TP 2-Statistiques univariées

1 Représentation de données

1.1 Série statistique

Vocabulaire des statistiques

- L'ensemble Ω des éléments dont on étudie les données est appelé **population**. Ses éléments sont appelés les **individus**.
- Un **échantillon** est un sous-ensemble (fini) de la population.
- L'**effectif** d'une population, d'un échantillon est le nombre d'individus de cette population, cet échantillon.

Définition 1 (Variable, série statistique)

- Une **variable** (ou caractère) est une application définie sur la population Ω. Une variable est dite quantitative lorsqu'elle est à valeurs réelles et qualitative sinon.
- Les valeurs prises par X sont appelées les **modalités**.
- La liste des valeurs prises par X est appelée une **série statistique**.

Exemple 1

Si on considère la population française, alors

1. *Le caractère qui donne le revenu annuelle est un caractère quantitatif.*
2. *Le caractère qui donne la catégorie socio-professionnelle est un caractère qualitatif.*

On peut représenter une série statistique de deux manières :

- en donnant la liste « brute » des valeurs prises par X : $[X(\omega), \omega \in \Omega]$;
- en donnant la liste des modalités **distinctes** prises par X affectées de leur effectif d'apparition (on dira que la série est groupée par modalités).

Exemple 2

La série statistique

Modalités	1	2	2	7	8	5	8	1

peut aussi être représentée par

Modalités	1	2	7	8	5
Effectifs	2	2	1	2	1

La commande `tabul`

La commande `tabul` prend en argument une série statistique et renvoie une matrice à deux colonnes : la première colonne contient les valeurs distinctes de la série statistique classées dans l'ordre décroissant (par défaut) ou croissant (option) et la seconde colonne les effectifs correspondants. Pour plus de détail, voir l'aide.

Exercice 1

1. *Avec la commande `grand` (voir l'aide Scilab), créer une série statistique contenant les résultats d'une simulation de* 100 *le lancers d'un dé à six faces équilibré.*
2. *Avec la commande `tabul`, trouver les effectifs d'apparition de chaque valeurs.*

1.2 Classes

On peut être amené à vouloir regrouper plusieurs modalités en « paquets » (par exemple, lorsque la série statistique prend un grand nombre de valeurs distinctes). En général, ces paquets sont des intervalles et on les appelle **classes** de la série.

Exemple 3

Si on reprend la série de l'exemple précédent :

Classes	$[1,2]$	$]2,6]$	$]6,8]$
Effectifs	4	1	3

La commande `dsearch`

La commande `dsearch` permet de regrouper une série statistique par classe. La syntaxe est la suivante

```
[ind,occ,non]=dsearch(X,C)
```

où

- X est un vecteur qui représente la série statistique ;
- C $= [c_1, c_2, \ldots, c_k]$ est un vecteur qui définit les classes : la première classe sera l'intervalle $[c_1, c_2]$ et pour $j \geqslant 2$ la j-ième classe sera $]c_j, c_{j+1}]$;
- `ind` est un vecteur de même longueur que X qui indique le numéro de la classe à laquelle appartient chaque élément de X ;
- `occ` est un vecteur de même longueur que C qui indique l'effectif de chaque classe ;
- `non` est le nombre d'éléments de X qui ne sont dans aucune classe.

Exemple 4

Recopiez le script suivant et exécutez-le.

```
L = [ 7 , 2 , 8 , 5 , 2 , 5 , 10 , 5 , 5 , 7 , 4 , 7 , 2 , 8 , 7 ]
c=linspace(2,10,5)
[ind,occ]=dsearch(L,c)
```

1. `c=[ , , , , ]` *, cela signifie que les classes qu'on considère sont*

   $I_1 =$, $I_2 =$, $I_3 =$, $I_4 =$

2. `ind=[ , , , , , , , , , , , , , , ]` *, cela signifie que*
   - `X(1)` *est dans la classe* _____
   - `X(2)` *est dans la classe* _____
   - *etc...*
3. `occ=[ , , , , ]` *, cela signifie que*
   - *l'effectif de la classe* I_1 *est* _____
   - *l'effectif de la classe* I_2 *est* _____
   - *etc...*

Exercice 2

1. *Avec la commande* `grand`*, définir un vecteur ligne* L *de* 100 *nombre aléatoire suivant une loi normale centrée réduite.*
2. *Que font les commandes* `min(L)` *et* `max(L)` *?*
3. *Avec la commande* `linspace` *construire un vecteur ligne* c *de* 11 *nombres régulièrement espacés entre* `min(L)` *et* `max(L)`*. On définit ainsi* 10 *classes, lesquels ?*
4. *Avec la commande* `dsearch` *déterminer l'effectif de chaque classe.*

1.3 Effectifs/Fréquences cumulées croissantes

Définition 2 (Effectifs cumulées croissants, fréquences cumulées croissantes)

Soit L une série statistique (quantitative) dont les modalités sont **rangées par ordre croissant**.

- L'**effectif cumulé croissant** d'une modalité est la somme des effectifs des modalités qui lui sont inférieures ou égales.
- La **fréquence** d'une modalité est le quotient $\frac{\text{Effectif de la modalité}}{\text{Effectif de la série}}$.
- La **fréquence cumulée croissante** d'une modalité est la somme des fréquences des modalités qui lui sont inférieures ou égales.

Exercice 3

1. *Recopier le script suivant*

```
n=input('entrez un nombre entier naturel n:')
x=floor(grand(1,n,'nor',5,1)
```

   Expliquer la deuxième ligne.
2. *Écrire une commande permettant d'obtenir la liste des effectifs cumulés croissants de la série* `x`.
3. *Tracer la ligne brisée représentant les fréquences cumulées croissantes en fonction des valeurs de la séries.*

1.4 Représentation graphique

Diagramme en barres

Pour représenter une série statistique , on peut utiliser un **diagramme en barres** : on place les modalités sur un axe horizontale et on dresse à la verticale de chacune une barre de hauteur égale à son effectif (ou sa fréquence).
Avec Scilab, on utilise la commande

```
bar(L,n)
```

où `L` est la liste des valeurs distinctes rangées par ordre croissant et `n` la liste des effectifs.

Exercice 4

Reprendre la série statistique de l'exercice précédent et tracer le diagramme en barre correspondant.

Diagramme circulaire

Pour représenter une série statistique , on peut utiliser un **diagramme circulaire** : chaque modalité (ou classe) est représentée par un secteur angulaire dont l'angle est proportionnel à l'effectif de la modalité (ou classe).
Avec Scilab, on utilise la commande

```
pie(n,['x1',...,'xp'])
```

où `n` est `n` la liste (de longueur p) des effectifs de chaque modalité distinctes et `x1`,..., `xp` les légendes.

Exercice 5

On considère la série statistique suivante

$$2,11,7,2,15,4,5,5,5,13,5,15,7,7,8,10,10,10,11,13,7,2,15,15$$

1. *Créer une liste* `L` *qui contient la série statistique.*
2. *Avec la commande* `tabul`*, construire la liste* `n` *des effectifs de chaque valeur distincte.*
3. *Créer le diagramme circulaire de la série (comme légende on prendra les modalités).*

Histogramme

Pour représenter une série statistique regroupée par classes, on peut utiliser un **histogramme**. Si les classe sont $]c_i, c_{i+1}]$, on place les c_i sur un axe horizontale et, pour chaque classe, on trace un rectangle dont la base est $]c_i, c_{i+1}]$ et dont l'aire est proportionnelle à l'effectif de la classe.
Avec Scilab, on utilise la commande

```
histplot(n,L)  ou  histplot(c,L)
```

où `L` est la série statistique, `n` le nombre de classe ou `c` le vecteur ligne définissant les classes.

Exercice 6

1. *Écrire une fonction qui prend en argument des entiers n et p, qui simule un échantillon de taille n d'une variable aléatoire suivant une loi normale centrée réduite $\mathcal{N}(0,1)$ et qui trace un histogramme de cette série avec p classes.*
2. *Tester la fonction pour $n = 10\,000$ et $p = 14$.*

2 Indicateurs de position

2.1 Mode

Définition 3 (Mode)

On appelle **mode** d'une série statistique toute valeur de la série correspondant au plus grand effectif (il peut y en avoir plusieurs).

Exemple 5

Pour la série $[7,2,8,5,2,5,10,5,5,7,4,7,2,8,7,7,2,8,5,2,5,10,5,5,7,4,7,2,8,7]$, 5 *et* 7 *ont le plus grand effectif (*4 *pour les deux). Les modes de la série sont donc* 5 *et* 7.

Exercice 7

On définit une série statistique avec les instructions suivantes

```
X=grand(1,100,'nor',0,4)
X=abs(floor(X))
```

1. *Utiliser la commande* `tabul` *pour tirer la liste par ordre croissant et obtenir l'effectif de chaque modalité.*
2. *Consulter l'aide de la commande* `max` *et en déduire une suite d'instructions permettant de déterminer le mode de la série.*

2.2 Moyenne

Définition 4 (Moyenne)

1. Soit $x = (x_i)_{1\leqslant i\leqslant n}$ une série statistique brute. La moyenne de la série est le nombre $\bar{x}$ défini par
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
2. Si la série est groupée par modalités $(y_i, n_i)_{1\leqslant i\leqslant p}$, alors
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i y_i.$$

Exercice 8

1. *Consulter l'aide de la fonction* `mean`.
2. *Calculer à l'aide de la fonction* `mean` *la moyenne de la série* X *de l'exercice précédent.*
3. *Définir une fonction* `moyenne` *qui prend en argument deux listes de même taille et qui renvoie la moyenne de la série statistique dont les modalités sont les éléments de la première liste et les effectifs ceux de la deuxième liste.*

2.3 Médiane

Définition 5 (Médiane)

On appelle **médiane** d'une série statistique tout nombre réel tel qu'au moins la moitié de l'effectif total ait une modalité inférieure ou égale et au moins la moitié de l'effectif total ait une modalité supérieure ou égale. En pratique, on trie la série par **ordre croissant** puis

- si l'effectif total N est pair, on prendra pour médiane la moyenne de la $\frac{N}{2}$-ième et $\left(\frac{N}{2}+1\right)$-ième valeur ;
- si l'effectif total N est impair, on prendra pour médiane la moyenne de la $\frac{N+1}{2}$-ième valeur.

Exercice 9

On considère la série suivante :

$$[7,2,8,5,2,5,10,5,5,7,4,7,2,8,7,7,2,8,5,2,5,10,5,5,7,4,7,2,8,7]$$

1. *Déterminer la médiane à la main.*
2. *Vérifier votre résultat à l'aide de la commande* `median`.

2.4 Quantiles

Définition 6 (Quartiles, déciles)

- Le **premier quartile** d'une série statistique est la plus petite valeur dont l'effectif cumulé croissant est supérieur ou égal à 25% de l'effectif total. Le **troisième quartile** d'une série statistique est la plus petite valeur dont l'effectif cumulé croissant est supérieur ou égal à 75% de l'effectif total.
- Pour $k \in \{1,\dots,p\}$, le k-ième décile est la plus petite valeur dont l'effectif cumulé croissant est supérieur ou égal à $10k$% de l'effectif total.

Exercice 10

Écrire une fonction qui prend en argument une série statistique brute et qui renvoie la valeur du premier et troisième quartile. Déterminer les quartiles de la série de l'exercice précédent.

Exercice 11

On considère les commandes :

```
n=input('entrez la valeur de n:')
x=grand(1,n,'poi',6)
```

1. *Écrire les commandes permettant de tracer le polygone des fréquences cumulées croissantes, c'est-à-dire la ligne brisée reliant les points dont l'abscisse est une modalité de* `x` *et l'ordonnée la fréquence cumulée croissante de cette modalité.*
2. *En déduire graphiquement la valeur des quartiles. Vérifier avec la fonction définie à l'exercice précédent.*

3 Indicateurs de dispersion

3.1 Étendue

Définition 7 (Étendue)

On appelle **étendue** d'une série statistique l'écart entre la plus grande et la plus petite modalité.

3.2 Variance et écart-type empiriques

Définition 8 (Variance et écart-type empiriques)

1. • Soit $x = (x_i)_{1\leqslant i\leqslant n}$ une série statistique brute. La **variance empirique** de la série est le nombre $V(x)$ défini par
$$V(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
• Si la série est groupée par modalités $(y_i, n_i)_{1\leqslant i\leqslant p}$, alors
$$V(x) = \frac{1}{n-1}\sum_{i=1}^{p} n_i(y_i - \bar{x})^2.$$
2. L'**écart-type empirique** de la série est la racine carrée de la variance empirique. On le note $\sigma(x)$.

⚠ Il ne faut pas confondre avec la formule de la variance "classique" où on divise par n et non $n-1$! La raison de ce changement sera explicité au second semestre lorsque nous parlerons d'estimateurs.

Exercice 12

Pour déterminer la variance empirique d'une série statistique brute `x`*, on utilise la commande* `variance(x)`.

1. *Calculer la variance empirique de la série statistique de l'exercice 9.*
2. *Écrire une fonction* `ecart_quadratique` *qui prend en argument une liste* $(x_i)_{1\leqslant i\leqslant n}$ *et qui renvoie le nombre* $\sum_{i=1}^{n}(x_i - \bar{x})^2$.
3. *Comparer les résultats des commandes*
   (a) `ecart_quadratique(x)/(length(x)-1)` *et* `variance(x);`
   (b) `ecart_quadratique(x)/length(x)` *et* `variance(x,'*',mean(x))`.

Exercice 13

Pour déterminer l'écart-type empirique d'une série statistique brute `x`*, on utilise la commande* `stdev(x)`.

1. *Calculer l'écart-type empirique de la série statistique de l'exercice 9.*
2. *Comparer les résultats des commandes*
   (a) `sqrt(variance(x))` *et* `stdev(x);`
   (b) `sqrt(variance(x,'*',mean(x)))` *et* `stdev(x,'*',mean(x))`.

Exercice 14

1. *On considère la série statistique*

```
X=grand(1,10000,'nor',25,3).
```

   Déterminer la moyenne, la variance et l'écart-type empiriques de `X`.
2. *Même question avec les séries* `X=grand(1,10000,'nor',25,sqrt(5))` *puis* `X=grand(1,10000,'nor',25,sqrt(37))`.
3. *Que remarque-t-on ?*

3.3 Écart inter-quantile

Définition 9 (Écart interquartile)

L'**écart interquartile** d'une série statistique est la différence entre le troisième quartile et le premier quartile.

4 Exercices

Exercice 15

1. *On considère la série statistique*

```
X=grand(1,10000,'nor',25,3).
```

   (a) *Écrire les commandes permettant le calcul de la médiane et de la moyenne de cette série.*

   (b) *Tracer l'histogramme en* 40 *classes de même amplitude pour des valeurs allant de* 15 *à* 35. *En déduire le mode de la série.*

2. *Mêmes questions avec un histogramme en* 15 *classes de même amplitude pour des valeurs allant de* 5 *à* 20 *avec la série* `X=grand(1,10000,'nor',13,sqrt(5))`.

3. (a) *Que remarquez-vous?*

   (b) *Rappeler la densité et l'espérance d'une variable aléatoire* X *suivant une loi* $\mathcal{N}(m,\sigma^2)$.

   (c) *Démontrer l'observation faite à la question 3.(a) concernant le mode.*

Exercice 16

Commencer par télécharger le fichier « TP2_data_exo16 » puis rentrer la commande suivante dans le fichier .sce, en remplaçant les ... par le chemin d'accès du fichier :

```
L=read('...\TP2_data_exo16',1,100).
```

Exécuter le fichier .sce : la variable L *contient un échantillon de taille* 100 *des salaires nets mensuels de salariés.*

1. *Quel est le salaire minimal/maximal de cet échantillon? En déduire l'étendue.*
2. *Déterminer le salaire médian puis le salaire moyen. Commenter.*
3. *On regroupes les données dans les classes suivantes*

$$I_1 = [0,1200] \quad , \quad I_2 =]1200,1600] \quad , \quad I_3 =]1600,2000] \quad , \quad I_4 =]2000,2400]$$
$$I_5 =]2400,2800] \quad , \quad I_6 =]2800,3200] \quad , \quad I_7 =]3200,5000]$$

   Déterminer l'effectif de chaque classe. A quelle classe appartient la 52*-ième valeur?*

4. *Représenter l'histogramme avec les classes définies ci-dessus.*
5. (a) *Écrire une fonction qui prend en argument une série statistiques sous forme de liste et qui renvoie le premier et neuvième décile de la série.*

   (b) *En déduire le premier et neuvième décile de la série* `L`.

Exercice 17

1. *Télécharger le fichier « TP2_data_exo17 ».*
2. *Construire le vecteur ligne* `classes`$= [0,1200,1300,\ldots,8600,8700]$.
3. *La commande* `effectifs=read('...\TP2_data_exo17',1,77)` *permet, en remplaçant les ... par le chemin d'accès du fichier, d'importer un vecteur ligne de même taille que* `classes` *qui contient des données de l'INSEE :*
   - *pour* $i = 1,\ldots,76$, `effectifs(i)` *correspond au nombre de salariés du privé dont le salaire mensuel net en équivalent temps plein en 2017 est compris entre* `classes(i)` *et* `classes(i+1)` *;*
   - `effectifs(77)` *correspond au nombre de salariés du privé dont le salaire mensuel net en équivalent temps plein en 2017 est supérieur à* `classes(77)`.

     (a) *Combien de salariés du privé ont un salaire compris entre* 2600 *et* 2700 *euros en* 2017*?*

     (b) *Avec la commande* `bar(classes,effectifs)`, *afficher la répartition des salaires dans le privé en* 2017 *: on obtient un diagramme en barres où la barre représentant une classe est alignée sur la*

borne inférieure de la classe.

5 Éléments du programme officiel

1. Les commandes suivantes, rencontrées durant ce TP, ainsi que leurs arguments sont exigibles :

   `cumsum` , `mean` , `max` , `min`

2. Nouvelle commande rencontrée : `grand`.
3. Compétences mises en jeu :
   - C1 : produit et interpréter des résumées numériques et graphiques d'une série statistiques ou d'une loi.
   - C6 : Porter un regard critique sur les méthodes d'estimation et de simulation.
4. Exigibles de première année :
   - `bar`, `histplot`