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**.

On peut représenter une série statistique de deux manières :

1. en donnant la liste « brute » des valeurs prises par X : $[X(\omega), \omega \in \Omega]$;
2. 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 1

Si on considère la population française, alors

1. *le caractère qui donne le revenu annuel est-il un caractère quantitatif ou qualitatif ?*
2. *Le caractère qui donne la catégorie socio-professionnelle est-il un caractère quantitatif ou qualitatif ?*

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 brute et renvoie une matrice à deux colonnes :

1. la première colonne contient les valeurs distinctes de la série statistique classées dans l'ordre décroissant (par défaut) ou croissant (option) ;
2. la seconde colonne contient les effectifs correspondants.

Pour plus de détails, voir l'aide.

► Consulter l'aide Scilab concernant la commande `grand` en tapant dans la console :

```
help grand
```

► Avec la command `grand`, créer une série statistique contenant les résultats d'une simulation de 100 le lancers d'un dé à six faces équilibré. Recopier la ligne de commande :

► Avec la commande `tabul`, trouver les effectifs d'apparition de chaque valeurs. En déduire le tableau d'effectifs :

1.2 Classes

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, on peut regrouper les valeurs en plusieurs classes. Par exemple :

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;
- $C = [c_1, c_2, \dots, c_k]$ 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 dont la longueur est égale à celle C moins un et qui indique l'effectif de chaque classe;
- `non` est le nombre d'éléments de X qui ne sont dans aucune classe.

► 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)
```

► Compléter :

```
C =
```

► En déduire les classes I_1, I_2, I_3 et I_4 :

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

► Compléter :

```
ind =
```

► En déduire dans quelle classe est `L(1)` puis `L(2)`.

► Compléter :

```
occ =
```

► En déduire l'effectif de la classe I_1, I_2, ...

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 $\dfrac{\text{Effectif de la modalité}}{\text{Effectif de la série}}$.
- La **fréquence cumulée croissante** d'une modalité est la somme des fréquences des modalités qui lui sont inférieures ou égales.

► Recopier le script suivant :

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

Expliquer ce que fait la deuxième ligne :

► Écrire une commande permettant d'obtenir la liste des effectifs cumulés croissants de la série `x` (on pourra utiliser la commande `cumsum` et consulter l'aide) :

- Tracer la ligne brisée représentant les fréquences cumulées croissantes en fonction des valeurs de la séries. Recopier la commande :

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 horizontal 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.

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

On considère la série statistique suivante

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

- Créer une liste `Liste` qui contient la série statistique.
- Avec la commande `tabul`, construire la liste `L` des valeurs distinctes rangées par ordre croissant et la liste `n` des effectifs de chaque valeur distincte.
- Tracer le diagramme en barres puis le diagramme circulaire de la série (pour le diagramme circulaire, on prendra comme légende les modalités). Recopier les lignes de commandes :

Histogramme

Pour représenter une série statistique regroupée par classes, on peut utiliser un **histogramme**. Si les classes sont $]c_i, c_{i+1}]$, on place les c_i sur un axe horizontal 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 classes ou `c` le vecteur ligne définissant les classes.

- Avec la commande `grand`, simuler un échantillon de taille 10 000 d'une variable aléatoire suivant une loi normale centrée réduite $\mathcal{N}(0,1)$ (**on n'oubliera pas de mettre un point-virgule à la fin de la ligne de commande pour ne pas afficher la liste!**).
- Tracer l'histogramme de cette série avec 14 classes et recopier les lignes de commandes :

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 4

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 (8 pour les deux). Les modes de la série sont donc* 5 *et* 7*.*

On définit une série statistique avec les instructions suivantes :

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

- Utiliser la commande `tabul` pour trier la liste par ordre croissant et obtenir l'effectif de chaque modalité.
- Consulter l'aide de la commande `max` et en déduire une suite d'instructions permettant de déterminer un mode de la série. Recopier ces instructions :

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
$$\bar{x}=\frac{1}{n}\sum_{i=1}^{p}n_i y_i.$$

- Consulter l'aide de la fonction `mean`.
- Calculer à l'aide de la fonction `mean` la moyenne de la série X définie précédemment.

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.

On considère la série suivante :

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

► Déterminer la médiane à la main.

► 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.

Scilab ne possède pas de commande donnant les quartiles d'une série au sens de la définition ci-dessus. Nous allons donc la créer en complétant la fonction suivante :

```
function res=quartiles(L)
    Tableau_croissant = ____________
    n = length(L)
    ecc =  ____________
    i = 1
    while ecc(i) < 25*n/100
        i = i+1
    end
    j = 1
    while ecc(j) < 75*n/100
        j = j+1
    end
    res = [____________ , ____________]
endfunction
```

► Compléter et recopier la fonction `quartiles` ci-dessus qui prend en argument une série statistique brute `L` et qui :
- détermine le tableau d'effectifs avec les modalités en ordre croissant et le stocke dans `Tableau_croissant`;
- détermine la liste des effectifs cumulés croissants et la stocke dans `ecc`;
- retourne la liste `res` constituée du premier quartile et du troisième quartile.

► Recopier les commandes et expliquer chaque ligne.

```
n = input('entrez la valeur de n:')
x = grand(1,n,'poi',6)
x_tab = tabul(x,'i')
ecc_x = cumsum(x_tab(:,2))
plot(x_tab(:,1), ecc_x)
```

► En déduire graphiquement la valeur des quartiles et vérifier avec la fonction `quartile`.

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)

Soit $x = (x_i)_{1\leqslant i\leqslant n}$ une série statistique brute.

1. • La **variance empirique** de la série est le nombre $\mathrm{V}(x)$ défini par

$$\mathrm{V}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

   • Si la série est groupée par modalités $(y_i, n_i)_{1\leqslant i\leqslant p}$, alors

$$\mathrm{V}(x) = \frac{1}{n-1}\sum_{i=1}^{p}n_i(y_i - \bar{x})^2.$$

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.

► Pour déterminer la variance empirique d'une série statistique brute `x`, on utilise la commande `variance(x)`. Calculer la variance empirique de la série statistique suivante :

x = [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]

► Compléter la fonction `ecart_quadratique` qui prend en argument une liste $(x_i)_{1\leqslant i\leqslant n}$ pour qu'elle renvoie le nombre $\sum_{i=1}^{n}(x_i - \bar{x})^2$.

```
function res=ecart_quadratique(x)
    moyenne = mean(x)
    res = 0
    for i = 1:length(x)
        res = res+ ________
    end
endfunction
```

► Comparer les résultats des commandes

1. `ecart_quadratique(x)/(length(x)-1)` et `variance(x)` ;
2. `ecart_quadratique(x)/length(x)` et `variance(x,'*',mean(x))`.

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

► Comparer les résultats des commandes

1. `sqrt(variance(x))` et `stdev(x)` ;
2. `sqrt(variance(x,'*',mean(x)))` et `stdev(x,'*',mean(x))`.

On considère la série statistique

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

- Déterminer la moyenne, la variance et l'écart-type empiriques de `X`.
- Même question avec les séries `X=grand(1,10000,'nor',25,sqrt(5))` puis `X=grand(1,10000,'nor',25,sqrt(37))`.
- 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 Exercice

Exercice 1

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* $\mathcal{N}(m,\sigma^2)$.

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`, `tabul`.
3. Compétences mises en jeu :
   - C1 : produire et interpréter des résumés numériques et graphiques d'une série statistique ou d'une loi.
   - C6 : Porter un regard critique sur les méthodes d'estimation et de simulation.
4. Exigibles de première année :
   - `bar`, `histplot`