

STATISTIQUES À UNE VARIABLE

Le but de ce TD est de faire des statistiques descriptives univariées.

I Commandes scilab de bases

Pour pouvoir tester les fonctions utilisées commençons par créer trois séries de données

```
serie1=grand(1000,1,"uin",0,10)
serie2=floor(grand(1000,1,"nor",0,1))
serie3=[100,89,67,103,78,123,90,98,104,126,100,89,100,68,134]
```

Vous pouvez modifier ces séries à votre guise.

I.1 Mesure de l'étendue

L'**étendue** est la différence entre la plus grande et la plus petite des valeurs de la série. C'est une façon de mesurer la dispersion de la série statistique (X_i) de donnée. Elle est définie par $\max_i(X_i) - \min_i(X_i)$. On peut utiliser les fonctions scilab `max` et `min`

Exemple :

```
disp(max(serie1)-min(serie1))
```

Exercice : Faire ce même calcul pour les deux autres séries.

I.2 Moyenne et médiane

Ces deux indicateurs de positions peuvent être calculés par scilab à l'aide des commandes `mean` et `median`.

Exemple : Pour la série 1

```
disp('moyenne et mediane de la serie 1')
disp(mean(serie1)),disp(median(serie1))
```

Exercice : Faire le même travail pour les séries 2 et 3. Pour les deux premières séries le résultat vous semble-t-il naturel ?

Exercice : Trouver une série (en la remplissant à la main avec une moyenne et une médiane « très différentes »).

I.3 Variance et écart type.

L'écart type qui est égale à la racine de la variance est un indicateur de dispersion peut être calculé par scilab à l'aide de la commande `stdev` (pour *standart deviation*).

Exemple : Pour la série 1 :

```
disp('ecart type et variance de la serie 1')
disp(stdev(serie1)),disp(stdev(serie1)^2)
```

Exercice : Faire le même travail pour les séries 2 et 3. Pour les deux premières séries, le résultat vous semble-t-il naturel ?

II Savoir faire

II.1 Utilisation de la fonction `tabul` : calcul des effectifs de chaque modalité

Commençons par un exemple

Exemple :

```
m=tabul(serie3)
disp(m)
```

```
134.    1.
126.    1.
123.    1.
104.    1.
103.    1.
100.    3.
 98.    1.
 90.    1.
 89.    2.
 78.    1.
 68.    1.
 67.    1.
```

On obtient une liste des valeurs prises par la série (**modalité**) et pour chaque valeur l'effectif de cette modalité.

Exercice 1.

Reprendre le même exemple en rajoutant l'option "i" c'est à dire **tabul**(serie3,"i"). De quoi "i" est il l'abréviation ?

Combiné à la fonction d'affichage **bar**, on peut obtenir le graphe de répartition de la série.

Exemple :

```
m=tabul(serie3,"i")
bar(m(:,2))
```

Explication : m est une matrice à deux colonnes, dans la première colonne se trouve les modalités de la série et dans la deuxième colonne se trouve les effectifs. C'est cette colonne qui nous intéresse m(:,2) permet de « l'extraire ».

Exemple :

```
bar(m(:,1),m(:,2))
```

Exercice 2.

Faire le même travail sur les deux autres séries

Remarque : Parfois on n'obtient pas un résultat satisfaisant, essayez le script suivant :

```
serie2p=grand(1000,1,"nor",0,3)
m=tabul(serie2p,"i")
bar(m(:,2))
```

Pourquoi obtien t'on un résultat si peu clair ?

II.2 Utilisation de la fonction dsearch : répartition en classes

Pour remédier a ce problème précédent on peut utiliser **dsearch** pour trier les valeurs en **classes**

Exemple :

```
[q,eff]=dsearch(serie2p,[-10 -5 0 5 10])
disp(eff)
```

Il faut lire le résultat de la manière suivante . Dans la classe [-5; 10[l'effectif est de 41 dans la classe [-5; 0[, l'effectif est de 465,...

Si on fait afficher le tableau **q**, on obtient une longue liste, l'entier *i* de cette liste indique le numéro de chaque classe à laquelle appartient l'élément *i* de la série statistique.

Plutôt que de donner explicitement toutes les limites des classes, on peut utiliser **linspace** pour créer les classes.

Exemple :

```
classes=linspace(-10,10,20)
[q,eff]=dsearch(serie2p,classes)
bar(eff)
```

Ou mieux

```
classes=linspace(-10,10,5)
[q,eff]=dsearch(serie2p,classes)
bar(classes(1:4),eff)
```

II.3 La fonction histplot

Cette fonction permet de simplifier le tracer de certain graphe.

Exemple :

```
histplot(10,serie2p)
```

Faire varier l'entier. Que représente t il ?

Exercice 3.

Tracer des histogrammes liés aux séries exemples.

III Exercices

III.1 Fréquence

Pour les trois séries statistiques représenter l'histogramme des **fréquences empiriques**. On rappelle que pour une valeur donnée ou pour une classe

$$f_i = \frac{n_i}{n_{total}}$$

où n_i est l'effectif de la modalité et n_{total} est l'effectif total.

III.2 Étude d'une série

On étudie la série suivante :

2, 6, 6, 5, 1, 1, 3, 3, 1, 1, 1, 1, 1, 4, 2, 3, 4, 5, 2, 2, 3, 1, 1, 3, 1, 4, 2, 2, 1, 3.

1. Représenter cette série avec un diagramme en barres.
2. Calculer, moyenne médiane et variance.
3. Classer les individus en trois classe $[0; 3[$, $[3; 5[$ et $[5; +\infty[$
4. Pouvez vous deviner quelle loi à servit à générer cette série.

IV Utiliser csvRead : Une série temporelle

On va étudier l'évolution du prix du café à la bourse de New-York. Vous devez télécharger le fichier "cafe.csv" qui est disponible sur le site dans la rubrique informatique/données et utiliser la commande

```
prix=csvRead('[chemin]\cafe.csv',';', 'double');
```

où vous devez remplacer [chemin] par l'emplacement de ce fichier.

Chaque nombre de cette série représente le prix du café à la fin d'un mois.

1. Calculer les caractéristiques de cette série.
2. À l'aide de dsearch, classer ces valeurs en 10 classes de largeur égale.
3. Représenter un histogramme de ces classes.
4. Avec plot représenter le prix en fonction du temps.
5. **Moyenne glissante à 12 mois.**

Pour "lisser" les variations rapides, on va calculer la moyenne sur 12 mois.

Si prix() est la suite des prix on définit cette moyenne par

$$\forall i \in ? \quad M(i) = \frac{1}{12} \sum_{k=0}^{11} \text{prix}(i - k)$$

- (a) Pouvez vous expliquer les termes de "moyenne" et "glissante"?
- (b) Recopier et compléter le programme suivant pour qu'il calcule cette moyenne.

```
L=length(prix)
```

```
M=zeros(L)
```

```
N=12
```

```
for i= :  
    M(i)=  
    for j=0:N-1  
        M(i)=
```

```
end  
  
end  
  
M=
```

- (c) Représenter le tableau obtenu.
- (d) Modifier le programme pour avoir la moyenne glissante à 6 mois.

V Le cours du NASDAQ

Le fichier NASDAQ.csv contient le cours de fermeture pour chaque jour depuis 1971.

1. À l'aide de csvRead charger ces en mémoire, on utilisera un tableau N.
2. Vous obtenez un tableau à deux colonnes, celle qui intéresse est la deuxième, pour l'isoler on peut utiliser la commande NAS=N(:,2).
3. À l'aide de la fonction plot afficher l'évolution des cours.
4. En s'inspirant de l'exercice précédent, écrire un programme qui affiche la moyennne glissante sur un mois et sur un an.
5. On cherche à connaitre quelle éa été la plus longue période de hausse continue. Compléter le programme suivant pour calculer cette période

```
L=length(NAS)  
plus_longue_hausse =0  
hausse=0  
for i=1:.....  
    if ..... then  
        hausse =hausse+1 //la hausse continue  
    else // on a fini une periode de hausse  
        plus_longue_hausse=max(.....,.....)  
        hausse=0  
    end  
end
```

```
disp(plus_longue_hausse)
```

6. S'inspirer du programme précédent pour calculer la plus longue période de baisse.
7. Écrire un programme qui calcule le nombre de joursou l'indice à augmenter, et le nombre de jours ou il a diminué

VI Une série économique

Le fichier salairesWH contient la série des salaires des personnels de la maison blanche en 2010.

1. Télécharger ce fichier et charger le dans Scilab.
2. Trouver le salaire minimal, le salaire médian, le maximum et le minimum.
3. Représenter cette série.
4. Trouver le salaire qui est gagné par le plus de personnes. (On pourra aller lire l'aide de la fonction `max` et utiliser `tabul`).