Mini-Projet

Sciences des données et la décision

Objectifs :

- Préparation des données.
- Calculer les paramètres statistiques d'une série statistique unidimensionnelle.
- Calculer les paramètres statistiques d'une série statistique à deux variables (correlation et ajustement).
- Représenter graphiquement une série statistique à une et deux variables (nuage de points, droite de regression histogramme, diagramme en boite,...).
- Calculer les probabilités d'une variable aléatoire à l'aide du logiciel Python.
- Estimation paramétrique ; pproximation des lois ; Test de Khi-2.

N.B :

1. Le Mini-projet devra être fourni sous forme d'un fichier zip (codes, compte rendu des résultats en Word ou latex, ... (6 pages maximum).
2. Le Mini-projet devra être rendu (groupes de 3 ou 4 étudiants) dans le même jour une fois la deuxième séance de Mini-Projet a été réalisée à 23h59 au plus tard (heure de Paris) en le soumettant via la plateforme Moodle.
3. Aucun délai supplémentaire ne sera accordé.
4. Les consignes ainsi que les notations du sujet doivent être respectées.
5. L'implication, la rigueur, ainsi que la discipline seront en grande partie tenues en compte lors de l'évaluation.
6. Votre travail sera évalué au fur et à mesure. N'hésitez donc pas à montrer votre travail à votre enseignant une fois un exercice entièrement terminé.
7. L'absence non justifiée ou la présence en classe sans ordinateur portable à une des trois séances seront pénalisées par un malus de 3 points (chaque séance).
8. Pour les retardataires, en cas de soucis particulier sous la plateforme Moodle, l'élève doit formuler un mail (joindre son travail également) à son professeur dans les 10 minutes qui suivent le deadline, en expliquant les raisons.

1 Méthodologie

- Créer un dossier nommé **Nom-Elèves-Mini-Projet** qui contiendra tous les scripts et toutes les fonctions à programmer.

---

1. École d'Ingénieurs de l'Air et de l'Espace et de la Mobilité Durable, 94200 Ivry-sur-Seine

— Il est fortement conseillé de travailler par écrit sur une feuille avant de commencer l'écriture du programme avec le langage Python.
— Il est aussi fortement conseillé de tester les lignes de commande au fur et à mesure de la construction de votre programme.
— Ajouter des commentaires pour améliorer la lisibilité de vos codes. Cela vous fera gagner beaucoup du temps lors d'un usage ultérieur.
— Dans tout ce travail, **on s'efforcera de ne pas utiliser de boucle for ou while**.

2 Statistique à deux variable : ajustement et correlation

Exercice 1 *Dans une ferme industrielle, le service vétérinaire veut modifier le régime alimentaire des vaches, dans le but d'augmenter la production laitière. Pour cela, on a choisi au hasard* 15 *vaches que l'on a nourries pendant un mois avec l'aliment habituel X et l'on a relevé pour chaque vache la production quotidienne moyenne du lait exprimée en Kg puis, on a nourri ces mêmes vaches pendant un mois avec le nouveau aliment Y et on a relevé de même la production quotidienne moyenne de chaque vache (voir fichier* ***dataMP.csv****).*

1. (a) *Charger le fichier* ***dataMP.csv*** *dans votre dossier.*
   (b) *Exporter ces données dans un tableau nomé* ***Data****, puis les afficher.*
   (c) *Charger X et Y.*
2. *Dessiner le nuage de points de cette distribution (Y en fonction de X).*
3. *Calculer les moyennes et les écarts-types des variables X et Y.*
4. *Calculer la covariance $cov(X, Y)$.*
5. *Calculer le coefficient de corrélation linéaire r. Que peut-on conclure ?*
6. (a) *Déterminer les paramètres de la droite de régression linéaire de Y en X puis dessiner cette droite. (dans le même repère que le nuage de points).*
   (b) *Retrouver ces paramètres à l'aide des formules obtenues en cours.*
7. *Vérifier que le point $G(\overline{x}, \overline{y})$ appartient à cette droite.*
8. (a) *Déterminer la variance résiduelle par deux méthodes.*
   (b) *Déterminer la variance expliquée par deux méthodes.*
   (c) *Vérifier bien l'équation de la variance.*

Exercice 2 *L'étude de la décharge d'un condensateur a apporté les résultats suivants :*

T (ms)	0	5	10	15	20	25	30	35	40	45	50
V	5,098	3,618	2,581	2,011	1,486	1,028	0,845	0,573	0,429	0,29	0,2

On suppose que la décroissance est exponentielle et suit la loi

$$V = V_0\, e^{-\frac{t}{\tau}} \tag{1}$$

où τ représente une constante de temps du circuit.

1. *Montrer que l'equation* (1) *peut s'écrire sous la forme $y = ax + b$ (préciser bien les choix des variables x et y).*
2. *Sauver les valeurs de x dans un tableau nommé* ***Xln****.*

3. *Sauver les valeurs de y dans un tableau nommé* ***Yln****.*
4. *En utilisant la méthode des moindres carrées, déterminer la droite de régression linéaire.*
5. *Évaluer V_0 et τ.*
6. *Calculer V pour $t = 53\ ms$*

Exercice 3 *On se propose d'examiner s'il existe un lien corrélatif entre la production agricole et la production industrielle pour la période 1944-1962.*

Année	*x (la production agricole)*	*y (la production industrielle)*
1944	*100*	*10*
1945	*61*	*50*
1946	*76*	*84*
1947	*74*	*99*
1948	*90*	*113*
1949	*93*	*122*
1950	*102*	*128*
1951	*98*	*143*
1952	*103*	*145*
1953	*110*	*145*
1954	*117*	*159*
1955	*118*	*172*
1956	*112*	*188*
1957	*115*	*204*
1958	*116*	*213*
1959	*121*	*220*
1960	*134*	*242*
1961	*130*	*254*
1962		*273*

1. *Saisir les valeurs de la production agricole dans un tableau nommé XPA.*
2. *Saisir les valeurs de la production industrielle dans un tableau nommé YPI.*
3. *Faire un nuage de points.*
4. *Déterminer le point moyen G et le placer sur le même graphique.*
5. *Déterminer la covariance entre x et y.*
6. *Calculer le coefficient de corrélation r.*
7. *Interpréter.*
8. (a) *Déterminer la droite de régression linéaire de y en x (a et b à déterminer).*
   (b) *Déterminer les valeurs ajustées $\hat{y}_i$ ainsi que les distances de chaque point par rapport à la droite d'ajustement obtenu par la méthode des moindres carrés.*
   (c) *Déterminer la variance résiduelle et la variance expliquée.*
   (d) *Tracer dans le même graphique cette droite.*
9. (a) *Déterminer la droite de régression linéaire de x en y (a' et b' à déterminer).*
   (b) *Déterminer les valeurs ajustées $\hat{x}_i$ ainsi que les distances de chaque point par rapport à la droite d'ajustement obtenu par la méthode des moindres carrés.*
   (c) *Déterminer la variance résiduelle et la variance expliquée.*
   (d) *Tracer dans le même graphique cette droite.*

(e) *Conclure.*

10. (a) *Rappeler la relation entre aa' et r.*
    (b) *Vérifier numériquement ce résultat.*
11. *Calculer la valeur de la production agricole en 1962.*

3 Statistique descriptive à une dimension

Exercice 4 *Afin d'étudier la structure de la population de gélinottes huppées (Bonasa umbellus) abattues par les chasseurs canadiens, une étude du dimorphisme sexuel de cette espèce a été entreprise. Parmi les caractères mesurés figure la longueur de la rectrice centrale (plume de la queue). Les résultats observés exprimés en millimètres sur un échantillon de 50 mâles juvéniles sont donnés par*

153	165	160	150	159	151	163	160	158	150	154	153	163	150
158	150	158	155	163	159	157	162	160	152	164	158	153	162
166	162	165	157	174	158	171	162	155	156	159	162	152	158
164	164	162	158	156	171	164	158						

1. *Déterminer la population.*
2. *Déterminer la variable statistique.*
3. *Cette variable est-elle discrète ou continue ?*
4. *Saisir ces valeurs dans un tableau nommé* ***BonasaUmbellus****.*
5. *Calculer la longueur minimum, la longueur maximum, l'étendue, la moyenne, la médiane, les quartiles, l'écart interquartile.*
6. *Déterminer les paramètres de dispersion (variance et écart-type) de cette série statistique. Commenter.*
7. *Représenter la boîte de dispersion de la série. Commenter.*
8. *Représenter graphiquement la répartition des longueurs par un histogramme.*

Exercice 5 *Dans une maternité, les pesées de 30 bébés, exprimées en kilogrammes, ont donné les résultats suivants :*

2,8	3	2,9	2,4	2,9	3,7	2,1	3,4	2,3	3,1	3.2	2.6	3.5	3.4	2.8
1.9	3,4	2,5	3.5	2,8	3,8	1,8	3	2	2,7	2, 6	2,8	1,9	2,9	2,6

Partie I :

1. *Saisir dans un tableur les poids de 30 bébés dans un fichier nommé* ***data.cvs*** *ou* ***data.dat*** *ou autre*
2. *Charger les données du fichier précédent dans un tableau nommé* $\mathbf{S}_1$.
3. *Organiser les valeurs ci-dessus sous forme d'une série statistique* $\mathbf{S}_1$.
4. *Représenter cette série par un diagramme en bâtons.*
5. *Calculer la moyenne et l'écart-type de cette série.*
6. *Représenter la boîte de dispersion de cette série et interpréter ?*

Partie II :

1. *Saisir la variable* ***Poids1*** *contenant les 15 premières valeurs des poids, puis calculer leur moyenne* $\mathbf{m}_1$.
2. *Saisir la variable* ***Poids2*** *contenant les 15 valeurs suivantes, puis calculer leur moyenne* $\mathbf{m}_2$.
3. *Déduire la moyenne totale des poids de 30 bébés.*

4 Lois de probabilités et approximations

Exercice 6 *Dans un aéroport, les portiques de sécurité servent à détecter les objets métalliques que peuvent emporter les voyageurs. On choisit au hasard un voyageur franchissant un portique.* 500 *personnes s'apprêtent à passer le portique de sécurité. On suppose que pour chaque personne la probabilité que le portique sonne est égale à* $0,031175$.
Soit X *la variable aléatoire donnant le nombre de personnes faisant sonner le portique, parmi les* 500 *personnes de ce groupe.*

1. *Donner la définition de la loi de Binomiale ainsi que sa moyenne et son écart-type.*
2. *Justifier rigoureusement que* X *suit une loi binomiale et donner ces paramètres.*
3. *Calculer l'espérance de* X. *Donner une interprétation au résultat obtenu.*
4. *Donner la valeur arrondie à* 10^{-4} *de la probabilité qu'au moins une personne du groupe fasse sonner le portique.*
5. *Donner la probabilité qu'au maximum* 3 *personnes fassent sonner le portique.*
6. *Peut-on calculer simplement* $P\big(X > 250\big)$ *?*
7. *Montrer qu'une approximation de la loi binomiale par une loi normale se justifie.*
8. *Calculer* $P\left(X > 50\right)$ *à l'aide de cette approximation.*
9. *A l'aide de cette approximation, calculer* $P\left(X = 70\right)$.
10. *Peut-on approximer cette loi par une loi de Poisson ? Justifier votre réponse.*

Exercice 7 *On sait par expérience qu'une certaine opération chirurgicale a* 90% *de chances de réussir. Cette opération est réalisée dans une clinique* 400 *fois chaque année. Soit* Y *le nombre de réussites dans une année. On considère que* Y *suit loi normale d'espérance* $\mu = 360$ *et de variance et* $\sigma^2 = 36$.

1. *Tracer la courbe de la densité de probabilité de* Y.
2. *Calculer la probabilité que la clinique réussisse au moins* 345 *opérations dans l'année.*
3. *Calculer la probabilité que la clinique rate plus de* 28 *opérations dans l'année.*
4. *L'assurance accepte de couvrir un certain nombre d'opérations ratées : ce nombre n'a que* 1% *de chances d'être dépassé. Quel est-il ?*

Exercice 8 *Dans une rue passante de Paris, on a mesuré le niveau de bruit en décibels (db) émis par* $n = 22$ *véhicules pris au hasard. Les données ordonnées sont les suivantes :*

54,8	*55,4*	*57,7*	*59,6*	*60,1*	*61,2*	*62,0*	*63,1*	*63,5*	*64,2*	*65,2*
65,4	*65,9*	*66*	*67,6*	*68,1*	*69,5*	*70,6*	*71,5*	*73,4*	*75*	*75,2*

1. *Donner la moyenne* m_e *ainsi que l'écart-type* σ_e *de cet échantillon.*
2. *Construire un histogramme.*
3. *Par une idée intuitive, peut-on assimiler le bruit par une loi normale ? justifier.*
4. *On suppose que cet échantillon est issu d'une loi normale* X *de paramètre* m *(espérance mathématique) et d'écart-type* σ.

   (a) *Estimer la moyenne* m *et l'écart-type* σ *pour* X.

(b) *Quelle est la loi de* $T = \dfrac{\overline{X} - m}{\dfrac{s}{\sqrt{n}}}$ *?*

(c) *Déterminer un intervalle de confiance pour la moyenne avec la confiance de* 95%.

(d) *Tracer dans le même graphique (précédemment), la courbe de la densité de probabilité de la variable aléatoire* X.

(e) *Estimer la probabilité que le niveau de bruit dépasse* 70 *db.*

(f) *Estimer la probabilité que le niveau de bruit est entre* 60 *db et* 75 *db .*

(g) *Déterminer* t_1 *tel que* $P(X < t_1) = 0,95$.

(h) *Déterminer* t_2 *tel que* $P(X \geq t_2) = 0,25$.

Exercice 9 *Sur un marché voisin du poulailler industriel, un marchand vend des oeufs de ferme. Le poids des* 100 *oeufs se répartissent de la façon suivante :*

Poids en grammes	< 55	$]55, 57]$	$]57, 59]$	$]59, 61]$	$]61, 63]$	> 63
Nombre	12	12	15	18	20	23

1. *Présenter cette distribution à l'aide d'un histogramme. De manière intuitive, peut-on conjecturer que cette distribution peut être modéliser par une loi normale ?*
2. *Déterminer la moyenne* m_e *et l'écart-type* σ_e *de cette distribution.*
3. *Le vendeur affirme qu'avec un seuil de signification de* 5%*, cette distribution peut être modélisée par une loi normale de paramètre* $m = m_e$ *(espérance mathématique) et d'écart-type*

$$\sigma = \sigma_e \sqrt{\frac{n}{n-1}}.$$

Le vendeur a-il pris la bonne décision ? Justifier.

5 Méli-mélo

Exercice 10 *(Télévision et espérance de vie) Les données contiennent des informations sur l'espérance de vie à la naissance, ainsi que le nombre de personnes par télévision et le nombre de personnes par physicien, pour les pays de plus de* 20 *millions d'habitants en* 1990 *(*40 *pays). Ces données se trouve dans le fichier "televisions.txt". Chaque ligne contient les informations sur un pays, délimitées par le caractère tabulation. Les variables sont les suivantes :*

Nom des variables	*Description*
pays	*nom des pays*
espvie	*espérance de vie à la naissance*
tv	*nombre de personnes par télévision*
phys	*nombre de personne par physicien*
espvieF	*espérance de vie àla naissance des femmes*
espvieH	*espérance de vie à la naissance des hommes*

1. (a) *Importer le fichier* ***televisions.dat*** *dans la table de données que l'on nommera televisions.*

   (b) *Visualiser les données et déterminer la nature des variables.*

2. *Proposer des résumés numériques pour chaque variable.*

3. (a) *Proposer des représentations graphiques pour les variables* ***tv*** *et* ***phys****.*
   (b) *Détecter les données atypiques ("outliers").*
4. *Construire des histogrammes et des boîtes à moustaches pour l'espérance de vie, puis pour les espérances de vie des hommes et des femmes.*
5. *Etudier les relations entre l'espérance de vie et le nombre de personnes par télévision. Pour ce faire, tracer un nuage de points entre les variables* ***espvie*** *et* ***tv****. Tracer à nouveau le nuage de points en supprimant les observations atypiques, puis le nuage de points entre les variable* ***espvie*** *et* ***log(tv)****.*
6. *Faire de même pour les variables* ***espvie*** *et* ***phys****.*
7. *Peut-on en déduire un lien de causalité entre ces variables ?*