# Estimation et tests statistiques, TD 5. Solutions

**Exercice 1** – Dans un centre avicole, des études antérieures ont montré que la masse d'un oeuf choisi au hasard peut être considérée comme la réalisation d'une variable aléatoire normale $X$, de moyenne $m$ et de variance $\sigma^2$. On admet que les masses des oeufs sont indépendantes les unes des autres. On prend un échantillon de $n = 36$ oeufs que l'on pèse. Les mesures sont données (par ordre croissant) dans le tableau suivant :

| | | | | | |
|---|---|---|---|---|---|
| 50, 34 | 52, 62 | 53, 79 | 54, 99 | 55, 82 | 57, 67 |
| 51, 41 | 53, 13 | 53, 89 | 55, 04 | 55, 91 | 57, 99 |
| 51, 51 | 53, 28 | 54, 63 | 55, 12 | 55, 95 | 58, 10 |
| 52, 07 | 53, 30 | 54, 76 | 55, 24 | 57, 05 | 59, 30 |
| 52, 22 | 53, 32 | 54, 78 | 55, 28 | 57, 18 | 60, 58 |
| 52, 38 | 53, 39 | 54, 93 | 55, 56 | 57, 31 | 63, 15 |

a) Calculer la moyenne empirique et l'écart-type empirique de cette série statistique. Tracer le boxplot et un histogramme.
b) Donner une estimation des paramètres $m$ et $\sigma$.
c) Donner un intervalle de confiance au niveau 95%, puis 98%, de la masse moyenne $m$ d'un oeuf.
d) Tester si la moyenne de cette variable est égale à 56.

a) $\bar{x} = 55.083$, $s = 2.683$, $Q1 = 53.29$, $Med = 54.96$, $Q3 = 56.5$.
Boxplot : $moust1 = 50.34$, $moust2 = 60.58$, un outlier=63.15.

Histogramme :

| | eff | largeur | hauteur |
|---|---|---|---|
| 50-52 | 3 | 2 | 1.5 |
| 52-54 | 11 | 2 | 5.5 |
| 54-56 | 13 | 2 | 6.5 |
| 56-58 | 5 | 2 | 2.5 |
| 58-64 | 4 | 6 | 0.67 |

b) $\bar{x}$ est une estimation de $m$, $s$ est une estimation de $\sigma$.
c) IC de niveau de confiance $1 - \alpha = 95\%$ pour $m$ :

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{36}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{36}}\right] = [54.207, 55.96]$$

car $z_{\alpha/2} = z_{0.025}$, $P[Z \leq 1.96] = 0.975$ quand $Z$ de loi $\mathcal{N}(0, 1)$, et donc $z_{\alpha/2} = 1.96$.
IC de niveau de confiance $1 - \alpha = 98\%$ pour $m$ :

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{36}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{36}}\right] = [54.043, 56.123]$$

car $z_{\alpha/2} = z_{0.001}$, $P[Z \leq 2.3263] = 0.99$ quand $Z$ de loi $\mathcal{N}(0, 1)$, et donc $z_{\alpha/2} = 2.3263$.

**Exercice 2** – On suppose que le poids d'un nouveau né est une variable normale d'écart-type égal à 0,5 kg. Le poids moyen des 49 enfants nés au mois de janvier 2004 dans l'hôpital de Charleville-Mézières a été de 3,6 kg.
a) Déterminer un intervalle de confiance à 95% pour le poids moyen d'un nouveau né dans cet hôpital.
b) Quel serait le niveau de confiance d'un intervalle de longueur 0,1 kg centré en 3,6 pour ce poids moyen ?

a) IC de niveau de confiance 95% pour le poids moyen :

$$\left[\bar{x} - 1.96\frac{\sigma}{7}, \bar{x} + z_{\alpha/2}\frac{\sigma}{7}\right] = [3.46, 3.74]$$

b)

$$\begin{aligned} P[\bar{X} - 0.05 \leq m \leq \bar{X} + 0.05] &= P\Big[\frac{-0.05}{\sigma/7} \leq \frac{\bar{X} - m}{\sigma/\sqrt{n}} \leq \frac{0.05}{\sigma/7}\Big] \\ &= 2F\Big(\frac{0.05}{0.5/7}\Big) = 2F(0.7) - 1 = 2 * 0.758 - 1 = 0.516 \end{aligned}$$

Le niveau de confiance est donc 0.516.

**Exercice 3** – On veut étudier la proportion $p$ de gens qui vont au cinéma chaque mois. On prend donc un échantillon de taille $n = 100$. Soit $N$ le nombre de personnes dans l'échantillon qui vont au cinéma mensuellement.
1) Quelle est la loi de $N$ ? Par quelle loi peut-on l'approcher et pourquoi ? En déduire une approximation de la loi de $F = N/n$.
2) On observe une proportion $f$ de gens qui vont chaque mois au cinéma. Donner la forme d'un intervalle de confiance pour $p$, de niveau de confiance $1 - \alpha$.
3) Applications numériques : $f = 0, 1$, $1 - \alpha = 90\%,\ 95\%,\ 98\%$.

1) On suppose que les personnes ont bien été interrogées indépendamment. Ainsi, on a un schéma de Bernoulli : une personne interrogée va au cinéma chaque mois —> SUCCES, sinon, ECHEC. Et donc $N$ suit une loi binomiale $\mathcal{B}(n = 100, p)$

$$P[X = k] = \binom{100}{k} p^k (1-p)^{100-k}, \quad k = 0, ..., 100$$

Comme $n \geq 20$, si $np > 5$ et $n(1-p) > 5$ (à vérifier lors de l'application numérique), on peut approcher cette loi par la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$, et donc $F$ suit approximativement la loi $\mathcal{N}\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big)$.
2) IC $[f - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, f + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}]$ où $P[Z \geq z_{\alpha/2}] = \alpha/2$, $Z$ de loi normale centrée réduite, $1 - \alpha$ est le niveau de confiance.
3) $f = 0.1$,
- $1 - \alpha = 90\%$, $z_{\alpha/2} = 1.645$, IC $[0.05, 0.15]$
- $1 - \alpha = 95\%$, $z_{\alpha/2} = 1.96$, IC $[0.04, 0.16]$
- $1 - \alpha = 98\%$, $z_{\alpha/2} = 2.326$, IC $[0.03, 0.17]$

**Exercice 4** – Un appareil de télécommunications reçoit un signal stocké à chaque (petite) unité de temps dans une suite de variables $(X_n)$. Cet appareil doit détecter un signal effectif, en le différenciant d'un bruit. On suppose que le bruit est une suite de variables indépendantes de loi normale de moyenne nulle et de variance 1. Pour un signal, la moyenne n'est pas nulle. Aujourd'hui on a observé une suite de 40 variables $(x_1, \ldots, x_{40})$, supposées indépendantes, de variance 1. La moyenne empirique vaut $0,6$. S'agit-il de bruit ? Construire un test pour répondre à cette question.

On veut tester $H_0 : m = 0$ contre $H_1 : m \neq 0$.
On utilise la statistique de test $Z = \frac{\bar{X}}{\sigma/\sqrt{n}}$.
Région de rejet : $|Z| > 1.96$ pour un risque 5%.
Ici, on observe $z_{obs} = \frac{0.6}{1/\sqrt{40}} = 3.79 >> 1.96$, donc on rejette $H_0$. On a bien un signal et de plus, la p-valeur vaut

$$P_{H_0}[|Z| > 3.79] = 2(1 - F(3.79)) = 0.0001$$

Le test est extrêmement significatif.

**Exercice 5** – On utilise une nouvelle variété de pommes de terre dans une exploitation agricole. Le rendement de l'ancienne variété était de 41.5 tonnes à l'ha. La nouvelle est cultivée sur 100 ha, avec un rendement moyen de 45 tonnes à l'ha et un écart-type de 11.25. Faut-il, au vu de ce rendement, favoriser la culture de la nouvelle variété ?

On veut tester $H_0 : m = 41.5$ contre $H_1 : m > 41.5$.
On utilise la statistique de test $Z = \frac{\bar{X}-41.5}{s/\sqrt{n}}$.
Région de rejet : $Z > 1.645$ pour un risque 5%.
Ici, on observe $z_{obs} = 3.11 > 1.645$, donc on rejette $H_0$. On a bien une amélioration significative du rendement et de plus, la p-valeur vaut

$$P_{H_0}[Z > 3.11] = 1 - F(3.11) = 0.00096$$

Le test est extrêmement significatif.

**Exercice 6** – Dans une agence de location de voitures, le patron veut savoir quelles sont les voitures qui n'ont roulé qu'en ville pour les revendre immédiatement. Pour cela, il y a dans chaque voiture une boîte noire qui enregistre le nombre d'heures pendant lesquelles la voiture est restée au point mort, au premier rapport, au deuxième rapport,..., au cinquième rapport. On sait qu'une voiture qui ne roule qu'en ville passe en moyenne 10% de son temps au point mort, 5% en première, 30% en seconde, 30% en troisième, 20% en quatrième, et 5% en cinquième. On décide de faire un test du $\chi^2$ pour savoir si une voiture n'a roulé qu'en ville ou non.
1) Sur une première voiture, on constate sur 2000 heures de conduite : 210 h au point mort, 94 h en première, 564 h en seconde, 630 h en troisième, 390 h en quatrième, et 112 h en cinquième. Cette voiture n'a-t-elle fait que rester en ville ?
2) Avec une autre voiture, on obtient les données suivantes : 220 h au point mort, 80 h en première, 340 h en seconde, 600 h en troisième, 480 h en quatrième et 280 h en cinquième.

On veut tester l'adéquation de notre échantillon à la loi discrète : $p_0 = 0.1$, $p_1 = 0.05$, $p_2 = 0.3$, $p_3 = 0.3$, $p_4 = 0.2$, $p_5 = 0.05$. On effectue un test du $\chi^2$. En fait, on veut tester $H_0 =$ la voiture n'a roulé qu'en ville, contre $H_1 =$ la voiture n'a pas roulé qu'en ville.
1) Pour la première voiture, on constate

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| eff obs $obs_i$ | 210 | 94 | 564 | 630 | 390 | 112 |
| eff th $th_i$ | 200 | 100 | 600 | 600 | 400 | 100 |

On calcule la distance du $\chi^2$.

$$D^2 = \sum_{i=0}^{5} \frac{(th_i - obs_i)^2}{th_i} = \frac{10^2}{200} + \frac{6^2}{100} + \frac{36^2}{600} + \frac{10^2}{600} + \frac{10^2}{400} + \frac{12^2}{100} = 6.21$$

Détermination du seuil : $P[\chi^2_5 > c] = 0.05 \Longrightarrow c = 11.07$.
Comme $D^2 = 6.21 < 11.07$, on ne peut pas rejeter $H_0$ : la voiture n'a roulé qu'en ville.
2) Pour la seconde voiture, on constate

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| eff obs $obs_i$ | 220 | 80 | 340 | 600 | 480 | 280 |
| eff th $th_i$ | 200 | 100 | 600 | 600 | 400 | 100 |

On calcule la distance du $\chi^2$.

$$D^2 = \sum_{i=0}^{5} \frac{(th_i - obs_i)^2}{th_i} = 458.67 >> 11.07$$

On rejette $H_0$ : la voiture n'a pas roulé qu'en ville. La p-valeur vaut 0. La décision ne fait pas de doute.

**Exercice 7** – Une chaîne d'agences immobilières cherche à vérifier que le nombre de biens vendus par agent par mois suit une loi de Poisson de paramètre $\lambda = 1,5$.
1) On observe 52 agents pendant un mois dans la moitié nord de la France. On trouve la répartition suivante : 18 agents n'ont rien vendu, 18 agents ont vendu 1 bien, 8 agents ont vendu 2 biens, 5 agents ont vendu 3 biens, 2 agents ont vendus 4 biens, et un agent a vendu 5 biens. Avec un test du $\chi^2$, chercher s'il s'agit bien de la loi de Poisson attendue.
2) Répondre à la même question avec les 52 agents dans la moitié sud de la France : 19 agents n'ont rien vendu, 20 agents ont vendu un bien, 7 agents 2 biens, 5 agents 3 biens et 1 agent 6 biens.

1) On veut comparer les effectifs observés avec les effectifs théoriques calculés à partir de la loi $\mathcal{P}(1.5)$.

| k | 0 | 1 | 2 | 3 | 4 | 5 | 6 et + |
|---|---|---|---|---|---|---|---|
| eff obs | 18 | 18 | 8 | 5 | 2 | 1 | 0 |
| eff th | 11.6 | 17.4 | 13 | 6.5 | 2.4 | 0.73 | 0.37 |

On doit regrouper les classes pour avoir un effectif théorique de 5 au minimum.

| k | 0 | 1 | 2 | 3 et + |
|---|---|---|---|---|
| eff obs | 18 | 18 | 8 | 8 |
| eff th | 11.6 | 17.4 | 13 | 10 |

La distance du $\chi^2$ vaut : $D^2 = 5.88$ et le seuil est 7.815 (on a 3 ddl). On accepte donc $H_0$ et on conclut que l'échantillon provient de la loi $\mathcal{P}(1.5)$.
2) $D^2 = 9.48$ et le seuil est le même, donc on rejette $H_0$ : l'échantillon ne provient pas de cette loi.

**Exercice 8** – On teste un médicament X destiné à soigner une maladie en phase terminale. On traite des patients avec ce médicament tandis que d'autres reçoivent un placebo ("contrôle"). On note dans la variable statut si les patients ont survécu plus de 48 jours. Voici le tableau obtenu

| | statut | |
|---|---|---|
| traitement | non | oui |
| contrôle | 17 | 29 |
| X | 7 | 38 |

Conclusion ?

$$\begin{aligned} D^2 &= \frac{\big(|17 - 46 * 24/91| - 0.5\big)}{46 * 24/91} + \frac{\big(|29 - 46 * 67/91| - 0.5\big)}{46 * 67/91} \\ &\quad + \frac{\big(|7 - 45 * 24/91| - 0.5\big)}{45 * 24/91} + \frac{\big(|38 - 45 * 67/91| - 0.5\big)}{45 * 67/91} \\ &= 4.335 > 3.8 \qquad \text{seuil pour 1 ddl} \end{aligned}$$

donc on rejette l'indépendance : la survie des patients dépend de leur traitement.

**Exercice 9** – On mesure la taille du lobe frontal de 30 crabes *Leptograpsus variegatus*. Voici les 30 longueurs obtenues :

12.6 12.0 20.9 14.2 16.2 15.3 10.4 22.1 19.8 15 12.8 20 11.8 20.6 21.3

11.7 18 9.1 15 15.2 15.1 14.7 13.3 21.7 15.4 16.7 15.6 17.1 7.2 12.6

Est-ce que cette variable suit une loi normale ?

On effectue un test du $\chi^2$ avec 6 classes de probabilité 0.1667.
Tout d'abord, pour la loi normale centrée réduite,

$$F^{-1}(0.1667) = -0.9661, \qquad F^{-1}(2 * 0.1667) = -0.4316, \qquad F^{-1}(3 * 0.1667) = 0$$

$F(4 * 0.1667) = 0.4316, \qquad F(5 * 0.1667) = 0.9661$

La transformation $x \longmapsto \sigma x + m$ permet de se ramener à la loi $\mathcal{N}(m, \sigma)$. Ici, les paramètres sont inconnus, on les estime donc par la moyenne empirique 15.45 et l'écart-type empirique 3.84, et on trouve les 6 classes délimitées par :

$$11.7402 \qquad 13.7927 \qquad 15.45 \qquad 17.1073 \qquad 19.1598$$

Pour ces classes, les effectifs observés sont 4-6-8-4-1-7 et les effectifs théoriques sont tous égaux à 5. Ainsi, $D^2 = 6.4$, ddl=3 et le seuil est 7.8. Donc on valide l'hypothèse de normalité des données.

**Exercice 10** – Tester l'adéquation à la loi normale $\mathcal{N}(5, 2)$ de l'échantillon suivant :

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.42 | 6.17 | 5.74 | 3.39 | 4.65 | 3.91 | 6.52 | 5.31 | 7.49 | 5.06 | 4.87 | 3.03 | 5.46 | 3.63 | 6.82 |
| 6.27 | 5.19 | 4.67 | 7.38 | 4.49 | 6.37 | 4.23 | 4.90 | 4.70 | 6.45 | 4.79 | 6.77 | 4.28 | 4.31 | 5.19 |

Pour cet exercice, on réalise un test d'adéquation du $\chi^2$ avec 6 classes d'effectifs théoriques 5. Les 6 classes sont obtenues à partir des 6 classes pour la loi normale centrée réduite trouvées dans l'exercice précédent et transformées par $x \longmapsto 5 + 2x$ :

$$3.0678 - 4.1368 - 5 - 5.8632 - 6.9322$$

Les effectifs observés sont $1 - 3 - 11 - 6 - 7 - 2$. Et $D^2 = 14$. Le seuil est 11.07. On rejette donc la loi proposée.