = \hat{p}_{III} = (1 - 2\,\hat{p}_I)/2$, 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

$$\sum_{j=1}^{4} \frac{(o_j - e_j)^2}{e_j},$$

qui sous $H_0$ suit la loi $\chi^2_\nu$ avec $\nu = 4 - 1 - 1 = 2$ (on a estimé un paramètre). La valeur observée de la statistique est $t_{obs} = 4.88$ qui est plus petite que le quantile $\chi^2_2(0.99) = 9.21$. Donc on ne rejette pas $H_0$.

*Dans cet exercice on n'a rejeté aucune hypothèse nulle (ni celle de la partie (a), ni celle de la partie (b)). Evidemment, les deux hypothèses nulles sont incompatibles, donc cela peut paraître bizarre. Mais rappelons-nous que "ne pas rejeter" n'est pas "accepter".*

**Exercice 3.** Cette situation peut d'abord paraître très différente de ce qu'on a fait avant. Mais en fait, elle est similaire à la partie (b) de l'exercice 2. On a

$H_0$ : les données viennent d'une loi normale,
$H_1$ : $H_0$ n'est pas vraie.

On observe 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 normal 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 $X \sim \mathrm{N}(0.173, 0.066^2)$, 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 fréquences observées et théoriques.

| | $o_j$ | $e_j$ |
|---|---|---|
| $\leq 0.1$ | 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

$$\sum_{j=1}^{5} \frac{(o_j - e_j)^2}{e_j}$$

qui, sous $H_0$, suit la loi $\chi^2_\nu$ 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.** On va tester

$H_0$ : le type du défaut est indépendant de sa localisation contre
$H_1$ : $H_0$ n'est pas vraie.

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 dénote par $O_{ij}$ 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

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

On estime les nombres attendus par

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

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 $(\sum_{j=1}^{n_2} o_{ij})/(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} o_{ij})$ et $p_j$ par $(\sum_{i=1}^{n_1} o_{ij})/(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} o_{ij})$.

La distribution asymptotique de la statistique $T$ sous $H_0$ est $\chi^2_\nu$ avec $\nu = (n_1-1)\times(n_2-1)$. On peut obtenir 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

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

est $t_{obs} = 8.08 > \chi^2_2(0.95) = 5.99$, 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 $\chi^2$ est possible, car le nombre d'observations est grand et tous les nombres attendus $e_{ij}$ sont plus grands que 5.