TP 1 : Statistiques descriptives univariées

L'étude statistique ressemble beaucoup à la théorie des probabilités. La différence fondamentale entre statistiques et probabilités est la suivante : **en probabilités**, on cherche à anticiper l'avenir : calculer nos chances de gagner à un jeu avant de jouer...etc...c'est à dire obtenir des informations avant d'effectuer une expérience aléatoire, alors qu'**en statistiques**, on cherche à tirer des informations après avoir effectué des expériences, souvent après avoir répété plusieurs fois la même expérience.
En probabilités on parle de calculs **théoriques** (induisent (anticipent) les résultats) alors qu'en statistiques, on parle de calculs **empiriques** (déduits de l'expérience).

1. Vocabulaire

• Une étude statistique concerne un ensemble Ω appelé **population** dont les éléments sont appelés **individus**.
L'étude consiste en l'analyse d'une **variable statistique X** observée sur les individus, appelée le **caractère observé**.

Une variable statistique X est **discrète** si les valeurs prises par X sont isolées les unes des autres, elle est dite **continue** si X peut prendre toutes les valeurs d'un intervalle de $\mathbb{R}$.

Exemples

1. Si l'on fait une étude statistique du nombre de buts par match durant la coupe du monde 2010
   — La **population** Ω étudiée est .....
   — Le **caractère observé** X est .....
   — Le caractère X est **discret** car il prend ses valeurs dans $\mathbb{N}$ qui est un ensemble discret.
2. Si l'on fait une étude statistique du taux de chômage des pays mondiaux :
   — La **population** Ω étudiée est .....
   — Le **caractère observé** X est .....
   — Le caractère X est **continu** car il prend ses valeurs dans $[0; 1]$ qui est un ensemble continu.

Contrairement aux probabilités, nous allons observer les valeurs prises par la variable X sur une grande population (i.e. simuler un grand nombre de fois la variable aléatoire X) et obtenir des informations sur X grâce à ces simulations (loi empirique : tableau des fréquences, moyenne empirique...) au lieu de l'étudier théoriquement avant de faire une expérience.

• Pour obtenir un renseignement exact sur X, il faudrait étudier tous les individus de la population. Lorsque cela n'est pas possible, on étudie seulement les individus d'une partie E de Ω appelée **échantillon observé**. L'entier $n = Card(E)$ est alors la **taille de l'échantillon**.

Exemple
On souhaite étudier le nombre X de buts marqués lors d'un match de foot. Seulement, nous n'avons ni le temps, ni l'envie (!) d'étudier tous les matchs de l'histoire, nous allons donc faire une étude sur un échantillon de 20 derniers matchs regardés :

1. Taper l'instruction suivante donnant la liste des nombres de buts marqués lors de ces matchs
   ```
   --> X=[3 1 5 3 2 7 0 1 0 3 2 4  4 0 3 3 2 5 3 1 ]
   ```
2. Taper l'instruction suivante :
   ```
   M=tabul(X,"i")
   ```
   Que fait l'instruction tabul ?
3. On appelle **modalités** de X la liste $x_1 < x_2 < ... < x_p$ des différentes valeurs prises par X, rangées dans l'ordre croissant. Taper une instruction en Scilab qui enregistre les modalités de X dans une matrice ligne x.
4. On note n_i l'**effectif** de la modalité x_i.
   Taper une instruction en Scilab qui enregistre la liste $[n_1, n_2, ..., n_p]$ des effectifs dans une matrice ligne n.
5. Quelle valeur renvoit l'instruction `sum(n)` ? Pourquoi ?

2. Description d'une série statistique et représentation graphique

a) Cas des variables discrètes

• On appelle **fréquence** de la valeur x_i le réel $f_i = \frac{\text{effectif}}{\textbf{effectif total}} = \frac{n_i}{n}$.
En pratique, c'est le taux de la population dont le caractère X prend la valeur x_i .

Remarque : on a $\sum_{i=1}^{n} f_i = 1$.

• On appelle **fréquence cumulée** de la valeur x_i le réel $p_i = \sum_{j \leq i} f_j$.
En pratique, c'est le taux de la population dont le caractère X prend une valeur inférieure ou égale à x_i.

Exemple

Nombre de buts	0	1	2	3	4	5	7
Effectif (nombre de matchs)	3	3	3	6	2	2	1
Fréquence							
Fréquence cumulée							

1. Remplir "à la main" le tableau ci-dessus.
2. Que fait l'instruction suivante ?

```
--> f=n / sum(n)
```

3. Que fait l'instruction suivante ?

```
--> fcc=cumsum(f)
```

4. **Représentation graphique :**
On peut représenter la série statistique d'une variable discrète à l'aide d'un diagramme en bâton des fréquences :

```
--> bar(x,f)
```

ou du diagramme en bâton des fréquences cumulées :

```
-->scf(1), bar(x,fcc)
```

Remarque : FONDAMENTALE
Les notions suivantes se correspondent en probabilités et en statistiques :

X variable aléatoire	X variable statistique
probabilité $P(X = x_i)$	fréquence f_i
fonction de répartition F_X	fréquence cumulée p_i

b) Cas général

Lorsqu'une variable X est continue, ou que le nombre de valeurs prises par une variable discrète X est trop grand, on répartit ces valeurs en classes.
La série statistique est alors caractérisée par ses **classes** (intervalles du type $]a_i, a_{i+1}]$ où les a_i sont croissants) et l'effectif n_i de chaque classe $]a_i, a_{i+1}]$.

Représentation graphique :
- On peut toujours représenter la série à l'aide d'un diagramme en bâton des effectifs (ou des fréquences) grâce à la fonction `bar` : on choisit alors comme modalités les centres c_i des classes.
- Mais le plus représentatif est l'histogramme des fréquences de la série statistique : on crée des rectangles de base $]a_i, a_{i+1}]$ et dont l'aire est proportionnelle à f_i.
Ceci se fait grâce à la fonction `histplot` :

Exemple
On souhaite représenter avec Scilab l'histogramme associé aux nombres de buts marqués.
• Première possibilité :

```
--> scf(2), histplot(7, X)
// hisplot répartit alors les nombres de buts en 7 classes de même longueur (c'est-à-dire ici [0,1],
//  ]1, 2], ]2,3], ]3,4], ]4,5], ]5,6] et ]6,7]) et calcule les fréquences correspondant à chaque
// classe puis trace l'histogramme correspondant.
```

Prenez le temps de bien comprendre les valeurs de fréquences obtenues !

• Seconde possibilité :

```
--> classes=[ -0.5,1.5,2.5,3.5, 7.5]
--> scf(3),  histplot(classes, X)
// histplot calcule alors les fréquences d'employés dans les classes demandées c'est-à-dire ici [-0.5,1.5],
// ]1.5, 2.5], ]2.5,3.5] et ]3.5,7.5].
```

De même, prenez le temps de bien analyser le diagramme obtenu !

On remarquera que l'instruction `histplot` n'est pas intéressante quand il s'agit d'un nombre de buts, car X a peu de modalités. Mais lorsqu'il s'agira de la taille d'individus par exemple, il sera nécessaire de les répartir en classes de tailles (car la variable est continue).

3. Indicateurs de position

a) Moyenne

• Si X est une variable aléatoire discrète prenant les valeurs x_1, x_2, ...x_p,on rappelle que son espérance (moyenne théorique) est définie par :

$$E(X) = \sum_{i=1}^{p} x_i P(X = x_i)$$

Par analogie, la moyenne (empirique) d'une variable statistique discrète X, notée $\overline{x}$, est définie par :

$$\overline{x} = \sum_{i=1}^{p} x_i \times f_i$$

• Si X est continue, on remplace une classe par son centre c_i.

Exemple
En Scilab, on peut calculer une moyenne de plusieurs façons :
• Sur une série statistique brute (non pondérée), comme `X`, on peut utiliser la fonction `mean` :

```
-->mean(X)
```

• Sur une série statistique pondérée, c'est-à-dire dont on connait la liste `x` des modalités et la liste `f` de leurs fréquences respectives, on peut utiliser la formule de $\overline{x}$ donnée ci-avant :

```
-->sum(x.*f)
```

b) Mode

Si X est une variable discrète, on appelle **mode** la valeur (ou les valeurs) du caractère dont l'effectif (ou la fréquence) est le plus grand.
Dans le cas d'une variable continue, on parle de **classe modale**.

Exemple
A l'aide des différents graphiques obtenus, déterminer le(s) mode(s) de la série statistique des buts de la coupe du monde.

c) Médiane

• Si X est une variable discrète prenant n valeurs : $v_1 \leq v_2 \leq ... \leq v_n$ (valeurs non groupées (non pondérées)), on appelle **médiane** un nombre réel m tel qu'il y ait autant de valeurs inférieures ou égales à m que de valeurs supérieures ou égales à m.
Ainsi,
— si $n = 2k + 1$, alors $m = v_{k+1}$
— si $n = 2k$, alors on choisit en général $m = \frac{v_k + v_{k+1}}{2}$.

En Scilab, l'instruction `median(X)` calcule la médiane d'une série statistique X non pondérée.

• Si X est une variable discrète pondérée, la médiane m est la modalité vérifiant : moins de 50% de la population prend des valeurs strictement inférieures à m et moins de 50% de la population prend des valeurs strictement supérieures à m.

Remarque

La médiane ne s'intéresse qu'à la valeur "centrale", sans tenir compte des valeurs extrémales.
La moyenne est au contraire "déformée" par les valeurs extrémales.

d) Quantiles

Soit X une variable discrète prenant n valeurs : $v_1 \leq v_2 \leq ... \leq v_n$ (valeurs non groupées (non pondérées)),
• La médiane de la série statistique est aussi appelée deuxième quartile de la série.
• On appelle **premier quartile**, noté Q_1 la modalité telle que moins d'un quart de la population prend des valeurs strictement inférieures à Q_1 et moins de trois quart de la population prend des valeurs strictement supérieures à Q_1.
C'est donc la médiane de la sous-série statistique formée en ne gardant que première moitié des valeurs v_i rangées dans l'ordre croissant.

• On appelle **troisième quartile**, noté Q_3 la modalité telle que moins de trois quart de la population prend des valeurs strictement inférieures à Q_3 et moins d'un quart de la population prend des valeurs strictement supérieures à Q_3.
C'est donc la médiane de la sous-série statistique formée en ne gardant que la seconde moitié des valeurs v_i rangées dans l'ordre croissant.

Remarque

On peut définir de même les déciles et les centiles :

— pour $k \in [|1;99|]$, le k-ième centile est la valeur c_k de la série pour laquelle moins de k % de la population prend des valeurs strictement inférieures à c_k et moins de $100-k$% de la population prend des valeurs strictement supérieures à c_k.
— pour $k \in [|1;9|]$, le k-ième décile est la valeur d_k de la série pour laquelle moins des k dixièmes de la population prend des valeurs strictement inférieures à d_k, moins des $10-k$ dixièmes de la population prend des valeurs strictement supérieures à d_k.

Exemples

1. Déterminer la médiane m, les quartiles Q_1 et Q_3 et les trois premiers déciles des nombres de buts de la coupe du monde 2010.

2. L'instruction `Y=gsort(X, 'g', 'i')` (respectivement `Y=gsort(X, 'g', 'd')`) classe le vecteur `X` par ordre croissant (respectivement par ordre décroissant) . Taper l'instruction :

   ```
   --> Y= gsort(X,'g','i')
   ```

   et utiliser le résultat pour lire graphiquement la médiane et les quartiles de `X`.
   Vérifier vos résultats à l'aide de la fonction `median`.

4. Indicateurs de dispersion

a) Variance et écart-type empiriques

• Si X est une variable aléatoire discrète prenant les valeurs x_1, x_2, ...x_p,on rappelle que sa variance (théorique) est définie par :

$$V(X) = E\left((X-E(X))^2\right) = \sum_{i=1}^{p}(x_i - E(X))^2 P(X = x_i) \text{ (par théorème de transfert)}$$

Mais le théorème de Koenig-Huygens permet de la calculer également par la formule :

$$V(X) = E(X^2) - E(X)^2 = \sum_{i=1}^{p} x_i^2 P(X = x_i) - E(X)^2 \text{ (par théorème de transfert).}$$

Par analogie, si X est une variable statistique discrète, la **variance empirique** de X est le nombre réel positif :

$$V = \sum_{i=1}^{p} (x_i - \overline{x})^2 f_i$$

On a aussi (Koenig-Huygens) :

$$V = \sum_{i=1}^{p} x_i^2 f_i - \overline{x}^2$$

Comme en probabilités, la variance mesure la tendance qu'a X à prendre des valeurs qui s'écartent de la moyenne. Il s'agit d'une **variance empirique**, calculée à partir de l'échantillon observé.
• Si X est une variable continue, on remplace une classe par son centre c_i.
• On appelle **écart-type empirique** de X le réel $\sigma = \sqrt{V}$.

Exemple
• En Scilab, la variance empirique de données brutes est logiquement calculée par les instructions `-->v=mean((X-mean(X)).^2)` ou indifféremment par (Koenig-Huygens) `-->mean(X.^2)-mean(X)^2` :

```
--> v=mean(X.^2)-mean(X)^2
```

On remarque aussi l'existence de la commande `st_deviation` (standard deviation : écart-type en anglais) que nous n'utiliserons pas. On calculera tout simplement la racine carrée de la variance.

• Sur une série statistique pondérée, c'est-à-dire dont on connait la liste `x` des modalités et la liste `f` de leurs fréquences respectives, on peut utiliser le théorème de transfert :

```
-->sum((x.^2).*f) -sum(x.*f)^2
```

b) Etendue et intervalle interquantiles

• L'**étendue** d'une série statistique est la distance entre la plus grande et la plus petite valeur de la série. En Scilab, Elle se calcule donc sur une série statistique X par l'instruction : `e= max(X)-min(X)`.

• L'**intervalle interquartile** est l'intervalle $]Q_1; Q_3[$: c'est l'intervalle des valeurs prises par le caractère si on exclut les 25% de la population les plus éloignés de la moyenne par valeurs inférieures et les 25% les plus éloignés par valeurs supérieures.
Le nombre $Q_3 - Q_1$ est appelé **écart interquartile**.
L'écart interquartile est donc l'étendue de la série statistique à laquelle on a enlevé les 50% de "cas pathologiques".

Remarque
On pourrait, selon les cas, décider d'enlever 20% de "cas extrêmes" : on regarderait alors l'intervalle interdécile $d_9 - d_1$, ou encore uniquement 2% de "cas extrêmes" : on regarderait alors $c_{99} - c_1$.

5. Exercices

Avant de commencer, effacer les anciennes données à l'aide de l'instruction `clear` et fermer toutes le fenêtres graphiques. Faire de même entre chaque exercice.

Exercice 1.
On souhaite étudier la loi d'une variable $Y = |\lfloor X \rfloor|$ où $X \hookrightarrow \mathcal{N}(0, 4)$. Pour cela, on réalise 100 simulation de la loi de Y et on construit le diagramme en bâton de l'échantillon obtenu :

1. La variable aléatoire Y est-elle discrète ou à densité ?
2. Taper dans la console les instructions suivantes, que font-elles ? Les valeurs de Y sont-elles cohérentes ?

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

3. Donner une valeur approchée de l'espérance et de la variance de Y.
4. Taper l'instruction `M=tabul(Y,"i")` et observer le résultat.

5. Mettre les modalités de Y dans une matrice colonne `x`, et les effectifs correspondants dans une matrice colonne `n`. Créer ensuite la matrice colonne `f` des fréquences des x_i.
6. Représenter graphiquement `f` en fonction de `x` à l'aide de la fonction `bar`.
   Donner une approximation de $P(Y = 0)$,$P(Y = 1)$, ..., $P(Y = 8)$.
7. Déterminer le mode de Y et le comparer avec la moyenne empirique et la médiane de Y.
8. Construire le vecteur $F = [f_1, f_1 + f_2, f_1 + f_2 + f_3, ..., 1]$ des fréquences cumulées de Y et tracer le diagramme des fréquences cumulées dans une nouvelle fenêtre.
9. Tracer le diagramme en bâton des fréquences cumulées sur un nouveau graphique.
   Lire sur le diagramme les trois quartiles Q_1, m et Q_3 puis les déciles de la série statistique.

Exercice 2.

1. Compléter et taper dans Scinote la fonction suivante qui effectue une simulation x d'une variable aléatoire X telle que X suit une loi normale $\mathcal{N}(13, 2)$ avec la probabilité 1/3 et une loi normale $\mathcal{N}(7, 2)$ avec la probabilité 2/3 :

   ```
   function x=simulX()
           if ************* then
                x=grand(1,1,'nor',13,2)
           else
               x=grand(1,1,'nor',7,2)
           end
   endfunction
   ```

   Exécuter le programme. Que se passe-t-il dans la console ? Pourquoi ? Que faut-il taper dans la console pour effectuer une simulation de la variable X ?
2. A la suite de ce programme, rajouter les instructions suivantes (après les avoir complété), qui créent et affichent une liste y de 100 simulations de la variable aléatoire $Y = \lfloor X \rfloor$:

   ```
   y=zeros(1,100)
   for k=1:100 do
       y(k)= **************
   end
   disp(y)
   ```

   Exécuter le programme et observer les résultats obtenus.
3. Etude empirique du 100-échantillon de la variable Y :
   — Faites afficher une valeur approchée de l'espérance et de l'écart-type de Y.
   — Tapez `-->gsort(y,'g','i')` et, à l'aide du résultat obtenu, déterminez "à la main" la médiane et les quartiles de l'échantillon y.
   — Tracez l'histogramme des fréquences de l'échantillon y avec les classes $[-0,5; 0,5]$, $]0,5; 1,5]$, ...$]19,5; 20,5]$. Vérifier la cohérence de l'histogramme avec les valeurs de Y.
   On remarque que, contrairement à la commande `bar` *, la commande* `histplot` *tabule toute seule les données.*
4. En observant le graphique, indiquer quels résultats étaient attendus au vu des paramètres et expliquer en quoi les paramètres étudiés ne permettent pas une description totalement satisfaisante de la série statistique y.

Exercice 3.
Simulation d'une loi usuelle :

1. On lance 6 pièces équilibrées et on note X le nombre de piles obtenus. Quelle est la loi de X ?
2. Remplacer les étoiles par ce qu'il faut pour que l'instruction suivante crée une liste x de 40 simulations de la variable aléatoire X.

   ```
   --> x=grand(1,40,***,***,***)
   ```

   Afficher le vecteur x.
3. Soit x un vecteur qui contient la série statistique et c un vecteur qui définit les classes, l'instruction

   ```
   [ind,occ,non]= dsearch(x,c)
   ```

renvoit plusieurs renseignements :
La sortie `non` renvoit le nombre de caractères qui n'appartiennent à aucune classe de c (permet juste de savoir si on a choisi les bonnes classes : s'il y a trop de valeurs qui n'appartiennent à aucune classe, il y a un problème !).
La sortie `occ` (la plus importante) renvoit les effectifs associés à chaque classe de c (elle classe donc les données dans la classe c choisie).
La sortie `ind` (non utilisée en pratique) renvoit la liste des numéros de la classe à laquelle appartiennent les éléments de la série x.

Taper l'instruction :

```
--> [I,O,N]=dsearch(x,[-0.5:1:6.5])
```

et vérifier la cohérence des résultats de la fonction `dsearch`.

4. Tracer le diagramme en bâton des fréquences de la série statistique x.
5. Tracez la représentation graphique en bâton de la loi binomiale de paramètre $(6, \frac{1}{2})$ au moyen de la fonction `binomial` (regarder le détail de la syntaxe de cette fonction dans l'aide).
Vous constaterez qu'en général ce graphique ne se superpose pas avec l'histogramme. Que pourrait-on modifier dans la simulation qui rendrait l'histogramme des fréquences plus proche de la loi théorique ?

Exercice 4. : un peu d'algorithmique
Le but de cet exercice est de compléter l'algorithme de définition de certaines fonctions (boites noires) Scilab ;

1. Compléter la fonction suivante qui prend en entrée une matrice M et renvoit en sortie la somme s des coefficients de la matrice M :

```
function s=SUM(M)
    [n,m]=size(M)
    s=0
    for i=1:n do
        for j=1:m do
            s=***************
        end
    end
endfunction
```

2. Compléter la fonction suivante qui prend en entrée un vecteur ligne x et renvoit en sortie le vecteur c des sommes cumulées de x :

```
function c=CUMSUM(x)
    n=length(x)
    c=zeros(1,n)
    c(1)=x(1)
    for i=2:n do
        c(i)=**********
    end
endfunction
```

3. Compléter la fonction suivante qui range la série statistique x dans l'ordre croissant et renvoit la médiane de x :

```
function m=MEDIANE(x)
    y=gsort(x,'g','i')
    n=length(y)
    if floor(n/2)==n/2 then
        m=*****************
    else
        m=*****************
    end
endfunction
```

Vous pourrez vérifier vos fonctions en comparant sur un exemple le résultat avec celui de la fonction associée prédéfinie en Scilab.

Liste des commandes exigibles : mean, sum, cumsum, tabul, bar, histplot, max, min, rand, grand