Statistiques

A. Série statistique à un caractère

On considère un ensemble E (la population) et on s'intéresse à une caractéristique (caractère statistique) que présente chaque élément de E.

Exemples
- on s'intéresse aux notes du devoir de synthèse n°1 en Maths des élèves d'une classe de 3è Année.
La population est l'ensemble des élèves et le caractère statistique est la note.
- on s'intéresse à la durée de vie des ampoules électriques produites par une usine .
La population est l'ensemble des ampoules, le caractère statistique est la durée de vie

On fait ensuite un relevé statistique relativement au caractère étudié, c'est à dire qu'on relève les différentes valeurs prises par le caractère; cela donne une liste de valeurs qui forme une série statistique. Le problème est de la présenter de façon parlante à l'aide de tableaux, de graphiques ou de paramètres statistiques.

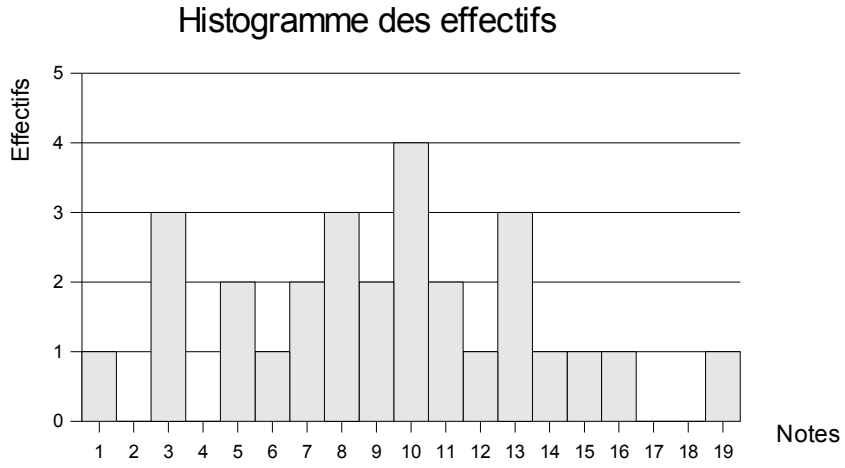
Exemple
Pour le devoir de synhèse n°1 en mathématiques d'une classe de 3è Année, on relève les notes :
7-13-11-5-12-10-8-6-13-19-9-3-5-3-13-10-10-15-16-1-3-9-8-14-7-11-10-8

1- Tableau des effectifs et histogramme

Une première opération consiste à regrouper les notes égales et à indiquer leur nombre, cela conduit à faire un tableau des effectifs.
Pour l'exemple du DS1 :

Notes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Effectifs	1	0	3	0	2	1	2	3	2	4	2	1	3	1	1	1	0	0	1

On peut représenter ce tableau par un histogramme :



On peut indiquer deux paramètres statistiques :
- le mode, c'est la valeur du caractère étudié qui a le plus grand effectif; pour le DS1, le mode

est 10; il s'agit d'un paramètre de position, il est destiné à montrer où se situent les valeurs les plus fréquemment observées.
- l'étendue, c'est la différence entre la plus grande et la plus petite valeurs observées; pour le DS1 l'étendue est 19 – 1 = 18; il s'agit d'un paramètre de dispersion, il est destiné à montrer comment les valeurs se distribuent autour d'une position centrale.
Ces deux paramètres sont très rudimentaires, le but de ce chapitre est d'en étudier d'autres :
- la moyenne et la médiane comme paramètres de position
- l'écart type et l'intervalle interquartile comme paramètres de dispersion.

2- Regroupement en classes

Pour faciliter la lecture du tableau des effectifs et de l'histogramme on peut regrouper les valeurs du caractère étudié en classes.
Pour le DS1, on peut par exemple considérer les classes [0;4[, [4;8[, [8;12[, [12,16[et [16;20[.
Cela donne un nouveau tableau des effectifs :

Classes	[0;4[	[4;8[	[8;12[	[12;16[	[16;20[
Effectifs	4	5	11	6	2

Et l'histogramme :

La classe modale est la classe [8;12[.

3- Fréquences

Lorsqu'on a à comparer des séries statistiques d'effectifs différents, on peut s'intéresser aux fréquences plutôt qu'aux effectifs.

La fréquence associée à une valeur du caractère étudiée est le quotient de son effectif par l'effectif total. Ainsi pour une valeur x_i d'effectif n_i, la fréquence est :

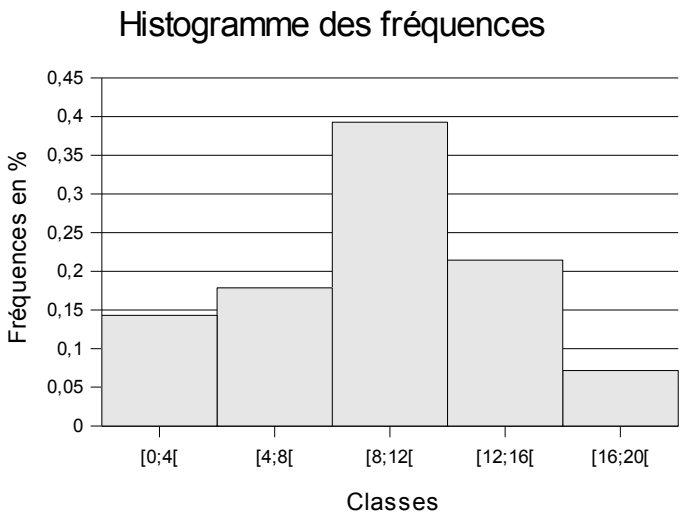
$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum n_i}$$

La fréquence peut être exprimé par un nombre entre 0 et 1 ou par un pourcentage.

Pour le regroupement en classes du DS1, on a le tableau des fréquences :

Classes	[0;4[	[4;8[	[8;12[	[12;16[	[16;20[	Total
Effectifs	4	5	11	6	2	28
Fréquences	0,14	0,18	0,39	0,21	0,07	1
Pourcentages	14%	18%	39%	21%	7%	100%

On obtient alors un histogramme qui a la même forme qu'avec les effectifs, mais avec une graduation standard de 0% à 100% qui permet de réaliser facilement des comparaisons entre séries d'effectifs différents.



B. Médiane et quartiles

Au lieu d'utiliser des classes formées par des intervalles de même longueur comme dans l'exemple précédent, nous pouvons considérer des classes de même effectif. En essayant de créer deux classes de même effectif, on obtient la notion de médiane; en essayant de créer quatre classes de même effectif, on arrive à la notion de quartiles.

1- Médiane

> La médiane sépare une série statistique en deux sous-séries de même effectif, l'une contient les valeurs les plus petites et l'autre les valeurs les plus grandes.

Pour déterminer la médiane d'une série de n valeurs :
- on range les valeurs du caractère étudié par ordre croissant
- si n est impair on prend la valeur située au milieu; si n est pair, on prend la moyenne des deux valeurs situées au milieu.

Exemple
Pour les notes du DS1 :
7-13-11-5-12-10-8-6-13-19-9-3-5-3-13-10-10-15-16-1-3-9-8-14-7-11-10-8
on obtient après rangement dans l'ordre croissant :
1-3-3-3-5-5-6-7-7-8-8-8-9-**9**-**10**-10-10-10-11-11-12-13-13-13-14-15-16-19
Il y a 28 valeurs; les deux valeurs du milieu sont la 14ème et la 15 ème qui sont 9 et 10; la médiane est donc 9,5.
Pour une moitié les notes sont inférieures à 9,5 et pour l'autre elles sont supérieures à 9,5.

2- Quartiles

> Les quartiles permettent de séparer une série statistique en quatre sous-séries de même effectif (à une unité près).
>
> Un quart des valeurs sont inférieures au premier quartile Q1.

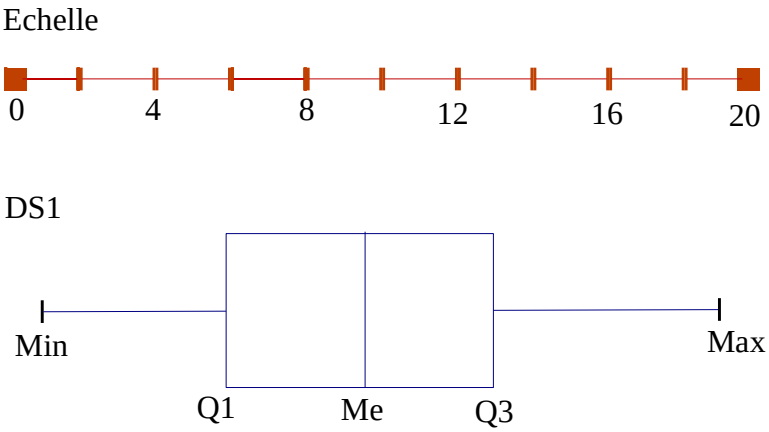
Un quart des valeurs sont supérieures au troisième quartile Q3.
Le deuxième quartile Q2 est aussi la médiane.

Exemple
Pour le DS1, reprenons les valeurs classées dans l'ordre croissant.
1-3-3-3-5-5-**6**-7-7-8-8-8-9-**9**-**10**-10-10-10-11-11-12-**13**-13-13-14-15-16-19
L'effectif total est de 28.
Comme 28/4 = 7, le 1er quartile est la 7ème valeur, donc Q1 = 6.
Comme 3 x 28/4 = 21, le 3ème quartile est la 21ème valeur, donc Q3 = 13.

3- Diagramme en boites

Les notes du DS1 peuvent être résumées par :
- le minimum Min = 1
- le 1er quartile Q1 = 6
- la médiane Me = 9,5
- le 3ème quartile Q3 = 13
- le maximum Max = 19

Ces 5 données permettent de construire un diagramme en boites :



4- Utilisation des effectifs ou des fréquences cumulées

Lorsque l'effectif d'une série statistique est important ou lorsqu'on a effectué des regroupements en classe, les méthodes vues précédemment pour déterminer médiane et quartiles ne sont pas très efficaces ou même pas envisageables.
On utilise alors un tableau des effectifs cumulés ou des fréquences cumulées, puis la représentation graphique en polygone des effectifs ou des fréquences cumulées.

Exemple
Reprenons les données du DS1 avec la répartition en classes :

Classes	[0;4[	[4;8[	[8;12[	[12;16[	[16;20[
Effectifs	4	5	11	6	2

Statistiques

Pour déterminer les effectifs cumulés, nous allons compter le nombre de valeurs inférieures à 0, 4, 8, 12, 16 et 20. Ceci nous donne le tableau :

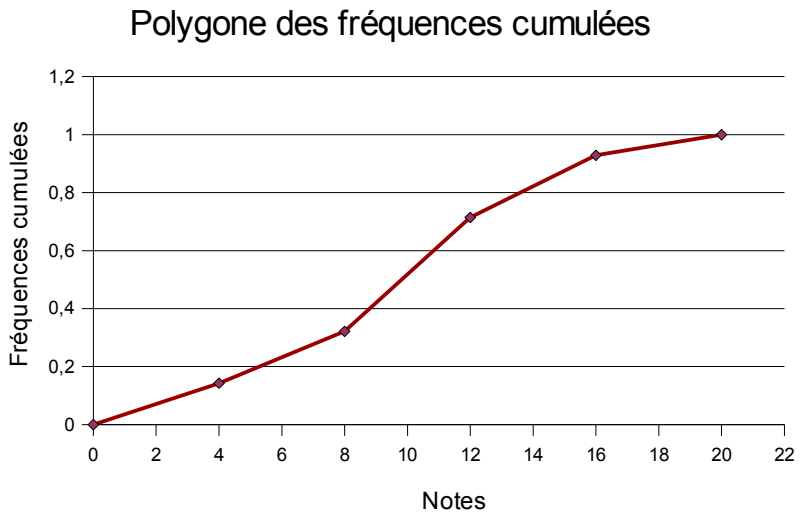
Notes	0	4	8	12	16	20
Effectifs cumulés	0	4	9	20	26	28

Avec les fréquences cumulées (en pourcentages), on obtient :

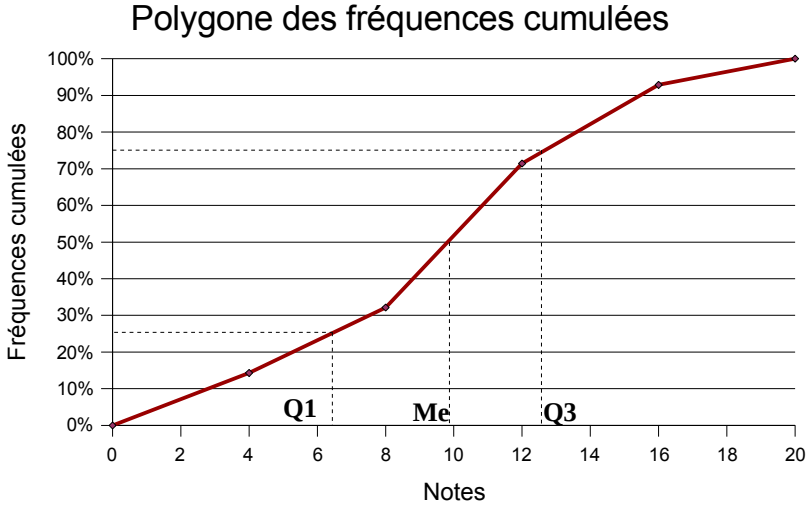
Notes	0	4	8	12	16	20
Effectifs cumulés	0	4	9	20	26	28
Fréquences cumulées	0%	14%	32%	71%	93%	100%

Le premier quartile correspond à une fréquence cumulée de 25% : il est entre 4 et 8.
La médiane correspond à une fréquence cumulée de 50% : elle est entre 8 et 12.
Le troisième quartile correspond à une fréquence cumulée de 75% : il est entre 12 et 16.

Pour déterminer plus précisément ces trois nombres nous pouvons utiliser la représentation graphique en polygone des fréquences cumulées :



Il suffit alors de lire sur le graphique les notes correspondants aux fréquences cumulées 25%, 50% et 75%.



On lit ici : Q1 = 6,5; Me = 9,9 et Q3 = 12,5.
Ces résultats sont voisins de ceux obtenus précédemment, mais moins précis; le regroupement en classes nous a fait perdre des informations.

C. Moyenne et écart type

La moyenne et l'écart type sont un paramètre de position et un paramètre de dispersion

souvent utilisées pour des séries statistiques dites « normales », c'est à dire relativement symétriques avec la plupart des valeurs autour de la moyenne.

1- Moyenne

La moyenne est le quotient de la somme des valeurs par le nombre de valeurs.
Ainsi, pour des valeurs x_i d'effectifs n_i :
- l'effectif total est $n = \sum n_i$
- la moyenne est $m = \dfrac{\sum n_i x_i}{n} = \dfrac{\sum n_i x_i}{\sum n_i}$ (la moyenne est parfois notée $\bar{x}$)

Exemple
Pour le DS1,

Notes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Effectifs	1	0	3	0	2	1	2	3	2	4	2	1	3	1	1	1	0	0	1

la moyenne est :

$$\frac{1\times1+3\times3+2\times5+1\times6+2\times7+3\times8+2\times9+4\times10+2\times11+1\times12+3\times13+1\times14+1\times15+1\times16+1\times19}{1+3+2+1+2+3+2+4+2+1+3+1+1+1+1}$$

$= \dfrac{259}{28} = 9{,}25.$

Si on utilise des regroupements en classe, on choisit comme valeur x_i correspondant à une classe son milieu.
Toujours pour le DS1 :

Classes	[0;4[	[4;8[	[8;12[	[12;16[	[16;20[
Effectifs	4	5	11	6	2

on obtient comme moyenne :

$$\frac{4\times2+5\times6+11\times10+6\times14+2\times18}{4+5+11+6+2} = \frac{268}{28} = 9{,}57.$$

Le regroupement en classes a provoqué une légère modification de la moyenne.

2- Variance et écart type

La variance permet de mesurer la dispersion autour de la moyenne : c'est la moyenne des carrés des écarts à la moyenne.
L'écart type est la racine carrée de la variance.
Ainsi :
- la variance est $V = \dfrac{\sum n_i (x_i - \bar{x})^2}{n}$
- l'écart type est $\sigma = \sqrt{V}$.

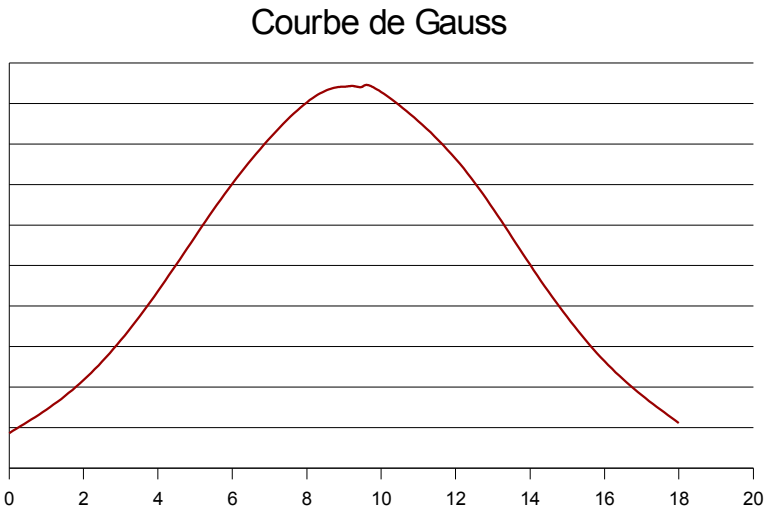
Exemple
Pour le DS1, la variance est 17,9 et l'écart type est 4,23.

Les calculatrices scientifiques et les tableurs nous donnent en général directement ces nombres.

3- Séries statistiques normales (ou gaussiennes)

On dit qu'une série statistique est normale ou gaussienne lorsque l'histogramme des effectifs

s'inscrit approximativement dans une courbe en cloche (courbe de Gauss). Il y a donc à la fois symétrie et regroupement autour de la moyenne qui correspond à l'axe de symétrie.



On dispose des résultats suivants pour les séries statistiques « normale » :

- approximativement 50% des données sont dans l'intervalle $\left[\bar{x}-\frac{2}{3}\sigma,\bar{x}+\frac{2}{3}\sigma\right]$
- approximativement 68% des données sont dans l'intervalle $\left[\bar{x}-\sigma,\bar{x}+\sigma\right]$
- approximativement 95% des données sont dans l'intervalle $\left[\bar{x}-2\sigma,\bar{x}+2\sigma\right]$
- approximativement 99,7% des données sont dans l'intervalle $\left[\bar{x}-3\sigma,\bar{x}+3\sigma\right]$

Exemple

Pour le DS1, nous avons trouvé la moyenne $\bar{x}$ =9,25 et l'écart type σ =4,23.

- l'intervalle $\left[\bar{x}-\frac{2}{3}\sigma,\bar{x}+\frac{2}{3}\sigma\right]$ est ici [6,43; 12,07]; il contient 14 notes sur 28, soit 50% des notes.
- l'intervalle $\left[\bar{x}-\sigma,\bar{x}+\sigma\right]$ est ici [5,02; 13,48]; il contient 18 notes sur 28, soit 64,3% des notes
- l'intervalle $\left[\bar{x}-2\sigma,\bar{x}+2\sigma\right]$ est ici [0,79; 17,71]; il contient 27 notes sur 28, soit 96,4% des notes
- l'intervalle $\left[\bar{x}-3\sigma,\bar{x}+3\sigma\right]$ est ici -3,44; 21,94]; il contient toutes les notes, soit 100% des notes.

La distribution des notes semble donc plutôt « normale ».