Feuille d'exercices sur les statistiques

Exercice n°1 (autocorrection)

Soit n un entier supérieur ou égal à 2.

Soit $(x_i ; y_i)$ pour $i \in [\![1, n]\!]$ une série statistique discrète telle que le moment d'ordre 2 et la variance de (x_i) soient non nuls.

La droite de régression au sens des moindres carrés est la droite d'équation y=ax+b avec (a,b) $\in \mathbb{R}^2$ qui minimise $\delta_{a,b} = \sum_{i=1}^{n}(y_i - ax_i - b)^2$

En considérant la fonction f définie sur $\mathbb{R}^2$ par $f(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$, on veut montrer que $\delta_{a,b} = \sum_{i=1}^{n}(y_i - ax_i - b)^2$ admet un minimum sur $\mathbb{R}^2$ et préciser pour quelles valeurs de a et de b, celui-ci est atteint.

1) Exprimer f(a,b) en fonction de $E(X^2)$, E(XY), E(X), E(Y) et $E(Y^2)$
2) Déterminer le point critique de f
3) Déterminer les coefficients de la matrice Hessienne
4) Conclure

On développe : $f(a,b) = n(E(Y^2)+a^2E(X^2)+b^2-2aE(XY)-2bE(Y)+2abE(X))$

Donc : $\frac{\partial f}{\partial a} = n(2aE(X^2)-2E(XY)+2bE(X))$ et $\frac{\partial f}{\partial b} = n(2b-2E(Y)+2aE(X))$

Après calculs : $a=\frac{E(XY)-E(X)E(Y)}{E(X^2)-E(X)^2}$ et b=E(Y)-aE(X).

De plus : $r = \frac{\partial^2 f}{\partial a^2} = 2nE(X^2)$ et $t = \frac{\partial^2 f}{\partial b^2} = 2n$ et $s = \frac{\partial^2 f}{\partial a \partial b} = 2nE(X)$

Or $rt-s^2=4n^2E(X^2)-4n^2E(X^2)=4n^2V(X)>0$ et $r>0$ donc minimum

Exercice n°2 (autocorrection)

Année	CA
1	3
2	7
3	10
4	18
5	21
6	35
7	53
8	82

Une entreprise réalise une étude sur l'évolution de son chiffre d'affaires, en millions d'euros, lors des 8 dernières années :

On note X la variable correspondant à l'année et Y la variable correspondant au chiffre d'affaires.

1) Représenter le nuage de points correspondant à (X,Y)
2) Donner les paramètres de la droite de régression de Y en X
3) L'ajustement précédent est-il acceptable ?

$E(X)=4,5$ $E(Y)=28,625$ $E(X^2)=25,5$ $E(Y^2)=1460.125$
$V(X)=25.5-4.5^2=5,25$
$V(Y)=1460.25-28.625^2=640.734$
$E(XY)=1461/8=182.625$
$Cov(X,Y)=E(XY)-E(X)E(Y) = 53.813$

Soit a=cov(X,Y)/V(X) = 10.25 et b=E(Y)-aE(X)=-17.5, soit y=10.25x-17.5

Le coeff de corrélation : cov(X,Y)/$\sqrt{V(X)V(Y)}$ = 0,928>0,75 donc acceptable !

Exercice n°3(autocorrection)

Après une épreuve de philosophie à un concours, on s'intéresse au lien existant pour une centaine de candidats pris au hasard, entre le temps consacré à la philosophie durant leur année scolaire et le résultat obtenu à cette épreuve.

En notant X leur note sur 20 et Y leur temps de travail en heures en philosophie durant l'année, on obtient les résultats suivants :

X Y	[0,20[	[20,50[	[50,100[	[100,200[
[0,4[	6	6	2	5
[4,8[	7	17	15	3
[8,12[	6	7	6	2
[12,16[	2	7	6	2
[16,20[	0	1	0	0

1) Déterminer la moyenne en philosophie de cet échantillon ainsi que le temps moyen consacré par cet échantillon.
2) Déterminer la droite de régression de X en Y.
3) Calculer le coefficient de corrélation linéaire de (X,Y). Conclure

E(X)=7.56 et E(Y)=55.15

$E(X^2)$=7344/100=73.44 , V(X)=$E(X^2)$-$E(X)^2$=16.286

$E(Y^2)$=481775/100=4817.5, V(Y)=1776.228

E(XY)=41170/100=411.7

Cov(X,Y)=-5.234, soit a=-0.003 et b=7.723, x=-0.003y+7.723, mais corrélation de -0.03

Exercice n°4

Un sondage portant sur 100 fumeurs choisis au hasard et avec remise au sein d'une population donnée, permet de constater que leur consommation moyenne est μ=9,37 cigarettes par jour avec un écart-type σ=2,14

1) Déterminer un intervalle de confiance à 95% de la consommation de cigarettes par jour au sein de cette population
2) Même question avec un intervalle à 99%.

Exercice n°5

A la veille du second tour d'une élection présidentielle opposant deux candidats A et B d'importances comparables.

Un institut de sondage désire estimer la probabilité que le candidat A soit élu.

On effectue l'hypothèse que les suffrages des différents électeurs sont indépendants et que chacun d'entre eux a une probabilité p de voter pour A à ce second tour.

1) L'institut de sondage désire obtenir une estimation à maximum $\pm 3\%$ du score du candidat A à ce second tour.
   a) Combien d'électeurs doit-il sonder au minimum pour obtenir un intervalle de confiance à 95% de son estimation
   b) Même question à 99%
   c) On considère que cet institut effectue un sondage auprès de 1068 électeurs parmi lesquels 552 se déclarent en faveur de A
   Déterminer l'estimation p_{obs} de p (sachant que tous les électeurs se prononcent), puis un intervalle à 95% puis à 99% de p.
2) Soit P la variable aléatoire égale à la prévision du score de A à ce second tour que l'on peut effectuer à l'aide du sondage précédent. On admet que P suit une loi normale.
   a) Déterminer les paramètres de la loi de P.
   b) Déterminer enfin la probabilité que A gagne cette élection.

Exercice n°6

Pour déterminer la teneur en potassium d'une solution, on effectue des dosages à l'aide d'une technique expérimentale donnée. On admet que le résultat d'un dosage est une variable aléatoire suivant une loi gaussienne $N(\mu, \sigma^2)$ dont l'espérance μ est la valeur que l'on cherche à déterminer, et dont l'écart-type σ est de 1 mg/litre si l'on suppose que le protocole expérimental a été suivi scrupuleusement.

Les résultats pour cinq dosages indépendants réalisés en suivant rigoureusement le protocole expérimental sont les suivants (en mg/litre) : 74.0, 71.6, 73.4, 74.3, 72.2.

1) Déterminer à partir de ces mesures un intervalle de confiance pour μ de niveau de confiance 95% et calculer l'intervalle observé
2) Quelle taille d'échantillon est nécessaire pour avoir au même niveau de confiance un intervalle de longueur inférieure à 0.1 mg/litre ?

Exercice n°7

Lors d'un sondage effectué en Ile de France, auprès de 550 personnes, il est apparu que 42 avaient de l'asthme. On se propose d'estimer par intervalle de confiance la probabilité p d'avoir de l'asthme en Ile de France. On note Z_i la variable aléatoire qui vaut 1 si la i-ème personne de l'échantillon sondé est atteinte et 0 sinon.
On admet que les variables $Z_1, \cdots, Z_n$ sont indépendantes et de même loi. On note $\overline{Z_n} = \frac{1}{n}\sum_{i=1}^{n} Z_i$
Comme $\overline{Z_n}$ est un estimateur consistant de p, on peut approcher la variance des Z_i par $\overline{Z_n}(1-\overline{Z_n})$
1)Rappeler le théorème central limite

2) Calculer des intervalles de confiance pour p, basés sur l'approximation gaussienne, de coefficients de sécurité asymptotiques 90%, 95% puis 99%. Calculer les intervalles observés.

Exercice n°8

Soit X une VAR d'espérance m et une variance v.
On dispose d'un n-échantillon $(X_1, \dots, X_n)$ de X. On appelle variance empirique de X la variable $W_n = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \overline{X_n}^2$, où $\overline{X_n}$ est la moyenne empirique de X.

1. Calculer $E(\overline{X_n})$ et $V(\overline{X_n})$,
2. En déduire $E(\overline{X_n}^2)$
3. Calculer $E(W_n)$et en déduire un estimateur sans biais de v.

Exercice n°9

Soit X une VAR de loi uniforme sur un intervalle [0, a] où a est un paramètre inconnu, et on dispose de $(X_1, \dots, X_n)$ un n-échantillon de X. On note $\overline{X_n}$ la moyenne empirique de X.

Rappels : Si X est une VAR de loi uniforme sur [0, a]. alors sa fonction de répartition est

$$F(x):\begin{cases} 0 \; si \; x \leq 0 \\ \frac{x}{a} \; si \; 0 \leq x \leq a \\ 1 \; si \; x \geq a \end{cases}$$

1. Soit $T_n = 2\overline{X_n}$. Montrer que T_n est un estimateur sans biais de a .
2. Calculer son risque quadratique.
3. Soit $T'_n = \max(X_1, \dots, X_n)$. Donner la fonction de répartition de T'_n.
4. En déduire une densité de T'_n, puis son biais et son risque quadratique.
5. Soit $T''_n = \frac{n+1}{n} T'_n$
   Déterminer son biais et son risque quadratique.
   Pour de grandes valeurs de n, quelle est le meilleur estimateur de a ?

Exercice n°10

Lors d'un sondage sur 100 personnes interrogées, 60 pensent voter pour A.

On modélise ce résultat par un échantillon $(X_1, X_2, \dots, X_{100})$ de variables indépendantes de même loi de Bernoulli de paramètre p. On cherche à déterminer un intervalle de confiance pour p au niveau de confiance 99%.

1. Déterminer l'espérance et la variance de la moyenne empirique $F = \frac{1}{100}\sum_{i=1}^{100} X_i$
2. On note F^* la variable centrée réduite associée à F.
   Par quelle loi peut-on approcher celle de F^* ? Déterminer t tel que $P(-t \leq F^* \leq t) \geq 0.99$
3. En déduire que $P(F - t\frac{\sqrt{p(1-p)}}{10} p \leq p \leq F + t\frac{\sqrt{p(1-p)}}{10}) \geq 0.99$
4. Montrer que pour tout $p \in [0, 1]$, $p(1-p) \leq \frac{1}{4}$et en déduire un intervalle de confiance pour p au niveau de confiance 0.99, puis en donner une estimation.

Exercice n°11

Afin d'étudier la proportion p de consommateurs satisfaits par un produit, on a interrogé 100 consommateurs. 56 d'entre eux ont déclaré être satisfaits par le produit.

Donner un intervalle de confiance à 95% de p.

Exercice n°12

Une usine fabrique des câbles. On suppose que la charge maximale supportée par un câble exprimée en tonnes est une variable aléatoire suivant une loi normale d'espérance m et de variance 0.5.

Une étude portant sur 50 câbles a donné une moyenne des charges maximales supportées égales à 12.2 tonnes.

1. Déterminer l'intervalle de confiance à 99% de la charge maximale moyenne de tous les câbles fabriquées par l'usine.

2. Quelle doit être la taille minimale de l'échantillon étudié pour que la longueur de l'intervalle de confiance à 99% soit inférieure ou égale à 0.2 ?

Exercice n°13

Les individus d'une population possèdent un caractère X qui suit une loi de probabilité dont la densité est donnée par $f_\theta(x) = kx^2/\theta^3$ pour $x \in [0, \theta]$ et $f_\theta(x) = 0$ sinon.

On s'intéresse au paramètre $\theta > 0$.

1. Déterminer la constante k pour que f_θ soit une densité.

2. On se donne $(X_1, \ldots, X_n)$ un échantillon de même loi que X. Montrer que la moyenne $\overline{X_n}$ n'est pas un estimateur sans biais de θ.

3. En déduire un estimateur sans biais de θ.

Exercice n°14

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes ayant pour densité : $f_\theta(x) = \frac{1}{2}(1 + \theta x)1_{[-1,1]}(x)$

1) Quelle contrainte doit-on imposer à θ pour que f_θ soit bien une densité de probabilité ?
2) Par la méthode des moments, donner un estimateur $\widehat{\theta}$ de θ.
3) Peut-on dire que $\widehat{\theta}$ est un bon estimateur de θ ?

Exercice n°15 (oraux ESSEC)

Soit θ un paramètre réel strictement positif.

Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et X une variable aléatoire ayant pour densité la fonction f définie par :

$$\forall t \in \mathbb{R} \quad f(t) = \frac{2t}{\theta^2} \cdot \mathbb{1}_{[0,\theta]}(t)$$

On considère $(X_i)_{i \in \mathbb{N}^*}$ une suite de variables aléatoires définies sur cet espace, indépendantes et identiquement distribuées, de même loi que X.

Pour tout $n \in \mathbb{N}^*$, on note $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. (a) Vérifier que f est une densité de probabilité. On notera X une variable aléatoire à densité de densité f.
   (b) Déterminer la fonction de répartition de la variable X.
   (c) Montrer que X admet une espérance et une variance et les calculer.
2. (a) Calculer $E(\overline{X_n})$ et $V(\overline{X_n})$.
   (b) Construire à partir de $\overline{X}_n$ un estimateur sans biais convergent du paramètre θ. On le notera T_n.
3. Appliquer le Théorème Central Limite à la variable $\overline{X_n}$.
   En déduire un intervalle de confiance asymptotique de θ, au niveau de confiance 95%.
   Si on note Φ la fonction de répartition d'une variable aléatoire qui suit une loi $\mathcal{N}(0,1)$, on donne, pour $t = 1,96$:
   $2\Phi(t) - 1 = 0,95$

Exercice n°16

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes identiquement distribuées de loi de Poisson de paramètre $\lambda > 0$.

Rappelons que pour tout entier positif k, $P(X_1 = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $E[X_1] = \lambda$ et $E[X_1^2] = \lambda + \lambda^2$.

Proposer un estimateur sans biais de λ et calculer son risque quadratique

Exercice n°17

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes identiquement distribuées de densité : $p_\theta(x) = \exp(\theta - x)1\{x \geq \theta\}$, $\theta > 0$.

On cherche dans cet exercice à estimer le paramètre de translation θ.

1. Proposer un estimateur sans biais de θ par la méthode des moments : on le notera $\widehat{\theta_n}$.

2. Calculer le risque quadratique de l'estimateur $\widehat{\theta_n}$.

3. On considère à présent l'estimateur $\widetilde{\theta_n}$ défini par $\widetilde{\theta_n}(X) = \inf_{1 \leq i \leq n} X_i$. Calculer sa loi.

4. L'estimateur $\widetilde{\theta_n}$ est-il sans biais ?

5. Calculer le risque quadratique de l'estimateur $\widetilde{\theta_n}$.

Exercice n°18

Soient $X_1, \ldots, X_n$, n variables aléatoires indépendantes de loi $B(\theta)$ (Bernoulli de paramètre θ).

1. Montrez que $X = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de θ efficace.

2. Trouver un estimateur sans biais de la variance des X_i de la forme $vb = \eta X(1 - X)$.

Exercice n°19 (oral ESSEC)

Soit $X_1, ..., X_n$ des variables aléatoires i.i.d. telles que :

$$P(X_1 = -1) = (1-p)^2, \quad P(X_1 = 0) = 2p(1-p), \quad P(X_1 = 1) = p^2$$

On cherche à estimer $p \in]0, 1[$. Montrer que

$$Z_n = \sum_{k=1}^{n} \frac{1 + X_k}{2n}$$

est un estimateur sans biais et convergent de p.

Exercice n°20 (Edhec)

Soit X une v.a.r. réelle admettant une espérance est m et une variance σ^2.
Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de la v.a.r. X.

1. Montrer que $T_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ est un estimateur sans biais de m.

2. On pose $V_n = \dfrac{1}{n}\sum_{k=1}^{n} (X_k - T_n)^2$.

a) Montrer que : $V_n = \dfrac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - (T_n)^2$.

b) Montrer que : $V_n = \dfrac{1}{n}\left(\sum_{k=1}^{n} (X_k - m)^2\right) - (T_n - m)^2$.

c) Montrer que $\mathbb{E}(V_n) = \dfrac{n-1}{n}\sigma^2$.

d) Construire, à partir de V_n, un estimateur sans biais $\widehat{V_n}$ de σ^2.

e) On dispose d'un échantillon de n observations $(x_1, x_2, \ldots, x_n)$ de la v.a.r. X. Donner une méthode pour obtenir une estimation ponctuelle de σ^2 à partir de ces observations.

Exercice n°21 (Essec 2009)

La sécurité routière fait une enquête sur le nombre d'accidents survenus par semaine sur un tronçon d'autoroute.
Soit X la v.a.r. égale au nombre d'accidents par semaine. On suppose que X suit une loi de Poisson de paramètre θ inconnu ($\theta \in]0, +\infty[$).
On se propose d'évaluer le paramètre $e^{-\theta} = \mathbb{P}([X = 0])$.
On note X_1, X_2, ..., X_n les résultats des observations faites pendant n semaines. On suppose $X_1, \ldots, X_n$ indépendantes et de même loi que X.

1. Pour tout $i \in [\![1, n]\!]$, on définit Y_i par : $Y_i = 1$ si $X_i = 0$, et $Y_i = 0$ sinon.
On note aussi : $\overline{Y_n} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.

a) Pour tout $i \in [\![1, n]\!]$, donner la loi de Y_i.

b) Montrer que $\overline{Y_n}$ est un estimateur sans biais de $e^{-\theta}$.

c) Calculer le risque quadratique de $\overline{Y_n}$.

d) Montrer que $\overline{Y_n}$ est un estimateur convergent de $e^{-\theta}$.