UNIVERSITÉ DE RENNES I
UFR DE MATHEMATIQUES

Licence Biologie Première Année
Probabilités et Statistiques

Devoir Maison

Exercice 1

Un astronome souhaite calculer, à l'aide d'un dispositif approprié, la distance d, en années-lumière (AL), entre la terre et une étoile. En raison des influences atmosphériques et d'inévitables erreurs de mesure, l'astronome prévoit de prendre plusieurs mesures et d'accepter leur moyenne comme estimation de la distance réelle d. Il y a des raisons de penser que les différentes valeurs mesurées correspondent à des variables aléatoires $X_1, \ldots, X_N$, indépendantes, identiquement distribuées, d'espérance commune d et de variance commune 4.

On cherche le nombre N de mesures que doit réaliser l'astronome afin d'obtenir une approximation de la distance d avec une marge d'erreur inférieure à 1/2 AL et un seuil de confiance de 95% (au moins).

1 - Soit $\overline{X} = \dfrac{1}{N}\sum_{k=1}^{N} X_k$. Déterminer, en justifiant le calcul, l'espérance et la variance de $\overline{X}$. Solution : $\mathbb{E}(\overline{X}) = \frac{1}{N}\mathbb{E}(\sum X_k) = \frac{1}{n}\sum_k \mathbb{E}(X_k) = \frac{1}{N}\cdot N \cdot d = d$. Pour la variance, nous avons vu en cours que pour des variables aléatoires *indépendantes* X, Y on a $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$. Ici, cela implique $\mathbb{V}(\overline{X}) = \frac{1}{N^2}\sum_k \mathbb{V}(X_k) = \frac{1}{N^2} N \cdot 4 = \frac{4}{N}$

2 - En utilisant le théorème central limite, donner une estimation du nombre N de mesures que doit réaliser l'astronome. Solution : Selon le TCL, si le nombre de mesures N est suffisamment important, on peut considérer que $\overline{X}$ est distribué comme une Gaussienne $\mathcal{N}(d, \frac{4}{N})$. On a vu en cours qu'un intervalle de confiance de niveau 95% est $[\overline{X} - z_{97,5\%}\cdot\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{97,5\%}\cdot\frac{\sigma}{\sqrt{n}}] = [\overline{X} - \frac{1{,}96\cdot 2}{\sqrt{n}}, \overline{X} + \frac{1{,}96\cdot 2}{\sqrt{n}}]$. L'astronome veut que $\frac{1{,}96\cdot 2}{\sqrt{N}} < \frac{1}{2}$, donc $N > (4\cdot 1{,}96)^2 = 61{,}4656$. Il faut donc 62 mesures !

Exercice 2

On considère l'échantillon statistique suivant :

$$1, 0, 2, 1, 1, 0, 1, 0, 0$$

1 - Calculer la moyenne et la variance empiriques de cet échantillon. Solution : moyenne empirique $\frac{1+2+1+1+1}{9} = \frac{2}{3}$, variance empirique $\frac{1^2+2^2+1^2+1^2+1^2}{9} - (\frac{2}{3})^2 = \frac{4}{9}$

2 - En supposant que les données de cet échantillon sont des réalisations d'une variable aléatoire de loi inconnue, donner une estimation non biaisée de l'espérance et de la variance de cette loi. Solution : on a vu en cours que le moyenne empirique est un estimateur non-biaisé de l'espérance, donc $\frac{2}{3}$ est une estimation non-biaisée de l'espérance. On a vu en TD qu'un estimateur non-biaisé de la variance est $\frac{n}{n-1}$ fois la variance empirique, donc $\frac{9}{8}\cdot\frac{4}{9} = \frac{1}{2}$ est notre estimation non-biaisée de la variance.

3 - On choisit de modéliser les valeurs de cet échantillon comme les réalisation d'une variable aléatoire de loi binomiale $\mathcal{B}(2, p)$.

a) Proposer un estimateur pour p basé sur la moyenne empirique. Solution : L'espérance de la loi $\mathcal{B}(2,p)$ est $2p$. Donc si $\hat{m}$ est notre estimateur de la moyenne, alors $\hat{p} = \frac{\hat{m}}{2}$ est un estimateur de p. Dans notre exemple, on a l'estimation $\hat{p} = \frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$.

b) Avec le même modèle, utiliser la variance empirique pour proposer un autre estimateur de p. Solution : la variance de la loi $\mathcal{B}(2,p)$ est $\mathbb{V} = 2p(1-p)$. Donc si $\hat{v}$ est notre estimation de la variance, alors une estimation $\hat{p}$ de p devrait satisfaire $\hat{v} = 2\hat{p}(1-\hat{p})$. On obtient donc *deux* estimations pour p, à savoir $\hat{p} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{\hat{v}}{2}}$. Dans le cas présent, $\hat{v} = \frac{1}{2}$ (voir partie 2 en haut). Nos deux estimations collapsent vers une seule $\hat{p} = \frac{1}{2}$.

4 - On choisit de modéliser les valeurs de cet échantillon par une loi de Poisson de paramètre λ. Quelle estimateur proposez-vous pour λ ? Solution : Ici, il y a plusieurs réponses raisonnables. Une loi de Poisson $\mathcal{P}(\lambda)$ est d'espérance λ, on pourrait donc utiliser la moyenne empirique (qui est un estimateur non-biaisé de l'espérance) comme estimateur de λ. On obtient ainsi une estimation $\hat{\lambda} = \frac{2}{3}$. En revanche, la variance de la loi $\mathcal{P}(\lambda)$ est aussi λ, on pourrait donc raisonnablement utiliser notre estimateur de la variance comme estimateur de λ, c.à.d., $\hat{\lambda} = \frac{1}{2}$.

Exercice 3

On a mesuré les dimensions d'une tumeur chez des souris traitées ou non avec une substance anti-tumorale. On a obtenu pour le groupe de souris témoins (échantillon de $n_1 = 30$ souris) : moyenne $m_1 = 7,075$ cm^2 et écart type $\sigma_1 = 0,576$ cm^2 et pour le groupe de souris traitées (échantillon de $n_2 = 28$ souris) : moyenne $m_2 = 5,850$ cm^2 et écart type $\sigma_2 = 0,614$ cm^2. On souhaite déterminer si la différence observée entre le groupe traité et le groupe témoin est significative en ayant recours à un test d'hypothèse.

On suppose d'une part que les données du groupe témoin correspondent à des réalisations de n_1 variables aléatoires $X_1, \dots, X_{n_1}$ indépendantes et identiquement distribuées et d'autre part que les données du groupe traité correspondent à des réalisations de n_2 variables aléatoires $Y_1, \dots, Y_{n_2}$ indépendantes et identiquement distribuées. On suppose de plus que les variables $X_1, \dots, X_{n_1}$ et $Y_1, \dots, Y_{n_2}$ sont indépendantes.

On souhaite tester l'hypothèse nulle $(H_0) : m_1 = m_2$ contre l'hypothèse alternative $(H_1) : m_1 \neq m_2$.

1 - Justifier que l'on peut considérer que de manière approchée $\overline{X} = \frac{1}{n_1}(X_1 + \cdots + X_{n_1})$ suit une loi normale $\mathcal{N}(m_1, \sigma_1^2/n_1)$ et que $\overline{Y} = \frac{1}{n_2}(Y_1 + \cdots + Y_{n_2})$ suit une loi normale $\mathcal{N}(m_2, \sigma_2^2/n_2)$. Solution : Le TCL dit exactement que, si toutes les variables aléatoires X_i suivent la même loi, de moyenne m_1 et de variance σ_1^2, et sont indépendantes, alors la variable aléatoire $\overline{X}$ suit, pour n_1 assez grand, approximativement une loi normale de moyenne m_1 et de variance $\frac{\sigma_1^2}{n_1}$, donc $\overline{X} \sim \mathcal{N}(m_1, \frac{\sigma_1^2}{n_1})$. Pour justifier qu'un échantillon de taille 30 est suffisamment grand pour pouvoir appliquer le TCL, il faudrait regarder la situation médicale en plus de détail – ceci dépasse le cadre du cours et de cette question.

2 - En admettant que la somme de 2 variables aléatoires indépendantes de loi normale suit encore une loi normale, déterminer la loi de la variable aléatoire $D = \overline{X} - \overline{Y}$ sous l'hypothèse (H_0). Solution : D'abord, $\mathbb{E}(D) = \mathbb{E}(X) - \mathbb{E}(Y) = m_1 - m_2 \overset{H_0}{=} 0$. L'observation clé maintenant est que, par hypothèse, les variables aléatoires $\overline{X}$ et $-\overline{Y}$ sont indépendantes. Par un théorème vu en cours, les variances de variables aléatoires indépendantes sont additives : $\mathbb{V}(D) = \mathbb{V}(\overline{X} + (-\overline{Y})) = \mathbb{V}(\overline{X}) + \mathbb{V}(-\overline{Y}) = \mathbb{V}(\overline{X}) + \mathbb{V}(\overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. En plus, on a admis que $D = \overline{X} + (-\overline{Y})$ suit une loi normale (approximativement), donc $D \sim \mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

3 - On note

$$T = \frac{D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Montrer que sous l'hypothèse (H_0), on a $\mathbb{P}(-z \leq T \leq z) \approx 1 - \alpha$ où z est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0, 1)$. En déduire le principe d'un test au niveau α pour l'égalité de 2 moyennes. Solution : Puisque $D \sim \mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, il vient $T \sim \mathcal{N}(0, 1)$, c.à.d. T suit une loi centrée réduite (approximativement). On a vu en cours et TD que ceci implique $\mathbb{P}(-z \leq T \leq z) \approx 1 - \alpha$, où z est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0, 1)$. On a donc un test très simple pour l'hypothèse H_0 (égalité des deux moyennes) : on calcule les moyennes empiriques $\overline{X}$ et $\overline{Y}$ et les écarts-type empiriques σ_1, σ_2 des deux échantillons. Si $\frac{\overline{X} - \overline{Y}}{\sqrt{blah\ blah}}$ est en-dehors de l'intervalle $[-z, z]$, on rejette l'hypothèse H_0, sinon on l'accepte.

4 - Calculer la valeur de la variable aléatoire T pour les échantillons observés. Qu'en déduit-on pour l'expérience considérée au seuil $\alpha = 5\,\%$. Solution : ici, $T = \frac{7,075 - 5,85}{\text{denominateur}}$, avec denominateur$= \sqrt{\frac{(0,576)^2}{30} + \frac{(0,614)^2}{28}}$. On trouve $T = 7,823$, ce qui largement est en-dehors de l'intervalle $[-z, z]$, avec $z = 1,96$. On rejette donc l'hypothèse nulle, et on accepte l'hypthèse alternative $m_0 \neq m_1$.