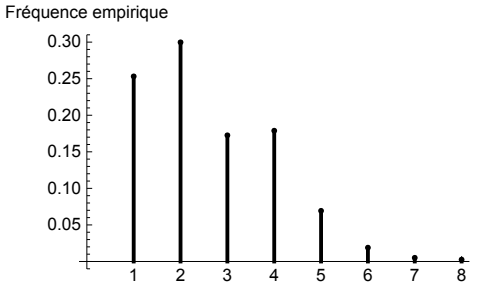
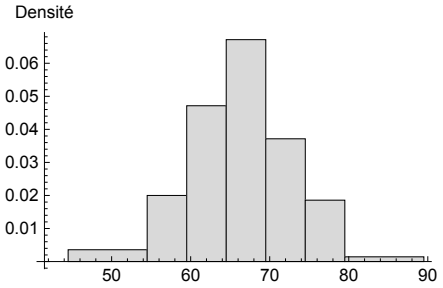
Applications des mathématiques

Statistique I

Statistique descriptive
Modèles statistiques





Version pour *Mathematica*

Edition 2017

Marcel Délèze

https://www.deleze.name/marcel/sec2/applmaths/csud/index.html

§ 0 Introduction

Avant propos

Les buts de ce cours sont

* d'introduire les notions de base de la statistique descriptive:
  fréquence, fréquence cumulée, moyenne, écart-type, ...;
* de traiter des données et de présenter des résultats au moyen de l'ordinateur.

§ 0.1 Variables discrètes, variables continues

Une statistique commence généralement par l'observation d'un ou plusieurs caractères à chacun desquels on associe une valeur numérique. Voici quelques assertions usuelles :

* cette famille a 3 enfants;
* cet homme a 52 ans;
* cet enfant mesure 123 cm;
* ce bébé pèse 5.2 kg;
* cet électeur a voté oui (code 0=non, 1=oui);
* cette cuve est vide (code 0=vide, 1=pleine).

Pour être traduites en langage mathématique, le sens de chacune de ces phrases doit être précisé:

* cette famille a exactement 3 enfants : $x = 3$;
* cet homme a eu 52 ans révolus mais pas encore 53 : $x \in [52;\ 53[$;
* cet enfant mesure 123 cm à un demi-centimètre près : $x \in [122.5;\ 123.5[$;
* ce bébé pèse 5.2 kg à 50 g près : $x \in [5.15;\ 5.25[$;
* cet électeur a voté oui; il n'y a que deux issues; ici $x = 1$ (exactement);
* cette cuve est vide à 5 % près : $x \in [0;\ 0.05[$ (x = taux de remplissage).

Variables discrètes (ou variables discontinues)

Parmi les grandeurs précédentes, certaines ont des valeurs exactes, chacune étant représentable par un point sur une droite:

* cette famille a exactement 3 enfants : **x=3**;
* cet électeur a voté oui; il n'y a que deux issues : code 0=non, 1=oui; ici, **x=1**.

Une variable discrète ne peut pas prendre toutes les valeurs intermédiaires. La valeur de la variable discrète est exacte. Par exemple, la famille observée a exactement 3 enfants; elle ne pourrait pas avoir 2.987 enfants.

Variables continues

Les autres caractères sont des variables continues, affectées d'une incertitude :

* cet homme a eu 52 ans révolus mais pas encore 53 : $x \in [52;\ 53[$;
* cet enfant mesure 123 cm à un demi-centimètre près : $x \in [122.5;\ 123.5[$;
* ce bébé pèse 5.2 kg à 50 g près : $x \in [5.15;\ 5.25[$;
* cette cuve est vide à 5 % près: $x \in [0;\ 0.05[$.

Pour le poids du bébé, l'arrondi à 5.2 kg peut traduire une imprécision de la mesure et désigner le centre d'un intervalle du type 5.20 ± 0.05. L'augmentation de la précision de la mesure d'une variable continue conduit généralement à un résultat un peu différent. Par exemple, un bébé peut

peser (5.18 ± 0.01) kg. Une grandeur continue peut prendre toutes les valeurs réelles situées dans un intervalle.

Une première distinction

Nous verrons que pour les variables discrètes,

on représente les *fréquences* par des *diagrammes à bâtons* et

la fonction de distribution est discontinue, en escalier.

tandis que pour les variables continues,

on représente les *densités* par des *histogrammes* et

la fonction de distribution est une courbe continue.

C'est pourquoi nous présenterons ces deux types de variables dans des paragraphes séparés:

§ 1 et § 3 Distributions à une variable discrète

§ 2 et § 4 Distributions à une variable continue

§ 0.2 Modèle d'urne

Pour introduire l'ensemble de la démarche statistique, partons d'un exemple type appelé "modèle d'urne". Considérons une urne contenant un grand nombre de boules noires et de boules blanches. On cherche la proportion des boules blanches. La probabilité de tirer une boule blanche est

$$p = \frac{\text{nombre de boules blanches dans l' urne}}{\text{nombre total de boules dans l' urne}}$$

Le nombre p étant inconnu, on extrait de l'urne un échantillon de 100 boules et on détermine la fréquence des boules blanches

$$f = \frac{\text{nombre de boules blanches dans l' échantillon}}{\text{nombre total de boules dans l' échantillon}}$$

$$\text{par exemple} \qquad f = \frac{46}{100}$$

La fréquence *f* peut être utilisée comme estimateur de la probabilité p

$$\hat{p} = f \qquad \text{ce qui donne} \qquad \hat{p} = \text{estimation de la probabilité} = \frac{46}{100}$$

Avec l'augmentation de la taille de l'échantillon, il est probable que l'erreur $(f - p)$ diminue. Si pour un échantillon de taille 10000 on a trouvé par exemple $f = \frac{4275}{10\,000}$, il y a de fortes chances que $\frac{4275}{10\,000}$ soit une meilleure estimation de *p* que $\frac{46}{100}$.

Nous devons donc distinguer

* des valeurs empiriques - telles que *f* - qui se rapportent à l'échantillon (valeurs observées);
* des valeurs théoriques - telles que *p* - qui se rapportent à la population et qui sont souvent inconnues;
* des estimations - telles que $\hat{p}$ - qui représentent des approximations de valeurs théoriques à partir de l'échantillon.

§ 0.3 La démarche statistique

Population

Dans une étude statistique, on s'intéresse à un caractère des éléments d'un ensemble bien défini:

les notes de l'ensemble des élèves d'une classe;

les âges de l'ensemble des chamois des Alpes;
les poids de l'ensemble des pains produits par une boulangerie;
les valeurs de l'ensemble des actions cotées à la bourse de New York à une date déterminée;
l'ensemble des mesures d'une grandeur effectuées par un physicien,
....

Un tel ensemble d'objets est appelé une population, quelle que soit la nature des éléments. Nous parlerons donc de "la population des pains produits par une boulangerie" ou de la "population des mesures d'une grandeur effectuées par un physicien". Chaque élément de la population est appelé *individu*.

On doit même souvent distinguer deux populations

* la population *réelle* d'où est tiré l'échantillon et qui est souvent finie;
* la population *théorique*, généralement infinie, qui idéalise la situation et sert d'arrière fond à la distribution théorique.

Variable statistique, variable aléatoire

La variable statistique est le caractère commun que l'on étudie et qui prend une valeur particulière pour chaque individu. Dans ce cours, nous allons ignorer les variables qualitatives et ne considérer que des variables quantitatives telles que

la taille de chaque individu,
le sexe de chaque individu,
le nombre d'enfants de chaque individu,
...

Alors que la variable *statistique* concerne un caractère *empirique*, c'est-à-dire observé, la variable *aléatoire* désigne ce même caractère commun sur la population *théorique*.

Échantillon

Lorsqu'il est difficile ou trop coûteux d'observer tous les individus d'une population, on choisit - généralement au hasard - une partie de la population appelée *échantillon*. Après avoir étudié l'échantillon (valeurs empiriques), la question est de savoir quelles conséquences on peut tirer sur l'ensemble de la population (estimation des valeurs théoriques).

Permanence statistique

L'expérience montre que, lorsqu'on choisit d'autres échantillons, on obtient certes d'autres résultats. Mais les résultats obtenus sont voisins et cohérents. Ils traduisent des caractéristiques stables de la population. Un échantillon peut fournir une *estimation* de certaines caractéristiques de la population dont il est tiré.

Loi de distribution théorique

Il arrive souvent que la population soit très nombreuse et que l'on n'en connaisse pas l'effectif exact. Les modèles probabilistes se réfèrent à la possibilité de répéter des observations une infinité de fois. En statistique, la population est une idéalisation mathématique : il s'agit d'une population, souvent infinie, dont la distribution détermine le modèle théorique. En résumé, le modèle statistique est défini par une *loi de distribution théorique*.

Les étapes de la démarche statistique

La démarche statistique comprend généralement les étapes suivantes:

1. **L'enregistrement et la présentation des données**

   L'acquisition des mesures se fait, suivant les cas, par observations, dénombrements, mesures, ...

   Les données sont mises sous la forme de tableaux, de fichiers informatiques, ...

2. **La réduction des données**

   2. 1 Le groupement des données

   Les données sont groupées en classes et peuvent être présentées graphiquement sous la forme

   - de diagrammes à bâtons dans le cas de variables discrètes ou
   - d'histogrammes dans le cas de variables continues.

   2. 2 Les paramètres empiriques

   Les données sont remplacées par un petit nombre de paramètres empiriques :

   - les paramètres de tendance centrale tels que la moyenne, la médiane, ...
   - les paramètres de dispersion tels que l'écart-type, l'écart interquartile, ...

3. **Le modèle statistique**

   On aimerait, un peu à la manière des physiciens, décrire le phénomène observé par une loi mathématique. On considère que l'échantillon étudié a été tiré d'une population infinie. Les caractéristiques de cette population dont décrites par une distribution de probabilités appelée "modèle théorique". Dans certains cas, la distribution théorique est inconnue. Dans d'autres cas, la distribution théorique a une forme bien déterminée. Par exemple, pour la variable "poids d'un individu", on peut prendre comme densité théorique une courbe en cloche de Gauss.

Dans ce cours introductif, l'accent est mis sur la *statistique descriptive* (§ 1 et § 2) et quelques *modèles théoriques* (§ 3 et § 4).

Exercice 0 - 1

Considérons un échantillon de n oeufs. Pour chacune des situations suivantes, dites s'il s'agit d'une grandeur statistique

qualitative
(quantitative) *discrète* ou
(quantitative) *continue*.

a) On mesure la masse de chaque oeuf, arrondie au gramme.

b) On classe les oeufs en trois catégories: les grands, les moyens et les petits.

c) On classe les oeufs en trois catégories:
ceux dont la masse est inférieure à 53 g;
parmi les restants, ceux dont la masse est inférieure à 63 g;
les autres.

d) On classe les oeufs en trois catégories:
classe C : ceux dont la masse appartient à l'intervalle $[33\,g;\ 53\,g[$;
classe B : ceux dont la masse appartient à l'intervalle $[53\,g;\ 63\,g[$;

classe A : ceux dont la masse appartient à l'intervalle [63 *g*; 83 *g*[.

Exercice 0 - 2

On considère l'ensemble de toutes les personnes qui ont obtenu le baccalauréat au Collège du Sud.
Pour chacune des situations suivantes, dites s'il s'agit d'une variable statistique

qualitative
(quantitative) *discrète* ou
(quantitative) *continue*.

a) On relève l'année où chaque étudiant a obtenu la maturité.

b) On relève l'âge auquel chaque étudiant a reçu la maturité.

§ 1 Statistique descriptive pour une une variable statistique discrète: distribution empirique discrète

Objectifs

Pour chaque notion étudiée (moyenne, médiane, écart-type, ...), le lecteur doit se préoccuper de savoir la calculer

1° à partir de données brutes, sans ordinateur;
2° à partir de données groupées, sans ordinateur;
3° à partir de données brutes, avec *Mathematica;*
4° à partir de données groupées, avec *Mathematica.*

Packages de l'auteur

- On peut consulter le mode d'emploi du package **Statistique**: https://www.deleze.name/marcel/sec2/applmaths/packages/aide/Statistique.pdf
- Avant d'utiliser le package, il faut le charger en donnant son adresse web:

```
Needs["Statistique`",
  nécessite
 "https://www.deleze.name/marcel/sec2/applmaths/packages/Statistique.m"]
```

Voici la liste des instructions disponibles :

```
Names["Statistique`*"]
  noms
```

```
{amplitudes, densiteContinue, densites, diagrammeBatons,
 diagrammeCumulatif, distributionContinue, distributionLissee, fctDensite,
 fctFrequenceCumulee, frequenceCumuleeContinue, frequenceCumuleeLissee,
 histogramme, InterpolatedQuantile, noeudsPolygonaux, polygoneDeDensite,
 quantileC, quantileLisse, sommesCumulees, StandardDeviationMLE, VarianceMLE}
```

- Le package **Tableaux** contient des commandes qui facilitent la présentation des données et résultats sous la forme de tableaux:

```
Needs["Tableaux`",
  nécessite
 "https://www.deleze.name/marcel/sec2/applmaths/packages/Tableaux.m"]
```

```
Names["Tableaux`*"]
  noms
```

```
{afficheTableau, afficheTableauTitre, arrondis, fusionneColonnes,
 fusionneLignes, fusionneTableaux, prodCart, prodCartTrans, tableauGraph}
```

- On peut consulter le mode d'emploi du package **Tableaux**: https://www.deleze.name/marcel/sec2/applmaths/packages/aide/Tableaux.pdf

Pour ne pas oublier d'exécuter ces instructions au début de chaque session de travail, il est conseillé de déclarer les instructions **Needs** comme étant des cellules d'initialisation. Pour ce faire,

sélectionnez les cellules voulues puis passez par le menu

Cell / Cell properties / Initialization cell

§ 1.1 Paramètres empiriques

Exemple 1 : nombre de personnes par ménage

Données brutes

En 1990, dans certains ménages privés du canton de Fribourg, on a compté le nombre de personnes dans le ménage. Dans la liste suivante, chaque nombre correspond à un ménage et indique le nombre de personnes dans le ménage.

```
x = {4, 1, 2, 3, 3, 4, 2, 2, 2, 2, 1, 3, 3, 1, 2, 4, 1, 2, 2, 2, 3, 4, 3, 4, 2, 3, 6, 2, 3, 2, 4, 2,
    2, 3, 2, 4, 3, 4, 1, 4, 2, 2, 4, 1, 1, 2, 2, 2, 1, 5, 1, 1, 3, 4, 2, 5, 2, 1, 2, 1, 2, 5,
    2, 2, 4, 2, 4, 2, 1, 5, 1, 4, 1, 5, 3, 3, 2, 5, 3, 3, 3, 4, 4, 3, 4, 2, 4, 4, 1, 1, 2, 1,
    2, 1, 1, 4, 5, 2, 4, 2, 5, 2, 2, 2, 4, 4, 4, 2, 2, 3, 3, 4, 1, 1, 3, 1, 2, 1, 2, 4, 3, 4,
    4, 3, 2, 2, 1, 2, 1, 1, 2, 3, 1, 4, 3, 1, 5, 1, 3, 4, 2, 1, 2, 1, 2, 3, 1, 2, 2, 3, 5, 3,
    3, 4, 3, 6, 2, 5, 4, 2, 2, 1, 1, 1, 2, 2, 3, 1, 3, 4, 4, 2, 3, 3, 4, 4, 3, 3, 4, 3, 3, 1,
    2, 1, 4, 3, 1, 1, 1, 1, 3, 1, 4, 2, 5, 3, 2, 1, 2, 2, 4, 6, 1, 1, 1, 4, 1, 2, 5, 2, 5, 1,
    1, 2, 3, 3, 4, 3, 4, 4, 2, 1, 1, 2, 3, 3, 2, 3, 4, 3, 4, 2, 1, 2, 1, 2, 2, 5, 2, 4, 1, 2,
    3, 2, 1, 3, 3, 1, 1, 2, 1, 4, 6, 4, 2, 1, 6, 2, 2, 2, 4, 3, 5, 3, 3, 1, 4, 4, 2, 2, 4, 3,
    4, 6, 2, 5, 3, 1, 1, 4, 4, 3, 4, 2, 2, 4, 4, 6, 2, 2, 3, 4, 1, 3, 3, 3, 2, 1, 1, 1, 1,
    4, 3, 2, 1, 4, 2, 2, 2, 1, 1, 1, 4, 1, 1, 1, 2, 2, 4, 1, 4, 3, 3, 5, 1, 1, 3, 1, 2, 1,
    5, 4, 1, 2, 2, 3, 4, 3, 2, 1, 2, 5, 1, 2, 4, 5, 2, 3, 2, 4, 2, 1, 2, 1, 5, 4, 1, 5, 2,
    1, 2, 2, 5, 2, 2, 2, 2, 2, 1, 1, 3, 3, 1, 4, 2, 2, 2, 2, 2, 4, 4, 1, 2, 1, 2, 2, 1, 1,
    1, 2, 1, 3, 4, 2, 2, 1, 5, 2, 4, 2, 3, 2, 4, 1, 5, 2, 1, 5, 1, 3, 1, 2, 3, 1, 4, 3, 4,
    3, 2, 3, 4, 1, 2, 4, 7, 2, 2, 2, 2, 1, 3, 2, 1, 4, 2, 3, 3, 2, 1, 5, 4, 2, 5, 3, 4, 2,
    3, 2, 2, 3, 1, 3, 4, 3, 1, 4, 2, 1, 1, 1, 1, 5, 3, 1, 2, 2, 4, 1, 2, 8, 1, 5, 3, 1, 2,
    1, 2, 2, 1, 2, 4, 5, 3, 1, 1, 4, 3, 4, 1, 2, 3, 4, 1, 1, 2, 1, 3, 1, 3, 7, 5, 2, 2, 1,
    4, 2, 4, 3, 4, 1, 1, 4, 4, 2, 3, 2, 1, 3, 1, 3, 5, 4, 2, 2, 1, 2, 2, 5, 5, 5, 3, 2, 4,
    4, 2, 4, 1, 2, 3, 1, 4, 5, 3, 1, 2, 2, 7, 2, 3, 3, 2, 2, 2, 7, 2, 1, 1, 2, 1, 1, 6, 1,
    2, 3, 2, 1, 1, 2, 3, 1, 4, 3, 4, 4, 5, 3, 3, 1, 2, 3, 2, 2, 3, 5, 3, 3, 5, 4, 5, 1, 6,
    2, 5, 2, 4, 3, 1, 5, 1, 1, 1, 2, 4, 3, 2, 4, 4, 2, 1, 1, 1, 2, 1, 1, 1, 1, 2, 2, 1, 2,
    2, 1, 2, 3, 5, 2, 2, 1, 3, 1, 2, 5, 1, 4, 3, 2, 1, 4, 1, 1, 2, 1, 5, 2, 3, 4, 4, 1, 3,
    5, 4, 1, 2, 1, 2, 4, 1, 6, 2, 3, 2, 3, 2, 2, 4, 2, 2, 2, 1, 2, 1, 2, 3, 2, 2, 3, 3, 2,
    1, 4, 4, 3, 1, 1, 2, 4, 2, 1, 2, 4, 4, 6, 2, 3, 4, 5, 4, 3, 1, 2, 1, 5, 4, 5, 1, 1, 1,
    4, 1, 2, 3, 4, 6, 2, 3, 2, 4, 4, 2, 4, 2, 1, 5, 2, 4, 4, 2, 2, 2, 3, 1, 2, 1, 1, 2, 1,
    1, 5, 2, 3, 1, 1, 2, 4, 2, 4, 2, 4, 3, 4, 4, 4, 6, 1, 2, 2, 1, 5, 3, 1, 1, 4, 4, 4, 2,
    1, 6, 3, 2, 2, 4, 1, 6, 1, 5, 3, 8, 4, 2, 3, 2, 3, 1, 4, 1, 3, 3, 1, 1, 1, 2, 2, 2, 2};
```

Pour obtenir les données numériques précédentes,
https://www.deleze.name/marcel/sec2/applmaths/csud/statistique_1/1-stat_I.nb

Ces données forment un échantillon. Plus généralement, un échantillon de taille n est une liste de *n* éléments notée

$$\vec{x} = \{x_1,\ x_2,\ \ldots,\ x_n\}$$

La taille de l'échantillon est

```
n = Length[x]
```
longueur

794

Données groupées

Groupement des données : modalités, effectifs

Dressons la liste des modalités (ou valeurs) qui apparaissent (ou pourraient apparaître) dans l'échantillon.

Pour notre échantillon, la liste des modalités est $\vec{c} = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

Avec *Mathematica*,

```
c = Union[x]
```
union

```
{1, 2, 3, 4, 5, 6, 7, 8}
```

Les modalités (ou valeurs) sont notées

$$\vec{c} = \{c_1, c_2, \ldots, c_k\}$$

où $k =$ nombre de modalités.

```
k = Length[c]
```
longueur

```
8
```

Pour chaque modalité, on peut calculer son effectif:

Nombre de personnes dans le ménage	Nombre de ménages
1	201
2	238
3	137
4	142
5	55
6	15
7	4
8	2

Pour notre échantillon, les effectifs de chaque modalité sont

$$\vec{n} = \{201, 238, 137, 142, 55, 15, 4, 2\}$$

Plus généralement, les effectifs des modalités sont notés

$$\vec{n} = \{n_1, n_2, \ldots, n_k\}$$

On a la relation

$$n_1 + n_2 + \ldots + n_k = n$$

c'est-à-dire

$$\sum_{j=1}^{k} n_j = n$$

Avec *Mathematica*, on peut calculer automatiquement les effectifs de la manière suivante.

```
? BinCounts
```

BinCounts[{x_1, x_2, …}] counts the number of elements x_i whose values lie in successive integer bins.
BinCounts[{x_1, x_2, …}, dx] counts the number of elements x_i whose values lie in successive bins of width dx.
BinCounts[{x_1, x_2, …}, {x_{min}, x_{max}, dx}] counts the number of x_i in successive bins of width dx from x_{min} to x_{max}.
BinCounts[{x_1, x_2, …}, {{b_1, b_2, …}}] counts the number of x_i in the intervals [b_1, b_2), [b_2, b_3), ….
BinCounts[{{x_1, y_1, …}, {x_2, y_2, …}, …}, $xbins$, $ybins$, …] gives an
array of counts where the first index corresponds to x bins, the second to y, and so on. ≫

```
effectifs = BinCounts[x, {Append[c, ∞]}]
```
compte des huc··· appose

```
{201, 238, 137, 142, 55, 15, 4, 2}
```

Distribution empirique : fréquences, diagramme à bâtons, fréquences cumulées

La fréquence d'une modalité désigne le rapport $\frac{\text{effectif de la modalité}}{\text{effectif total}}$.

Calculons la fréquence de chaque valeur

```
freq = effectifs / n
```

$$\left\{\frac{201}{794}, \frac{119}{397}, \frac{137}{794}, \frac{71}{397}, \frac{55}{794}, \frac{15}{794}, \frac{2}{397}, \frac{1}{397}\right\}$$

```
N[freq]
```
valeur numérique

```
{0.253149, 0.299748, 0.172544, 0.178841, 0.0692695, 0.0188917, 0.00503778, 0.00251889}
```

La fréquence $f_1 = \frac{201}{794} \simeq 0.253$ signifie que, sur 794 ménages, il y en a 201 qui ne comportent qu'une seule personne; en d'autres termes, le 25.3 % des ménages compte une seule personne.
Les fréquences des modalités sont notées

$$\vec{f} = \{f_1, f_2, \ldots, f_k\}$$

Les fréquences ont les propriétés suivantes

$$f_j = \frac{n_j}{n} = \text{fréquence de } c_j$$

$$0 \le f_j \le 1 \qquad \text{et} \qquad \sum_{j=1}^{k} f_j = 1$$

En effet, pour notre échantillon,

$$0 \le 201 \le 794 \quad \Longrightarrow \quad 0 \le \frac{201}{794} \le 1 \qquad \text{et}$$

$$\frac{201}{794} + \frac{238}{794} + \frac{137}{794} + \frac{142}{794} + \frac{55}{794} + \frac{15}{794} + \frac{4}{794} + \frac{2}{794} = \frac{794}{794} = 1$$

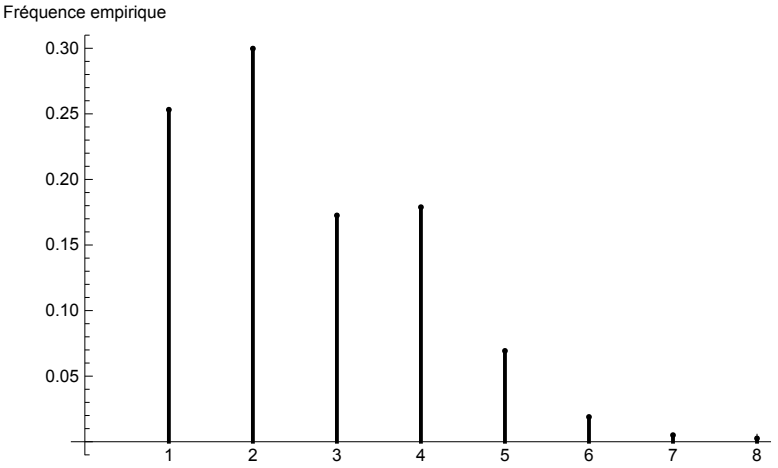
Plus généralement,

$$0 \le n_j \le n \quad \Longrightarrow \quad 0 \le \frac{n_j}{n} \le 1 \quad \Longrightarrow \quad 0 \le f_j \le 1 \qquad \text{et}$$

$$\sum_{j=1}^{k} f_j = \sum_{j=1}^{k} \frac{n_j}{n} = \frac{1}{n} \sum_{j=1}^{k} n_j = \frac{1}{n} n = 1$$

Pour une variable discrète, les fréquences se représentent par un *diagramme à bâtons*

```
diagrammeBatons[c, freq, AxesLabel → {None, "Fréquence empirique"}]
                                titre d'axe      aucun
```



Fréquence d'un intervalle

La fréquence des modalités de l'intervalle] *a*, *b*], notée *f*(] *a*, *b*]), est définie comme suit

$$f(]\,a,\ b]) = \text{somme des fréquences } f_i \text{ des modalités } c_i \text{ telles que } a < c_i \leq b$$

Plus simplement - mais abusivement - cette expression est notée *f*] *a*, *b*] et est appelée *fréquence de l'intervalle*] *a*, *b*].

Dans notre exemple,

$$f]\,2.5;\,4.3] = f_3 + f_4 = \frac{137}{794} + \frac{142}{794} = \frac{279}{794} \simeq 0.351385$$

Fréquences cumulées (ou distribution empirique)

Calculons maintenant les sommes de fréquences

$$F_1 = f_1;\quad F_2 = f_1 + f_2;\quad F_3 = f_1 + f_2 + f_3;\ \ldots$$

```
freqCumulees = Accumulate[freq]
               accumule
```

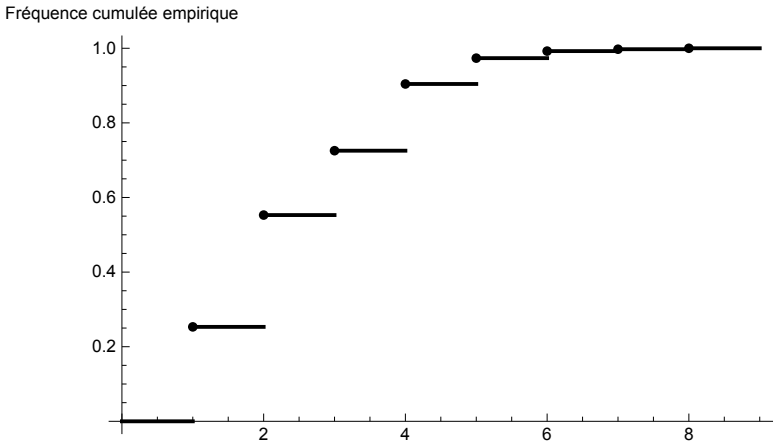
$$\left\{\frac{201}{794}, \frac{439}{794}, \frac{288}{397}, \frac{359}{397}, \frac{773}{794}, \frac{394}{397}, \frac{396}{397}, 1\right\}$$

```
N[freqCumulees]
valeur numérique
```

```
{0.253149, 0.552897, 0.725441, 0.904282, 0.973552, 0.992443, 0.997481, 1.}
```

Lorsque la variable statistique est discrète, les fréquences cumulées se représentent par une fonction discontinue, en escalier.

```
diagrammeCumulatif[c, freq, AxesLabel → {None, "Fréquence cumulée empirique"}]
                                    titre d'axe      aucun
```



Plus généralement, la fréquence cumulée (aussi appelée fonction de distribution empirique) est une fonction qui a les propriétés suivantes

$$\begin{array}{ll} F(t) = 0 & \text{pour } t < c_1 \\ F(t) = f_1 & \text{pour } c_1 \le t < c_2 \\ F(t) = f_1 + f_2 & \text{pour } c_2 \le t < c_3 \\ \quad \ldots & \quad \ldots \\ F(t) = f_1 + f_2 + \ldots + f_{k-1} & \text{pour } c_{k-1} \le t < c_k \\ F(t) = 1 & \text{pour } c_k \le t \end{array}$$

$$\boxed{0 \le F(t) \le 1 \qquad \text{et} \qquad F \text{ croissante}}$$

Interprétation :

$$F(t) = \frac{\text{nombre de ménages dont le nombre de personnes est } \le t}{\text{nombre total de ménages}}$$

$$= \text{fréquence des ménages dont le nombre de personnes est } \le t$$

Sommes de fréquences consécutives

La somme de fréquences consécutives est égale à la différence de deux valeurs de la fonction de distribution. Par exemple,

$$f_3 + f_4 + f_5 = (f_1 + \ldots + f_5) - (f_1 + f_2) = F(c_5) - F(c_2)$$

Plus généralement, pour $i < j$, on a

$$f_i + \ldots + f_j = (f_1 + \ldots + f_j) - (f_1 + \ldots + f_{i-1}) = F(c_j) - F(c_{i-1})$$

où, pour le cas $i = 1$, on pose $f_0 = 0$.

La fréquence l'intervalle $]a,\ b]$ peut s'exprimer au moyen de la fréquence cumulée F:

$$\boxed{f]\,a,\ b] = F(b) - F(a)}$$

Dans le dernier exemple,

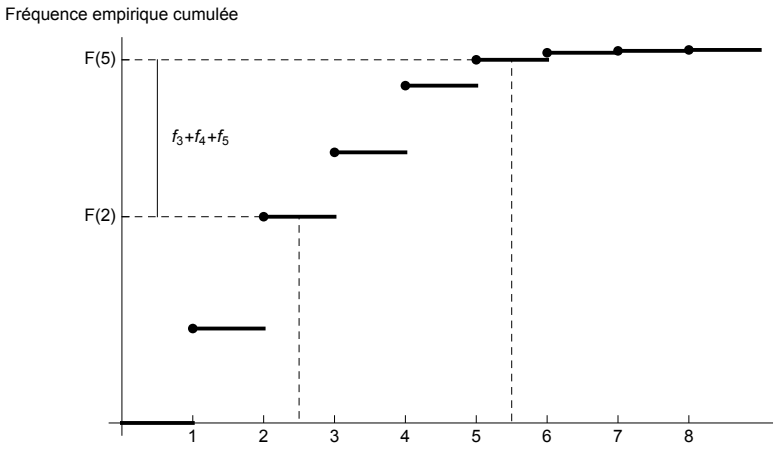
```
freq[[3]] + freq[[4]] + freq[[5]]
```

$$\frac{167}{397}$$

```
freqCumulees[[5]] - freqCumulees[[2]]
```

$$\frac{167}{397}$$

$$f]\,2,\,5] = f_3 + f_4 + f_5 = F\,(5) - F\,(2)$$



On peut aussi voir dans la figure précédente que

$$f]\,2.5,\,5.5] = F\,(5.5) - F\,(2.5) = F\,(5) - F\,(2) = f]\,2,\,5] = f_3 + f_4 + f_5$$

Mesures de tendance centrale : mode, médiane, moyenne

Le mode (plus précisément le mode empirique) est la modalité la plus fréquente. Dans notre échantillon, la modalité la plus fréquente est 2, ce qu'on note $M_o = 2$.
Avec *Mathematica*,

```
mo = Commonest[x]
```

le plus commun

```
{2}
```

Lorsque plusieurs modalités ont une fréquence maximale, on dit que la distribution est *multimodale*. Par exemple,

```
Commonest[{1, 2, 2, 3, 4, 4, 5}]
```

le plus commun

```
{2, 4}
```

Plus généralement,

$$M_o\,(x) = \{c_j \mid n_j = \max\,\{n_1,\, n_2,\, \ldots,\, n_k\}\}$$

Définition de la médiane (au sens de quantile $\frac{1}{2}$) lorsque les valeurs sont discrètes

On appelle ***médiane*** *(au sens de quantile* $\frac{1}{2}$*) la modalité* $Q_{\frac{1}{2}} \in \{c_1,\ c_2,\ \ldots,\ c_k\}$ *pour laquelle la distribution empirique F vaut* $\frac{1}{2}$ *ou pour laquelle la fonction* $t \mapsto \left(F(t) - \frac{1}{2}\right)$ *change de signe.*

Remarque pour initiés

La médiane (dans le sens usuel) est le quantile interpolé $\frac{1}{2}$ c'est-à-dire calculé avec la fréquence cumulée continue par morceaux, comme nous le verrons dans le § 2. Pour une variable statistique continue, la médiane (au sens usuel) et le quantile (interpolé) $\frac{1}{2}$ coïncident. Pour une variable statistique discrète, la médiane (au

sens usuel) et le quantile (non interpolé) $\frac{1}{2}$ sont deux notions un peu différentes. Mais, dans tous les cas, les quantiles $\frac{1}{2}$ (interpolé ou non) sont des médianes (au sens large). Aussi le lecteur débutant est-il dispensé d'entrer dans ces fines distinctions.

Illustrons la définition par deux exemples. Pour l'échantillon {4, 7, 19}, la distribution empirique vérifie

$$F(t) = \frac{1}{3} \quad \text{pour} \quad 4 \le t < 7 \qquad \left(F(t) - \frac{1}{2}\right) < 0$$

$$F(t) = \frac{2}{3} \quad \text{pour} \quad 7 \le t < 19 \qquad \left(F(t) - \frac{1}{2}\right) > 0$$

L'équation F (t) = $\frac{1}{2}$ n'a pas de solution. Par contre, F (t) - $\frac{1}{2}$ change de signe en $t = 7$. La médiane est donc $Q_{\frac{1}{2}} = 7$.

Plus généralement, dans le cas où n est impair, la médiane d'une liste *ordonnée* de n nombres $\{x_1, x_2, ..., x_n\}$ est le terme situé au milieu de la liste

$$x_{\frac{n+1}{2}}$$

Pour l'échantillon {3, 5, 9, 17}, la distribution empirique vérifie

$$F(t) = \frac{1}{4} \quad \text{pour} \quad 3 \le t < 5$$

$$F(t) = \frac{1}{2} \quad \text{pour} \quad 5 \le t < 9$$

$$F(t) = \frac{3}{4} \quad \text{pour} \quad 9 \le t < 17$$

L'équation F (t) = $\frac{1}{2}$ possède une infinité de solutions réelles $t \in [5; 9[$ mais une seule solution est une modalité : $t = 5$. La médiane est donc $Q_{\frac{1}{2}} = 5$.

Plus généralement, dans le cas où n est pair, la médiane d'une liste *ordonnée* de n nombres (éventuellement avec répétitions) $\{x_1, x_2, ..., x_n\}$ est le dernier terme de la première moitié de la liste

$$x_{\frac{n}{2}}$$

Avec *Mathematica*, dans ce cours,

* pour une distribution discrète, nous n'utiliserons pas Median[...] mais **Quantile[..., $\frac{1}{2}$]**;
* pour une distribution continue, nous n'utiliserons pas Quantile[..., $\frac{1}{2}$] mais Median[...] ou InterpolatedQuantile[..., $\frac{1}{2}$].

```
? Quantile
```

Quantile[*list*, *q*] gives the q^{th} quantile of *list*.
Quantile[*list*, {q_1, q_2, …}] gives a list of quantiles q_1, q_2, ….
Quantile[*list*, *q*, {{*a*, *b*}, {*c*, *d*}}] uses the quantile definition specified by parameters *a*, *b*, *c*, *d*.
Quantile[*dist*, *q*] gives a quantile of the symbolic distribution *dist*. ≫

```
Quantile[{4, 7, 19}, 1/2]
```

quantile

7

```
Quantile[{3, 5, 9, 17}, 1/2]
```
quantile

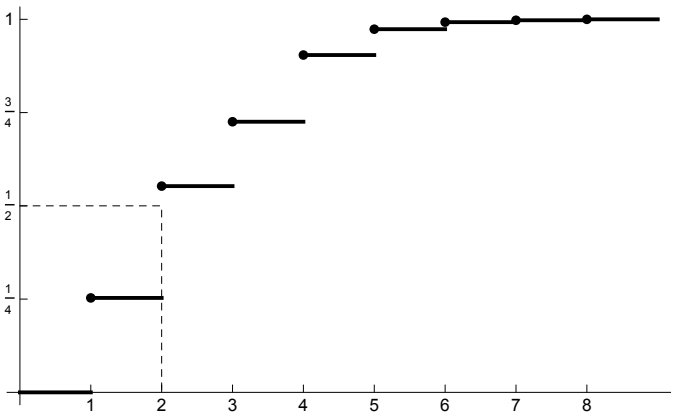
5

Interprétation graphique de la médiane

```
me = Quantile[x, 1/2]
```
quantile

2

Fréquence empirique cumulée



A gauche de la médiane se trouvent environ 50 % des valeurs observées.
A droite de la médiane se trouvent environ 50 % des valeurs observées.

Pratiquement, pour déterminer la médiane, il est plus commode de repérer dans le tableau des fréquences cumulées la modalité pour laquelle le cap de $\frac{1}{2}$ est atteint :

x	$F(x)$
...	...
c_{j-1}	$F_{j-1} < \frac{1}{2}$
c_j	$F_j \geq \frac{1}{2}$
...	...

On en déduit la valeur de la médiane

$$M_e = c_j$$

La moyenne (plus précisément la moyenne arithmétique empirique) peut être définie directement à partir des données brutes :

$$m = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

```
m = Mean[x]; N[m]
```
valeur mo··· valeur

2.60076

La moyenne empirique peut être calculée sur les données groupées.

$$m = \bar{x} = \frac{1}{794}\,(4 + 1 + 2 + 3 + 3 + \ldots + 2) =$$

$$\frac{1}{794}\,(1 * 201 + 2 * 238 + 3 * 137 + 4 * 142 + 5 * 55 + 6 * 15 + 7 * 4 + 8 * 2) =$$

$$1 * \frac{201}{794} + 2 * \frac{238}{794} + 3 * \frac{137}{794} + 4 * \frac{142}{794} + 5 * \frac{55}{794} + 6 * \frac{15}{794} + 7 * \frac{4}{794} + 8 * \frac{2}{794}$$

En général,

$$m = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} c_j\, n_j = \sum_{j=1}^{k} c_j\, \frac{n_j}{n} = \sum_{j=1}^{k} c_j\, f_j$$

$$\boxed{m = \bar{x} = \sum_{j=1}^{k} c_j\, f_j}$$

En *Mathematica*, il est avantageux d'utiliser le produit scalaire, symbolisé par un point, pour réaliser ce type de calcul

```
m = c.freq; N[m]
```
valeur

```
2.60076
```

Le mode, la médiane et la moyenne sont des *mesures de tendance centrale*.

Mesures de dispersion : étendue, écart-type, variance, intervalle interquartile

L'étendue est

```
etendue = Max[c] - Min[c]
```
maximum minimum

```
7
```

L'étendue ne donne généralement pas une bonne idée de la dispersion des données car elle ne se fonde que sur des valeurs extrêmes souvent peu représentatives.

Plus significatifs sont les écarts à la moyenne. Considérons quatre échantillons de nombres entiers

```
{4, 4, 4, 4, 4, 4}
{3, 5, 3, 5, 3, 5}
{2, 6, 2, 6, 2, 6}
{2, 4, 6, 2, 4, 6}
```

Ces quatre échantillons ont la même moyenne arithmétique $m = 4$.
Il diffèrent par leurs écarts à la moyenne $x_i - m$

```
{0, 0, 0, 0, 0, 0}
{-1, 1, -1, 1, -1, 1}
{-2, 2, -2, 2, -2, 2}
{-2, 0, 2, -2, 0, 2}
```

On appelle *mesure de dispersion* toute manière d'évaluer l'importance de ces écarts à la moyenne.

Pour mesurer la dispersion, on utilise parfois l'écart moyen qui est défini comme la *moyenne arithmétique des valeurs absolues des écarts à la moyenne*

$$\frac{1}{n}\sum_{i=1}^{n} | x_i - m | = \sum_{j=1}^{k} | c_j - m | f_j$$

Pour les quatre échantillons ci-dessus, on obtient respectivement

$$0$$
$$1$$
$$2$$
$$\frac{8}{6} = \frac{4}{3} \approx 1.333$$

L'usage de l'écart moyen est problématique et peu répandu. Les mesures de dispersion les plus courantes sont l'écart-type et la variance. Pour calculer l'écart-type, on utilise la moyenne quadratique au lieu de la moyenne arithmétique. La *moyenne quadratique* des nombres $\{e_1, e_2, e_3, ..., e_n\}$ est définie comme suit

$$\sqrt{\frac{e_1^2 + e_2^2 + e_3^2 + \ldots e_n^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (e_i)^2}$$

L'écart-type empirique ou écart standard empirique (non corrigé) est égal à la *moyenne quadratique des écarts à la moyenne* :

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - m)^2}$$

Pour les quatre échantillons précédents, l'écart-type est respectivement

$$0$$
$$\sqrt{\frac{6}{6}} = 1$$
$$\sqrt{\frac{24}{6}} = 2$$
$$\sqrt{\frac{16}{6}} \approx 1.633$$

Quoique la moyenne quadratique puisse être assez différente de la moyenne arithmétique, l'écart-type mesure bien la dispersion des individus autour de la moyenne. Pour notre échantillon "taille des ménages fribourgeois", on obtient

```
s = StandardDeviationMLE[x]; N[s]
                             valeur
```

```
1.38653
```

Le carré de l'écart type est appelé variance. La variance empirique est donc la *moyenne arithmétique des carrés des écarts*

$$v = s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^2$$

La variance et l'écart-type (non corrigés) peuvent être calculés sur les données groupées

$$v = \frac{1}{794}\left((4-m)^2 + (1-m)^2 + (2-m)^2 + (3-m)^2 + (3-m)^2 + \ldots + (2-m)^2\right) =$$
$$\frac{1}{794}\left((1-m)^2 * 201 + (2-m)^2 * 238 + (3-m)^2 * 137 + (4-m)^2 * 142 +\right.$$
$$\left.(5-m)^2 * 55 + (6-m)^2 * 15 + (7-m)^2 * 4 + (8-m)^2 * 2\right) =$$
$$(1-m)^2 * \frac{201}{794} + (2-m)^2 * \frac{238}{794} + (3-m)^2 * \frac{137}{794} + (4-m)^2 * \frac{142}{794} +$$
$$(5-m)^2 * \frac{55}{794} + (6-m)^2 * \frac{15}{794} + (7-m*)^2\ \frac{4}{794} + (8-m)^2 * \frac{2}{794}$$

En général,

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n}\sum_{j=1}^{k}(c_j - m)^2 n_j = \sum_{j=1}^{k}(c_j - m)^2 \frac{n_j}{n} = \sum_{j=1}^{k}(c_j - m)^2 f_j$$

$$\boxed{v = \sum_{j=1}^{k}(c_j - m)^2 f_j}$$

$$\boxed{s = \sqrt{v} = \sqrt{\sum_{j=1}^{k}(c_j - m)^2 f_j}}$$

En *Mathematica*, il est avantageux d'utiliser le produit scalaire, symbolisé par un point, pour réaliser ce type de calcul

```
var = (c - m)^2.freq;
s = √var ; N[s]
```

valeur num

```
1.38653
```

Pour des variables discrètes, la moyenne, l'écart standard et la variance peuvent être calculés indifféremment sur les données brutes ou sur les données groupées. Le groupement des données ne modifie pas les données.

Quantiles et intervalle interquartile

Le nombre q étant donné $(0 < q < 1)$, on appelle *quantile* la valeur $Q_q \in \{v_1, v_2, .., v_k\}$ pour laquelle la distribution empirique F vaut q ou pour laquelle la fonction $t \mapsto (F(t) - q)$ change de signe.

L'intervalle interquartile est égal à la différence entre le quantile $\frac{3}{4}$ et le quantile $\frac{1}{4}$.

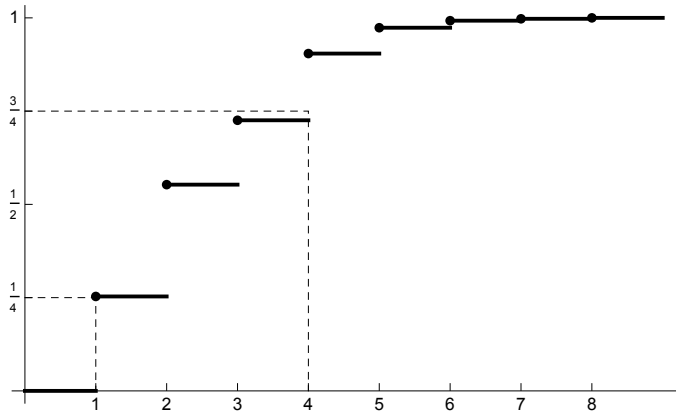
```
interQuartile = Quantile[x, 3/4] - Quantile[x, 1/4]
```

quantile quantile

```
3
```

Dans l'intervalle interquartile est situé approximativement le 50 % des observations.

Fréquence empirique cumulée



Modèle statistique

La variable aléatoire (= caractère observé) est le nombre de personnes dans le ménage, noté *X*.
La distribution théorique est inconnue. Les fréquences empiriques

```
N[freq, 4]
```
valeur numérique

```
{0.2531, 0.2997, 0.1725, 0.1788, 0.06927, 0.01889, 0.005038, 0.002519}
```

n'ont pas toutes la même fiabilité. Parce que les deux dernières sont fondées sur des effectifs inférieurs à 5, elles sont entachées d'une grande incertitude. Les six premières fréquences peuvent être interprétées comme des estimations des probabilités

$\hat{p}_1$ = 0.2531 = estimation de la probabilité qu' un ménage soit constitué d' une seule personne;

...

$\hat{p}_6$ = 0.01889 = estimation de la probabilité qu' un ménage soit constitué de 6 personnes.

L'espérance mathématique $\mu = E(X)$ désigne la "moyenne théorique" de la variable aléatoire *X*. Dans notre exemple, sa valeur exacte nous est inconnue. Si l'échantillon est *représentatif* de la population, la moyenne empirique peut être utilisée comme estimation de l'espérance mathématique

$\hat{\mu}$ = estimation de l' espérance mathématique du nombre de personnes par ménage = 2.6

Autre exemple

La manière de calculer le mode, la médiane, la moyenne, l'écart-type et l'écart interquartile dépend
- de la forme des données: brutes, groupées;
- de la forme du calcul: formules mathématiques, algorithmes, fonctions statistiques de *Mathematica*, fonctions générales (non statistiques) de *Mathematica*, etc.

Ces diverses formes sont présentées au moyen d'un autre exemple; pour les consulter, accédez au site
https://www.deleze.name/marcel/sec2/applmaths/csud/statistique_1/annexes/1-1-complements.zip
puis décompresser pour obtenir les cinq compléments *1-1_compl1_mode.nb, 1-1_compl2_mediane.nb, 1-1_compl3_moyenne.nb, 1-1_compl4_ecart-type.nb, 1-1_compl5_ecart-interquartile.nb*

Exercice 1 - 1 [Sans ordinateur]

Un hameau compte 10 personnes actives dont

9 gagnent 4 000 Fr par mois et

1 gagne 50 000 Fr par mois.

Déterminez

a) la moyenne, la médiane et le mode;

b) l'étendue, l'écart-type et l'écart interquartile.

Exercice 1 - 2 [Sans ordinateur]

Dans une petite localité, on a relevé le nombre de pièces de chaque appartement.

Nombre de pièces	Nombre d'appartements
1	48
2	72
3	96
4	64
5	39
6	25
7	3

a) Représentez graphiquement

le diagramme en bâtons;

la distribution empirique.

b) Calculez ou déterminez

la moyenne arithmétique;

le mode;

la médiane.

c) Calculez ou déterminez

l'étendue;

la variance;

l'écart-type;

l'intervalle interquartile.

Exercice 1 - 3 [Avec Mathematica]

Mêmes questions que dans l'exercice précédent.

Exercice 1 - 4 [Sans ordinateur]

Démontrez que la variance est égale à la différence entre la moyenne des carrés et le carré de la moyenne

$$v = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - m^2$$

Lien vers les corrigés des exercices

https://www.deleze.name/marcel/sec2/applmaths/csud/corriges/statistique_1/1-stat-I-cor.pdf