Mathématiques pour la biologie (math39)
Licence Biologie - Université Aix Marseille
Feuille 4 : échantillons et estimations

Sauf indication du contraire toutes les estimations de moyennes et de variances données dans les énoncés sont obtenues par les formules :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad ; \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Exercice 1/
Dans une étude immunologique, des scientifiques ont greffé de la peau de souris sur deux lots de souris. Le premier lot de 10 souris greffées n'a eu aucun traitement. Le deuxième lot de 10 souris a été traité par un antigène de transplantation. La survie des allogreffes a été mesurée en jours et les résultats de cette expérience sont les suivants :

Témoins	Traitées
$n_A = 10$ souris	$n_B = 10$ souris
$\bar{X}_A = 9.32$ jours	$\bar{X}_B = 10.56$ jours
$S_A^2 = 0.91$	$S_B^2 = 1.10$

1) Quelles sont les variables étudiées et les paramètres estimés ?
2) Que peut-on dire des valeurs de $\bar{X}_A$ et $\bar{X}_B$ si l'on prend n_A et n_B de plus en plus grands.
3) Par quelles lois pourrait-on approcher $\bar{X}_A$ et $\bar{X}_B$?
4) Proposez un estimateur sans biais de la moyenne de la prolongation de survie avec le traitement (donnez sa formule et sa valeur). Rappelez ce que signifie "sans biais".

Exercice 2/
Les tours thoraciques de n=80 vaches laitières ont été mesurés avec un ruban bovimétrique de manière indépendante. Les caractéristiques de l'échantillon sont :

$$\bar{X} = 220cm \quad ; \quad S^2 = 1252cm^2$$

On suppose que le tour thoracique est normalement distribué de moyenne m et de variance σ^2.

1) Quelle loi suit la variable aléatoire $\bar{X}$?

2) Comment se ramène-t-on à une loi normale centrée réduite ?

Exercice 3/
Une maladie atteint une proportion $p_A \in [0,1]$ ($1 \Leftrightarrow 100\%$) d'arbres d'une région A et une proportion $p_B \in [0,1]$ d'arbres d'une région B. On observe indépendamment 80 arbres de la région A et 100 arbres de la région B et le tableau suivant nous donne le nombre d'arbres malades parmi ces observations :

Région A	Région B
$n_A = 80$	$n_B = 100$
malades=11	malades=9

1) Quelle loi suit le nombre d'arbres malades (sur les 80 observées) de la région A ?

2) Donnez une estimation de p_A et p_B.

3) Quelle sont les moyennes et les variances de ces estimateurs ?

4) Quelle sont les limites de ces estimateurs quand n_A et n_B deviennent infiniment grands ?

4) Comment pourrait-on approximer les lois de ces estimateurs ?

Exercice 5/
La durée de vie d'une molécule est modélisée par une variable aléatoire continue X qui suit une loi exponentielle dont la densité s'écrit :

$$f(x) = \theta \exp(-\theta x), \qquad x \in (0; +\infty),$$

où θ est un paramètre inconnu, strictement positif, que l'on cherche à estimer. On rappelle que

$$\int_0^{+\infty} x^n \exp(-\theta x) dx = \frac{n!}{a^{n+1}}.$$

(1) Montrer que

$$P(X > t + h | X > t) \quad = \quad P(X > h).$$

Justifier l'expression de *loi sans mémoire*.

(2) On observe n durées de vie que l'on note $X_1, \cdots, X_n$ (supposées indépendantes). On propose d'estimer le paramètre θ par

$$\hat{\theta} \quad = \quad 1/\bar{X}.$$

Calculer la durée de vie moyenne et expliquer le choix de cet estimateur.

Exercice 6/
Dans une étude nutritionnelle en Europe on modélise la consommation journalière d'un certain type de produit alimentaire par une loi uniforme continue sur $[a, b]$ avec $a \leq b$ paramètres inconnus. On observe ainsi de manière indépendante n consommations notées $U_1, \cdots, U_n$.

(1) Quels estimateurs proposeriez-vous pour a et b ?

(2) Soit $y > 0$ un seuil fixé comme étant une consommation excessive. En notant $Z_i = \mathbb{I}_{U_i > y}$ montrer que les Z_i suivent une loi de Bernoulli.

3) Montrer qu'alors N_y, le nombre d'observations U_i $(i = 1, \cdots, n)$ inférieures à y, suit une loi binomiale.

4) Trouver un estimateur de la probabilité de consommer une quantité supérieurs à y.

Exercice 7/
Un agriculteur possède un champ carré dont il veut estimer la superficie. Quand il mesure un côté de son champ, il sait que l'erreur expérimentale de la mesure est une variable aléatoire supposée centrée et de variance σ^2. Il réalise une première mesure de ce côté et trouve une valeur $x_1 = 510$ mètres. Il en déduit une superficie de $s_1 = 26.01$ hectares. Il réalise une deuxième mesure et trouve alors $x_2 = 490$, d'où une valeur de la superficie $s_2 = 24.01$. Il abandonne ses mesures et réfléchit pour savoir quelle est la bonne façon

de procéder. Doit-il adopter comme estimation de la surface s_1, s_2, ou une estimation combinant les deux mesures, telle que :

$$\begin{aligned} s_3 &= \frac{s_1 + s_2}{2} = 25.01 \\ s_4 &= (\frac{x_1 + x_2}{2})^2 = 25 \\ s_5 &= x_1 x_2 = 24.99 \end{aligned}$$

Faut-il recommencer ses mesures jusqu'à ce qu'il trouve deux résultats identiques, ou bien combiner intelligemment n mesures pour construire des estimations du type s_3 ou s_4 (généralisées à ces n mesures) ?

Afin d'aider l'agriculteur à résoudre son problème, étudier les cinq estimateurs proposés, ainsi que les estimateurs construits à partir de n mesures. Sous l'hypothèse, admise dans la suite, que la loi des erreurs de mesure est une loi $\mathcal{N}(0, \sigma^2)$, on calculera notamment : les espérances et les variances de chacun des estimateurs.

Indication : si $X \sim \mathcal{N}(m, \sigma^2)$ *alors* $Var(X^2) = 2(\sigma^4 + 2m^2\sigma^2)$.