BUT GEII - S4
Florent ARNAL
florent.arnal@u-bordeaux.fr

Génie électrique et informatique industrielle / iut de BORDEAUX

TD/TP de statistique descriptive

Statistique descriptive univariée

Exercice 1 :
Voici les résultats obtenus au cours du semestre par 11 étudiants :

```
10.25 13 16.5 9.75 12.25 12.75 14.25 12 17.5 12.25 13.5
```

1. Proposer deux représentations graphiques de ces notes.
2. Déterminer la médiane et les quartiles de ces notes.
3. Calculer la moyenne et l'écart-type de ces notes.

Exercice 2 : [Calcul de moyenne et écart-type]

1. Sur un demi-groupe TP constitué de $n_1 = 12$ étudiants, on a relevé une moyenne $\overline{x_1} = 10$.
   Sur l'autre demi-groupe de TP, de $n_2 = 13$ étudiants, on a relevé une moyenne $\overline{x_2} = 12$.
   Déterminer la moyenne du groupe constitué de ces deux demi-groupes.
2. Sur un échantillon de taille $n_1 = 10$, on a relevé un écart-type $s_1 = 2$.
   Sur un nouvel échantillon, de même moyenne, de taille $n_2 = 20$, on a relevé un écart-type $s_2 = 1$.
   Quel est l'écart-type s de ces prélèvements ?

Exercice 3 : [Écart-type et étendue]
Cet exercice est lié à des notions utilisées lors d'essais interlaboratoires pour estimer la "qualité" de méthodes d'essais.
On s'intéresse à des mesures (analyses) sur un produit en effectuant 2 répétitions dans un laboratoire.
On considère donc un échantillon constitué de 2 valeurs y_1 et y_2.

1. Exprimer la SCE et la variance observée s^2 de la série de mesures $(9; 11)$ ainsi que son étendue R.
2. Déterminer une relation entre la variance observée (empirique) s^2 d'une série de mesures $(y_1; y_2)$ et son étendue R.
3. En déduire une relation entre l'écart-type observé et l'étendue R.

Exercice 4 :
La nutrition du jeune adulte est à ce jour un sujet largement inexploré dans le monde. La grande cohorte d'étudiants i-Share (Internet-based Students Health and Research Enterprise), en cours de constitution à Bordeaux, offre une opportunité de mieux connaître les habitudes alimentaires de cette population, afin de mieux définir des stratégies de santé publique dans le futur. Les données sont accessibles via le fichier 'DonneesIShare.csv' disponible sur moodle (avec un descriptif de l'enquête).

1. Combien y a-t-il de personnes âgées de 21 ans qui ont participé à l'enquête ?
2. Dans cette partie, on ne considère que les personnes âgées de 21 ans.
   (a) Représenter graphiquement le nombre de fruits mangés quotidienneemnt.
   (b) Déterminer le nombre moyen ainsi que le nombre médian de fruits mangés par ces personnes.
3. Déterminer un résumé paramétrique du nombre de fruits consommés quotidiennement pour chaque catégorie d'âge.
4. Déterminer une représentation graphique (à l'aide de boxplots) du nombre de fruits consommés en fonction de l'âge.
5. Proposer au moins deux représentations graphiques pour illustrer les données liées à l'IMC des étudiants de cette étude.
   Il est conseillé de chercher des informations sur l'IMC (Indice de masse corporelle) sur des sites tels que https://www.has-sante.fr/upload/docs/application/pdf/2011-10/annexe_1_table_dindice_de_masse_corporelle.pdf.

Exercice 5 :
Les données suivantes correspondent aux salaires nets mensuels d'apprentis de première et deuxième année d'une formation de l'enseignement supérieur.

Année 1	Année 2
1040.00	1040.00
1284.01	1231.53
847.00	900.00
720.00	1150.00
1000.62	876.00
661.94	970.00
800.00	986.00
969.00	1100.00
900.00	930.00
900.00	963.00
747.00	1200.00
785.00	901.00
903.00	1049.45
907.20	1014.96

Commenter ces résultats.

Exercice 6 :
On s'intéresse à une production d'eau minérale naturelle de quantité nominale 100 cl.
Le cahier des charges impose que le volume des bouteilles soit compris entre 99 cl et 101 cl. En outre, l'entreprise souhaite que la moyenne de la production soit égale à la quantité nominale annoncée. L'entreprise a prélevé un échantillon de 20 bouteilles et a obtenu les résultats suivants :

```
 [1]  99.9  99.4  99.9  99.9  99.6  99.0  99.2  99.4  99.5 100.2  99.7  99.3
[13]  99.2  99.4  99.4  99.4  99.6  99.2  99.6  99.1
```

Commenter les résultats obtenus à l'aide de résumés paramétriques et graphiques.

Exercice 7 :
Une entreprise souhaite mettre en place une nouvelle méthode d'analyse pour détecter les taux d'adultération. Pour ce faire, sur 10 pâtés différents, un opérateur effectue le dosage avec la méthode actuelle et la méthode "testée".
Les résultats observés sont les suivants :

	Methode_actuelle	Methode_testee
1	4.8	5.2
2	5.1	5.1
3	5.0	5.3
4	5.0	5.4
5	4.8	5.6
6	4.9	5.1
7	4.9	5.5
8	5.0	5.3
9	4.8	5.2
10	5.2	5.2

1. Proposer une représentation graphique adaptée.
2. Que pouvez-vous dire de la nouvelle méthode ?

Exercice 8 : Diagramme de Pareto
Créer le diagrammme de Pareto associé aux données suivantes :

Machine	1	2	3	4	5	6	7	8
Nombre de pannes	26	13	35	24	20	52	5	2

Exercice 9 :
Cet exercice a pour objectif d'aborder, au travers d'un exemple, les notions essentielles en statistiques à deux variables.
Afin de procéder à l'étalonnage d'un nouvel appareil de mesures, on effectue 5 mesures (grandeurs obtenues) avec cet appareil associées à 5 valeurs de référence X (grandeurs théoriques).
Pour chaque mesure effectuée, on note Y la valeur obtenue.
A noter que les valeurs observées ont toujours été supérieures aux valeurs attendues. Les résultats sont consignés dans le tableau suivant :

Valeurs de référence x_i	0,2	0,5	1	2	4
Écarts y_i	0,283	0,676	1,311	2,631	5,231

1. Représenter graphiquement cette série de données.
   Un ajustement affine vous paraît-il judicieux ?
2. Déterminer les coefficients de corrélation linéaire et de détermination entre les variables X et Y.
   Interpréter ces valeurs.
3. Déterminer une équation de la droite d'ajustement linéaire de Y en X.
4. Calculer les estimations $\hat{Y}$.
5. Déterminer les différents résidus et les représenter en fonctions de X.
6. En déduire la SCE_{res} puis la SCE_{exp}.
7. Retrouver, en utilisant un quotient de SCE, la valeur du coefficient de détermination.
8. Pour une valeur mesurée (observée) de 6,5, donner un ordre de grandeur de la mesure attendue (théorique).

Exercice 10 :
On considère une série statistique double (X, Y) associée aux valeurs suivantes :

X	-3	-2	-1	0	1	2	3
Y	2	1	0	-1	0	1	2

1. Proposer une représentation graphique de cette série.
2. Déterminer la valeur du coefficient de corrélation et de la covariance entre X et Y.
3. (a) Déterminer une équation de la droite de régression de Y en X.
   Donner une équation "plus précise" en utilisant, notamment, les questions précédentes.
   (b) On pose $Z = |X|$.
      i. Représenter la série (Z, Y).
      ii. Déterminer une expression de Y en fonction de X.

Exercice 11 :
Cet exercice a pour objectif d'aborder, au travers d'un exemple, les notions essentielles en statistiques à deux variables.
Dans le tableau suivant, on a reporté le nombre (moyenne lissée sur une semaine) de nouveaux cas quotidiens atteints par le virus de la Covid-19, toutes les quinzaines, du 15 juillet au 30 octobre 2020. La date 1 correspond au 15 juillet, la date 2 au 30 juillet, ..., la date 8 au 30 octobre.

```
date         1    2    3    4    5     6     7     8
contamines 906 1319 2841 5006 8511 11771 19721 41558
```

On note X la variable égale à la date et Y la variable égale au nombre de contaminés.

1. Représenter ces informations à l'aide d'un graphique adapté.
2. Un ajustement affine du nombre de contaminés en fonction de la date est-il pertinent ?
3. On convient de noter Z la variable définie par $Z = \ln Y$.
   (a) Représenter le nuage de points associé à X et Z
   (b) Un ajustement affine de Z en X est-il pertinent ?
   (c) Déterminer une équation de la droite d'ajustement de Z en X.
   (d) En déduire une expression de Y en fonction de X.
   (e) En admettant que la progression "naturelle" se soit poursuivie (sans confinement), donner une estimation du nombre quotidien de cas qu'il y aurait eu autour du 30 novembre 2020.

Exercice 12 :
Un laboratoire teste l'utilisation d'un nouveau produit bio dans le but d'éradiquer les pucerons dans des cultures de roses. Le test porte sur 10 parcelles de 100 m^2 indépendantes, isolées les unes des autres et présentant le même niveau d'attaque par les pucerons. On souhaite étudier la corrélation entre Y le nombre moyen de pucerons par plant et X la quantité de produit utilisé en ml/m^2 sur la parcelle. Les mesures sont faites 3 jours après la pulvérisation.

Parcelle	1	2	3	4	5	6	7	8	9	10
X	15	20	25	11	34	20	16	14	18	11
Y	5	$3,5$	$2,7$	$7,4$	$1,8$	$3,5$	$4,6$	5.5	4	$7,4$

Comparer la pertinence d'un ajustement affine et d'un ajustement puissance obtenu en utilisant un ajustement affine de $\ln Y$ en $\ln X$.
Déterminer, si cela semble cohérent, une estimation de la quantité de produit à épandre par mètre carré pour que le nombre moyen de pucerons par plan soit inférieur ou égal à 3.