STATISTIQUES

L'étude du vocabulaire et des notions de statistique du programme de 1ère L sera illustrée par un exemple: voir la fiche d'accompagnement du travail à réaliser sur tableur et la fiche d'exemples associés à ce travail.
Vous trouverez sur cette fiche de cours uniquement les définitions générales que les différents exemples étudiés permettront d'éclairer.

I) Vocabulaire élémentaire

• **Population**: Ensemble étudié.

• **Individus**: Éléments de la population.

• **Caractère étudié ou variable statistique**: Propriété étudiée dans la population.

Les valeurs que peut prendre un caractère s'appellent les **modalités**.
De façon générale, un caractère peut être :

• **quantitatif** quand les valeurs sont numériques (mesures physiques, physiologiques, sociologique, démographiques, économiques, …)
Le caractère est dit **discret** lorsqu'il ne peut prendre qu'un nombre fini de valeurs numériques: *c'est le cas de l'exemple étudié ici.*
Le caractère est dit **continu**, lorsqu'il peut prendre une infinité de valeurs numériques: *par exemple, la taille d'un élève est un caractère de type quantitatif continu*. Dans cette situation il est commode de regrouper les valeurs du caractère dans des classes: *par exemple, on va regrouper les tailles des individus dans des classes d'amplitude 1 cm.*

• **qualitatif** quand les valeurs ne peuvent être ni ordonnées ni ajoutées (groupe sanguin, couleur des yeux, vote pour un candidat).
Pour des raisons de facilité de traitement informatique ou mathématique, on cherche à se ramener à des caractères quantitatifs par un codage.

• **L'effectif d'une modalité** est le nombre d'individus de la population possédant cette valeur du caractère.

• **L'effectif total** est le nombre d'individus de la population: C'est la somme des effectifs de chaque modalité.

• **La série statistique des effectifs** est la fonction qui, à chaque valeur du caractère (modalité), associe l'effectif de cette modalité.
Elle est le plus souvent définie à l'aide d'un tableau. *Exemple: voir tableau des données statistiques.*

Comme pour les fonctions, les séries statistiques peuvent être représentées graphiquement.

• **La série statistique des effectifs cumulés** est la fonction qui à chaque modalité associe la somme des effectifs des modalités de valeurs inférieures ou égale à cette modalité.

• **La série statistique des fréquences** est la fonction qui, à chaque valeur du caractère, associe la fréquence de la classe de ce caractère.

• **La série statistique des fréquences cumulées** est la fonction qui, à chaque valeur du caractère, associe la fréquence cumulée de la classe de ce caractère. Même méthode que pour les effectifs cumulés.

II) Indicateurs de position

• **Le mode** d'une série statistique est la (ou les) modalités ayant le plus grand effectif.

• **La médiane** d'une série statistique est la valeur centrale de la série statistique: Il y a autant d'effectif avant la médiane qu'après, c'est à dire que les modalités inférieures à la médiane correspondent à 50 % de l'effectif total et les modalités supérieures à la médiane correspondent aux autres 50 % de l'effectif total. C'est pour cela que la médiane est peu sensible aux valeurs extrêmes, ce qui n'est pas le cas de la moyenne !

De façon plus précise:
On ordonne la série des données statistiques par ordre croissant.
Si l'effectif total de la série est impair (de taille: 2n + 1), la médiane est la valeur du terme de rang n + 1 dans cette série ordonnée.
Si l'effectif total de la série est pair (de taille: 2n), la médiane est la moyenne des valeurs des termes de rang n et n + 1 dans cette série ordonnée.

• **La moyenne**

1) Si le caractère étudié est ***discret***, défini par :

Valeur du caractère	x_1	x_2	x_3	.....	x_p
Effectifs	n_1	n_2	n_3	.....	n_p
Fréquences	f_1	f_2	f_3	.....	f_p

La population a pour effectif total: $N=n_1+n_2+n_3+.....+n_p$

• La **moyenne** de cette série statistique est le nombre $\overline{x}$ défini par:

$$\overline{x}=\frac{n_1x_1+n_2x_2+n_3x_3+.....+n_px_p}{N}=f_1x_1+f_2x_2+f_3x_3+.....+f_px_p$$

2) Si le caractère étudié est de type **continu**, il est alors possible de regrouper ses valeurs dans des classes (intervalles les contenant):

Avec p intervalles: $[c_1;c_2[$, $[c_2;c_3[$, $[c_3;c_4[$, $[c_4;c_5[$,....., $[c_p;c_{p+1}]$ dont les centres sont:

$x_1=\frac{c_1+c_2}{2}$, $x_2=\frac{c_2+c_3}{2}$, $x_3=\frac{c_3+c_4}{2}$, $x_4=\frac{c_4+c_5}{2}$, ..., $x_p=\frac{c_p+c_{p+1}}{2}$

Pour calculer sa moyenne, on l'assimile à un caractère discret dont les valeurs sont les centres des intervalles et l'on utilise alors la formule précédente.

Remarque importante*: Moyenne et médiane ne sont pas liée. Voici quelques exemples qui vous montrent que tout est possible…*

Série statistique	**Médiane m et moyenne $\overline{x}$**	**Commentaires**
1 ; 1; 1 ; 2 ; 2	m = 1 , $\overline{x}$ = 1,4	La médiane est la plus petite valeur de la série
1 ; 1; 2 ; 2 ; 2	m = 2 , $\overline{x}$ = 1,6	La médiane est la plus grande valeur de la série
1 ; 1; 2 ; 3 ; 3	m = 2 , $\overline{x}$ = 2	La médiane est égale à la moyenne
1 ; 1; 2 ; 4 ; 4	m = 2 , $\overline{x}$ = 2,4	La médiane est inférieure à la moyenne
1 ; 1; 3 ; 4 ; 4	m = 3 , $\overline{x}$ = 2,6	La médiane est supérieure à la moyenne

III) Indicateurs de dispersion

• **L'étendue** d'une série statistique est la différence entre la valeur minimum et la valeur maximum du caractère étudié.

• **Les quartiles:**

Premier quartile: Valeur de la série statistique (triée dans l'ordre croissant) qui partage l'effectif total en deux parties de la façon suivante:

$\frac{1}{4} = 25\ \%$ de l'effectif est inférieur ou égal à ce premier quartile.

$\frac{3}{4} = 75\ \%$ de l'effectif est supérieur à ce premier quartile.

Lorsqu'il n'est pas possible d'obtenir exactement le partage 25 % - 75 % , on prend pour premier quartile, la plus petite valeur q_1 de la série telle qu'au moins 25 % de l'effectif total soit inférieur ou égal à q_1.

Deuxième quartile: Valeur de la série statistique (triée dans l'ordre croissant) qui partage l'effectif total en deux parties de la façon suivante:

$\frac{1}{2} = 50\ \%$ de l'effectif est inférieur ou égal à ce deuxième quartile.

$\frac{1}{2} = 50\ \%$ de l'effectif est supérieur à ce deuxième quartile.

Le deuxième quartile est donc la médiane de la série statistique. La façon détaillée de la déterminer à déjà été étudiée.

Troisième quartile: Valeur de la série statistique (triée dans l'ordre croissant) qui partage l'effectif total en deux parties de la façon suivante:

$\frac{3}{4} = 75\ \%$ de l'effectif est inférieur ou égal à ce troisième quartile.

$\frac{1}{4} = 25\ \%$ de l'effectif est supérieur à ce troisième quartile.

Lorsqu'il n'est pas possible d'obtenir exactement le partage 75 % - 25 %, on prend pour troisième quartile, la plus petite valeur q_3 de la série telle qu'au moins 75 % de l'effectif total soit inférieur ou égal à q_3.

• **L'intervalle inter quartiles:**

Intervalle qui regroupe la moitié centrale de l'effectif total, c'est à dire situé entre le premier et le troisième quartile.

• **L'écart inter quartiles:**

C'est l'étendue entre le premier et le troisième quartile, c'est à dire l'amplitude de l'intervalle inter quartiles.

• **Les déciles:**

Utilisés surtout pour décrire les séries statistique à fort effectif (les partages se font en dixièmes de l'effectif total). On s'intéresse essentiellement au premier et dernier déciles ($9^{ème}$).

Premier décile: Valeur de la série statistique (triée dans l'ordre croissant) qui partage l'effectif total en deux parties de la façon suivante:

$\frac{1}{10} = 10\ \%$ de l'effectif est inférieur ou égal à ce premier décile.

$\frac{9}{10} = 90\ \%$ de l'effectif est supérieur à ce premier décile.

Lorsqu'il n'est pas possible d'obtenir exactement le partage 10 % - 90 % , on prend pour premier décile, la plus petite valeur d_1 de la série telle qu'au moins 10 % de l'effectif total soit inférieur ou égal à d_1.

Dernier décile: Valeur de la série statistique (triée dans l'ordre croissant) qui partage l'effectif total en deux parties de la façon suivante:

$\frac{9}{10} = 90\ \%$ de l'effectif est inférieur ou égal à ce dernier décile.

$\frac{1}{10} = 10\ \%$ de l'effectif est supérieur à ce dernier décile.

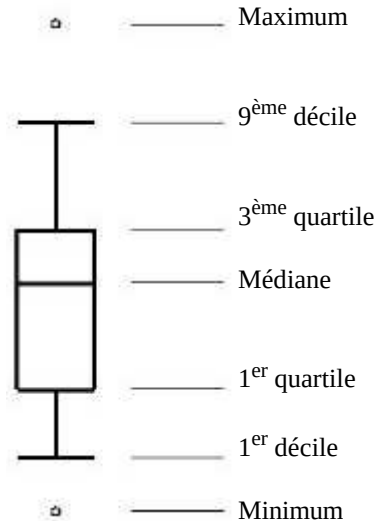
Lorsqu'il n'est pas possible d'obtenir exactement le partage 90 % - 10 % , on prend pour $9^{ème}$ décile, la plus petite valeur d_9 de la série telle qu'au moins 90 % de l'effectif total soit inférieur ou égal à d_9.

Remarque:

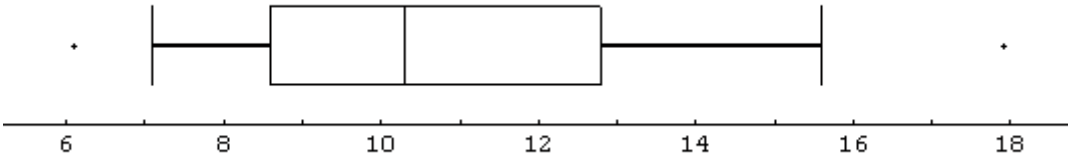
Comme pour l'intervalle inter quartile, on évalue l'**intervalle inter décile** qui contient les 80 % centraux de la population totale. Il est situé entre le premier et le dernier décile.

- **Le diagramme en boîte ou « Boîte à moustache » ou « Boîte à pattes » :**

Il s'agit, d'une façon imagée, de décrire la répartition de l'effectif total à l'aide de sa médiane, son 1^{er} et son $3^{ème}$ quartile, son 1^{er} et son $9^{ème}$ décile, son maximum et son minimum, de la façon représentée ci-dessous:



Afin que les informations soient lisibles, il est nécessaire de donner une graduation associée à la boîte, comme le montre l'exemple ci-dessous:



Remarques: Parfois, le minimum et le maximum ne sont pas indiqués. Parfois aussi, pour des séries à petits effectifs, pour les bouts des pattes, on remplace le 1^{er} décile par le minimum et le $9^{ème}$ décile par le maximum. Il arrive aussi parfois que la moyenne soit signalée par une croix afin de la situer par rapport à la médiane....

- **L'écart-type:**

C'est le réel positif σ défini par :

$$\sigma^2 = \frac{n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+n_3(x_3-\bar{x})^2+.....+n_p(x_p-\bar{x})^2}{N}$$

Le nombre σ^2 est appelé la ***variance*** de la série statistique.

• **Distribution normale - Données gaussiennes - Plages de normalité:**

Dans de nombreux domaines, les séries statistiques portant sur un très grand nombre de données conduisent à des graphiques (polygone des effectifs ou histogramme des effectifs) de même forme régulière et symétrique qui sont très proches d'une "courbe en cloche" appelée "courbe de Gauss", en hommage au grand mathématicien allemand Karl-Friedrich Gauss (1777-1855) qui a tant apporté aux mathématiques.

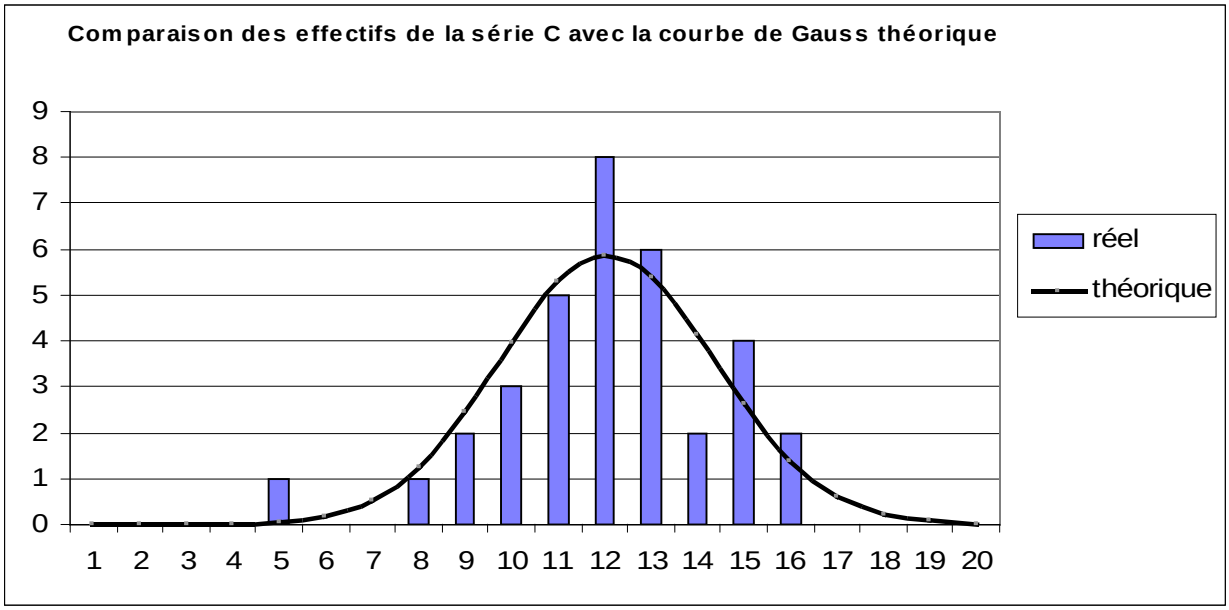
De telles données sont dites "gaussiennes" et la loi mathématique associée s'appelle la "loi normale". Les courbes obtenues sont à peu près symétriques autour de la moyenne $\bar{x}$ de la série statistique.

Pour des séries de ce type, on observe que:

- **Environ 95 % des valeurs du caractère étudié sont situées dans l'intervalle $[\bar{x}-2\sigma\,;\bar{x}+2\sigma]$**
  Cet intervalle est appelé plage de normalité à 95 %.
- **Environ 99 % des valeurs du caractère étudié sont situées dans l'intervalle $[\bar{x}-3\sigma\,;\bar{x}+3\sigma]$**
  Cet intervalle est appelé plage de normalité à 99 %.

Exemple à traiter en utilisant les données de la série C

Les données de notre exemple sont issues d'une population dont l'effectif total est très petit (du point de vue statistique) et n'ont pas tout à fait les caractéristiques d'une distribution de type "gaussien" (distribution "normale"). Le graphique ci-dessous, comparant les distributions pratiques et théoriques des notes montre quelques fluctuations:



En résumé:

Avec des séries statistiques de type gaussien, on peut s'attendre à trouver environ 95 % des observations qui fluctuent à moins de 2 écarts-type de la moyenne $\bar{x}$ de la série statistique. Cela signifie aussi que, toute valeur de cette série a 95 % de chance de se trouver dans l'intervalle $[\bar{x}-2\sigma\,;\bar{x}+2\sigma]$ et donc 5 % de chance de se trouver à l'extérieur de cette plage de normalité à 95 %.

Avec des séries statistiques de type gaussien, on peut s'attendre à trouver environ 99 % des observations qui fluctuent à moins de 3 écarts-type de la moyenne $\bar{x}$ de la série statistique. Cela signifie aussi que, toute valeur de cette série a 99 % de chance de se trouver dans l'intervalle $[\bar{x}-3\sigma\,;\bar{x}+3\sigma]$ et donc 1 % de chance de se trouver à l'extérieur de cette plage de normalité à 99 %. On peut donc être pratiquement certain que la quasi totalité des valeurs de cette série statistique est situé à moins de trois écart-types de la moyenne de la série.