# Feuille d'exercices n$^o$ 5

**Tests d'hypothèses, test d'ajustement du $\chi_2$**

**Exercice 1.** Le candidat A veut se présenter à l'élection uniquement si plus de 50 % des électeurs comptent voter pour lui. Il fait faire un sondage pour évaluer ses chances. Il accepte une marge d'erreur de 5 % (c'est-à-dire une probabilité de 5 % de se présenter même si l'opinion réelle de la population ne lui est pas favorable).
a) Sur 1000 personnes interrogées, 549 disent vouloir voter A. Que peut-on en déduire sur les chances de A ? Le candidat A va-t-il se présenter ?
b) Même question si 520 personnes sur 1000 disent vouloir voter A.
*Solution : Soit $p$ la probabilité qu'un électeur préfère A. $H_0$ : $p = 1/2$. $H_1$ : $p > 1/2$.*
*Soit $(X_1, \ldots, X_n)$ un $n$-échantillon de loi $b(p)$, et*

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

*la moyenne empirique.*
*a) Ici $n = 1000$ et $M_n(\omega) = 549/1000 = 0,549$.*
*Par le théorème central limite,*

$$\frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \xrightarrow{loi} \mathcal{N}(0,1).$$

*Donc*

$$P\left(\frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \geq a\right) = P\left(M_n \geq p + \frac{a\sqrt{p(1-p)}}{\sqrt{n}}\right) \simeq P(N \geq a)$$

*où $N$ est de loi $\mathcal{N}(0,1)$. $P(N \geq a) \simeq 0,05$ pour $a \simeq 1,645$.*
*Remarque : vu les hypothèses, on cherche plutôt une zone de rejet de la forme $M_n \in [A, +\infty[$, car une valeur élevée de $M_n$ est peu probable sous $H_0$.*
*Pour $p = 1/2$ et $n = 1000$, $p + a\sqrt{\frac{p(1-p)}{n}} \simeq 0,526$.*
*Donc, si $H_0$ est vraie, $P(M_n \geq 0,526) \leq 0,05$. D'où une zone de rejet $R = [0.526; +\infty[$. Or $M_n(\omega) = 0,549 \notin R$, donc on rejette $H_0$ et on conclut que $H_1$ est vraie, pour le seuil de test $\varepsilon = 0.05$. Le candidat A décide donc de se présenter.*
*Remarque : on peut aussi tester $H_0$ : $p = 1/2$ contre $H_1'$ : $p \neq 1/2$. Dans ce cas, on considère plutôt un intervalle bilatère. On a $P(-c \leq N \leq c) \simeq 0.95$ pour $c = 1.96$, ce qui donne*

$$P\left(p - c\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq M_n \leq p + c\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \simeq P(0.456 \leq M_n \leq 0.543) \simeq 0.95$$

*On a la même conclusion.*
*b) Si seulement 520 personnes sur 1000 répondent oui, on conclut que $H_0$ est vraie.*

**Exercice 2.** On s'intéresse au problème des algues toxiques qui atteignent certaines plages. Après étude, on constate que 10% des plages sont atteintes par ces algues et on veut tester l'influence de rejets chimiques sur l'apparition de ces algues. On observe 100 plages proches de zones de rejet chimiques, et on compte le nombre de plages atteintes par l'algue : on constate que 20 plages sont atteintes par l'algue. Pouvez-vous répondre à la question, pour le seuil $\varepsilon = 0.05$ : « Les rejets chimiques modifient-ils le nombre de plages atteintes ? »

*Solution : Posons $H_0$ = "les rejets chimiques ne modifient pas le nombre de plages atteintes par les algues", et $H_1$ l'hypothèse inverse.*
*Notons $p_0 = 0.1$ la proportion de plages atteintes par l'algue parmi toutes les plages ; $p$ la proportion de plages atteintes par l'algue en cas de rejets chimiques. Alors $H_0$ s'écrit $p = p_0$ et $H_1$ s'écrit $p \neq p_0$ (ou $p > p_0$).*
*Pour tout $1 \leq i \leq 100$, soit $X_i$ la variable aléatoire qui vaut 1 si la plage est atteinte, 0 sinon. Les $X_i$ suivent une loi de Bernoulli $b(p)$. On suppose qu'elles sont indépendantes. On note $M_n = \frac{1}{n}\sum_{i=1}^{100} X_i$ (fréquence des plages atteintes).*
*L'observation nous donne $m_n = M_n(\omega) = 0,2$ .*
*Par le théorème central limite,*

$$\frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \xrightarrow{loi} \mathcal{N}(0,1).$$

*Donc*

$$P\left(-c \leq \frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \leq c\right) = P\left(p - c\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq M_n \leq p + c\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \simeq P(-c \leq N \leq c)$$

*où $N$ suit la loi $\mathcal{N}(0,1)$. On a $P(-c \leq N \leq c) \simeq 0.95$ pour $c = 1.96$. Sous l'hypothèse $H_0$ ($p = p_0 = 0.1$), cela donne $P(0.0712 \leq M_n \leq 0.1288) \simeq 0,95$ ; autrement dit, la zone d'acceptation de $M_n$ pour le test est $A =\in [0.0712; 0.1288]$. On constate que la fréquence observée $m_n = 0.2$ est dans la zone de rejet ($0.2 \notin A$). On peut donc rejeter $H_0$ et conclure, au risque 0.05, que les rejets chimiques modifient le nombre de plages atteintes par l'algue.*
*Remarque : cet exercice (ainsi que le précédent) peut aussi être fait avec le test d'ajustement du $\chi^2$ (à 1 degré de liberté).*

**Exercice 3.** A l'issue d'une expérience de 1000 tirages, un générateur de chiffres aléatoires a donné les résultats suivants :

| chiffre | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| nombre d'apparitions | 123 | 87 | 103 | 90 | 110 | 81 | 108 | 85 | 123 | 90 |

Tester au seuil de 5% l'hypothèse selon laquelle le générateur simule de façon satisfaisante un tirage uniforme sur les entiers $\{0, \ldots, 9\}$.
*Solution : Notons $X$ le résultat d'un tirage d'un entier entre 0 et 9 à l'aide de ce générateur et $\mu = (p_0, \ldots, p_9)$ sa loi (c'est-à-dire $P(X = i) = p_i$). On cherche à tester l'hypothèse $H_0$ : "$\mu$ est la loi uniforme sur $\{0, \ldots, 9\}$" contre l'hypothèse $H_1$ : "$\mu$ n'est pas la loi uniforme sur $\{0, \ldots, 9\}$" au niveau 5%. La loi uniforme sur $\{0, \ldots, 9\}$ est $\nu = (q_0, \ldots, q_9)$ avec $q_i = \frac{1}{10}$ pour $0 \leq i \leq 9$.*
*Notons $N(i)$ le nombre d'apparitions du chiffre $i$ sur $n = 1000$ tirages de chiffres à l'aide de ce générateur (nombres du tableau) et $Z = \sum_{i=0}^{9} \frac{(N(i)-nq_i)^2}{nq_i} = \sum_{i=0}^{9} \frac{(N(i)-100)^2}{100}$.*
*On rejette l'hypothèse $H_0$ au niveau 5% si la valeur observée $z_{obs} = Z(\omega)$ est telle que $P(Z \geq z_{obs}) \leq 0.05$ sous l'hypothèse $H_0$. Comme pour tout $i \in \{0, \ldots, 9\}$, $1000p_i = 100$ est suffisamment grand, on peut approximer la loi de $Z$ sous $H_0$ par celle du $\chi^2$ à 9 degrés de liberté. Donc, on peut approximer $P(Z \geq z_{obs})$ par $P(Y \geq z_{obs})$, où $Y$ suit la loi $\chi^2(9)$. A partir des valeurs observées pour les v.a. $N^{(i)}$ qui sont données dans le tableau, on obtient $z_{obs} \simeq 21.86$. D'après la table de valeurs numériques de $\chi^2(9)$, $P(Y \leq t) = 0.95$ pour $t = 16.919$. La zone d'acceptation est $A = Z \leq t$. Comme $z_{obs} > t$, on rejette $H_0$ au seuil 0.05.*
*Alternative : on voit avec la table que $P(Y \leq z_{obs})$ est compris entre 0.99 et 0.995. Donc on rejette $H_0$ au seuil de 5%, et aussi au seuil de 1% (mais pas au seuil de 0.5%).*

**Exercice 4.** Le domaine vital d'un élan se compose de feuillus ($25,8\%$ de la superficie), de forêts mixtes ($38\%$ de la superficie), de résineux ($25,8\%$ de la superficie) et d'un marécage ($10,4\%$ de la superficie). Dans ce domaine, l'élan a été localisé à 511 reprises au cours de l'année : 118 fois dans les feuillus, 201 fois dans les forêts mixtes, 110 fois dans les résineux et 82 fois dans le marécage.
L'élan fréquente-t-il indifféremment les quatre types de végétation ?
*Solution : Test du $\chi^2$ d'ajustement à 3 degrés de liberté.* $n = 511$.

$$z_{obs} = \frac{(118 - 0.258n)^2}{0.258n} + \frac{(201 - 0.38n)^2}{0.38n} + \frac{(110 - 0.25.8n)^2}{0.258n} + \frac{(82 - 0.104n)^2}{0.104n} \simeq 21,0.$$

$P(\chi^2(3) > 21) < 0.001$. *Réponse : non.*

**Exercices supplémentaires**

**Exercice 5.** On considère un dé à 6 faces. On le lance $n$ fois avec $n = 315672$. On compte le nombre $N$ de fois qu'il tombe sur 5 ou 6, on trouve $N = 106602$ (soit une fréquence de 0.3377). On voudrait savoir si le dé est équilibré. Pour cela, on appelle $p$ la probabilité de tomber sur 5 ou 6, et on considère les hypothèses :
$H_0 : p = \frac{1}{3}, \qquad H_1 : p \neq \frac{1}{3}.$
Faites un test au seuil $\varepsilon = 0.001$ pour savoir si $p = \frac{1}{3}$. Le dé est-il équilibré ?
*Solution : Intervalle d'acceptation au seuil* $0.001$ *:* $I = [p - b\sqrt{\frac{p(1-p)}{N}}, p + b\sqrt{\frac{p(1-p)}{N}}]$ *avec* $b = 3.3$, *d'où* $I = [0.3286; 0.3381]$. *La fréquence mesurée n'est pas dans* $I$, *donc on conclut* $H_1$. *Le dé n'est pas équilibré.*
*Remarque : la fréquence mesurée paraît assez proche de* $\frac{1}{3}$, *mais la grande valeur de* $n$ *permet de faire un test avec une grande précision et de rejeter l'hypothèse selon laquelle la proba de tomber sur 5 ou 6 est* $\frac{1}{3}$.

**Exercice 6.** La théorie de Mendel prédit qu'en croisant 2 types particuliers de plantes, on obtient des plantes de 4 types, qu'on note T1, T2, T3, T4, avec une loi de probabilité $\nu = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$. A la suite d'expériences, on observe 154 plantes de type T1, 44 de type T2, 63 de type T3 et 21 de type T4. Tester, au seuil de $\varepsilon = 0.05$, si cette observation est conforme à la théorie de Mendel.
*Solution : Test du* $\chi^2$. *Si* $Y$ *suit une loi* $\chi^2(3)$, *on a* $P(Y \leq t) = 0.95$ *pour* $t = 7.18$. *On calcule* $z_{obz} \simeq 4.21 < t$. *Donc on conclut que l'observation est conforme à la théorie de Mendel.*

**Exercice 7.** Une population de souris présente un taux de cancer spontané de 20%. On se demande si un traitement donné modifie ce taux. On applique ce traitement à 100 souris. On en trouve 11 cancéreuses.
a) Le traitement est-il efficace au seuil de 1% ? Au seuil de 5% ?
b) Sur combien de souris l'étude doit-elle porter pour qu'une différence de 2% entre la proportion de cancéreux dans l'échantillon et dans la population soit significative au seuil de 5% ?
*Solution : a) Ici, ce qui nous intéresse, c'est de savoir si le traitement diminue ou non la fréquence du cancer. On choisit donc de faire un test unilatéral sur les fréquences. On teste l'hypothèse* $H_0$ *: l'échantillon est tiré d'une population où la proportion de malades est 0.2, contre l'hypothèse* $H_1$ *: l'échantillon est tiré d'une population où la proportion de malades est* $< 0.2$. *On applique le TCL... On rejette l'hypothèse* $H_0$ *au seuil de 5%. En revanche, on ne peut pas rejeter* $H_0$ *au seuil de 1%. Autrement dit, si on déclare que le traitement est efficace, on a moins de 5% mais plus de 1% de chances de se tromper.*
*b) L'énoncé concerne la situation où la fréquence de la maladie dans l'échantillon est* $p = 0.18$. *Si la taille de l'échantillon est* $n$ *(assez grand),* $(M_n - 0.2)/\sqrt{p(1-p)/n}$ *suit à peu près une loi normale standard. On applique le TCL... on trouve* $n \geq 1082.4$. *On conclut qu'il faut étudier au moins 1083 souris.*

**Table (partielle) pour une v.a. $X$ de loi normale $N(0,1)$**

| $t$ | $P(0 \le X \le t)$ | $P(-t \le X \le t)$ |
|---|---|---|
| 0.3 | 0.118 | 0.236 |
| 0.6 | 0.226 | 0.451 |
| 0.68 | **0.25** | **0.5** |
| 0.8 | 0.288 | 0.576 |
| 1.26 | 0.396 | 0.792 |
| 1.32 | 0.407 | 0.813 |
| 1.645 | 0.45 | **0.90** |
| 1.96 | 0.475 | **0.95** |
| 3 | 0.499 | 0.998 |
| 3.3 | 0.4995 | 0.999 |

**Table (partielle) pour une v.a. $X$ de loi $\chi_2(9)$**

| $p$ | $t$ tel que $P(X \le t) = p$ |
|---|---|
| 0.800 | 12.2421 |
| 0.850 | 13.2880 |
| 0.900 | 14.6837 |
| 0.950 | 16.9190 |
| 0.975 | 19.0228 |
| 0.980 | 19.6790 |
| 0.990 | 21.6660 |
| 0.995 | 23.5894 |
| 0.999 | 27.8772 |

**Table (partielle) pour une v.a. $X$ de loi $\chi_2(1)$**

| $p$ | $t$ tel que $P(X \le t) = p$ |
|---|---|
| 0.800 | 1.6424 |
| 0.850 | 2.0723 |
| 0.900 | 2.7055 |
| 0.950 | 3.8415 |
| 0.975 | 5.023 |
| 0.980 | 5.411 |
| 0.990 | 6.634 |
| 0.995 | 7.8794 |
| 0.999 | 10.8276 |

**Table (partielle) pour une v.a. $X$ de loi $\chi_2(3)$**

| $p$ | $t$ tel que $P(X \le t) = p$ |
|---|---|
| 0.800 | 4.6416 |
| 0.850 | 5.3170 |
| 0.900 | 6.2514 |
| 0.950 | 7.8147 |
| 0.975 | 9.3484 |
| 0.980 | 9.8374 |
| 0.990 | 11.3449 |
| 0.995 | 12.8382 |
| 0.999 | 16.2662 |