Statistiques

L'étude et la comparaison de séries statistiques menées en classe de seconde se poursuivent avec la mise en place de nouveaux outils dans l'analyse de données.
Ce chapitre de statistiques va tout d'abord vous rappeler les notions vues l'année dernière puis approfondir un petit peu...
Oui, juste un petit peu.

www.mathsbook.fr

I - Rappels de statistiques

Dans cette première partie, nous allons revoir les notions de base de statistiques, les langages et les premiers calculs. Rien de bien méchant.

1 - Rappels de base et langage statistique

On commence par un tableau qui rappelle de cours de statistiques de seconde. Oui oui, se sont des rappels.

En langage statistique	En langage mathématiques	En langage courant
Population	Ensemble P	Ensemble des objets, des individus, mesures étudiées
Individu	Elément de P	Objet, individu, mesure
Caractère Caractère quantitatif Caractère qualitatif	Application f $f : P \to \mathbb{R}$ $f : P \to A (A non inclus dans \mathbb{R})$	Aspect des objets étudiés Les données sont numériques Les données ne sont pas numériques
Modalité Caractère quantitatif discret Caractère qualitatif continu	Image par f des éléments de P $f(P)$ sont des valeurs isolées $f(P)$ est un intervalle	Valeur que peut prendre le caractère Valeurs isolées que l'on peut dénombrer Toute valeur d'un intervalle
Classe	Partie de l'ensemble des modalités	Regrouper des modalités
Effectif Effectif total Effectif cumulé croissant	n_i $N = n_1 + n_2 + ... + n_k$ $n_1 + n_2 + ... + n_p$	Nombre d'individu ayant la même modalité Nombre total d'objets Somme des effectifs
Mode Caractère discret Mode Caractère continu Classe modale Mode		 Valeur du plus grand effectif Valeur du plus grand effectif Centre de la classe modale
Fréquence	$F = \frac{n_i}{N}$	Rapport : effectif d'une valeur par effectif total
Etendue		Différence entre les valeurs extrêmes

2 - Médiane

Pas mal de rappels déjà. On continue avec la définition de la médiane.

Définition de la médiane : C'est la valeur du caractère qui permet de partager la population N en deux groupes de même effectifs. On distingue deux cas : celui d'un caractère quantitatif discret et celui d'un caractère quantitatif continu.

Cas d'un caractère quantitatif discret :
- Si N est impair : la médiane est la valeur du caractère observé au rang $\frac{N+1}{2}$.
- Si N est pair : la médiane n'est pas définie, mais on convient de prendre pour médiane la moyenne des caractères observés au rang $\frac{N}{2}$ et $\frac{N}{2}+1$.

Cas d'un caractère quantitatif continu : on construit la courbe des fréquences cumulées et la médiane est l'antécédent de $0,5$.

Je vais vous donner un exemple simple du cas d'un caractère quantitatif discret.

Exemple : Les notes d'un élève de première sont les suivantes : 3, 5, 12, 14 et 18.
On dénombre cinq notes distinctes, donc un nombre impair de notes.
La médiane est donc la valeur du rang 3. En effet, on applique bêtement la formule précédente :

$$\frac{5+1}{2} = 3$$

D'où : la médiane est 12.

Maintenant, si l'on rajoute la note de 15 à l'élève. On aurait donc les notes suivantes : 3, 5, 12, 14, 15 et 18.
La on est dans le cas d'un nombre de notes pair. On va prendre la moyenne des rang $\frac{N}{2}$, soit 12, et $\frac{N}{2}+1$, soit 14. Ce qui nous donne :

$$\frac{12+14}{2} = \frac{26}{2} = 13$$

La médiane est donc 13.

3 - Moyenne arithmétique pondérée

Une petite définition pour commencer.

Définition de la moyenne arithmétique pondérée : La moyenne arithmétique pondérée, que l'on note $\bar{x}$, est donnée par la formule suivante :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + ... + n_kx_k}{n_1 + n_2 + ... + n_k} = \frac{1}{N} \times \sum_{i=1}^{i=k} n_ix_i$$

Avec $N = n_1 + n_2 + ... + n_k$ et n_i l'effectif de la valeur x_i.

4 - Exemples

Bon, maintenant on va s'exercer un peu sur des exemples pour bien clarifier toutes les notions que l'on vient d'aborder.

Voici donc deux exemples complets à savoir faire et refaire.

Exemple : Etude d'une série statistique à caractère discret :

Dans une classe de 25 élèves de première, les résultats à un contrôle de mathématiques sont les suivants :

7 ; 9 ; 15 ; 11 ; 10 ; 10 ; 16 ; 7 ; 8 ; 14 ; 15 ; 9 ; 10 ; 10 ; 14 ; 15 ; 18 ; 12 ; 8 ; 14 ; 8 ; 8 ; 10 ; 11 ; 15.

Alors, déjà, quelle est la population, le caractère et les valeurs prises par ce dernier ?

...

Eh bien, allez-y ? Vous connaissez la réponse, j'en suis sûr !

Bon, je vous aide.
La population est l'ensemble des contrôles de mathématiques.
Le caractère étudié est la note obtenue par chaque élève de première de cette classe.
Les valeurs prises par le caractères sont les entiers compris entre 7 et 18 (les valeurs des notes quoi).

On va résumer les notes dans l'ordre croissante, l'effectif, l'effectif cumulé et la fréquence dans un tableau :

Note	7	8	9	10	11	12	14	15	16	18
Effectif	2	4	2	5	2	1	3	4	1	1
Effectif cumulé	2	6	8	13	15	16	19	23	24	25
Fréquence	0,08	0,16	0,08	0,2	0,08	0,04	0,12	0,16	0,04	0,04

Normalement, si vous avez bien compris et bien appris toutes les formules précédentes, vous saurez sans aucun problème retrouver toutes les valeurs de ce tableau.

Je l'explique un peu quand même.
La première ligne correspond aux notes des élèves au contrôle de maths. Ca, pas de problème je pense.
La deuxième ligne correspond au nombre de chacune des notes. Par exemple, 2 personnes ont obtenu 7 au contrôle, 4 ont eut 8, etc.
La troisième ligne, c'est la même chose, sauf qu'on compte cette fois-ci combien de personne au eut la note ou moins, soit : 8 personnes ont eut 9 ou moins, etc. On retombe bien sur le nombre total d'élèves, à savoir 25, à la fin.
La dernière ligne, c'est la fréquence. Vous avez la formule un peu plus haut. Pas besoin de réexpliquer.

Calculons maintenant l'étendue, le mode et la médiane.

Calcul de l'étendue : Je vous rappelle que l'étendue est la différence entre la valeur maximale et la valeur minimale, soit ici 11 :

$$18 - 7 = 11$$

Calcul du mode : C'est la valeur qui correspond au plus grand effectif, c'est-à-dire ici la note qui a été obtenue par le plus d'élève. Il s'agit de... 10 ! Oui, 10, obtenue par cinq élèves.
Calcul de la médiane : On a un nombre impair de notes, donc on applique la formule suivante :

$$\frac{25+1}{2} = \frac{26}{2} = 13$$

La médiane est donc la note obtenue par le 13^{me} élève. C'est là que va nous service la ligne des effectifs cumulés.
On lit aisément que le 13^{me} élève a eut 10 à son contrôle de maths, la médiane est donc ici de 10.

Exemple : Etude d'une série statistique à caractère continu :

Dans un lycée, nous avons relevé la taille des élèves et les avons regroupées dans le tableau suivant :

Taille en cm	$[150;160[$	$[160;165[$	$[165;170[$	$[170;175[$	$[175;180[$	$[180;200[$
Effectif	34	42	80	59	50	35

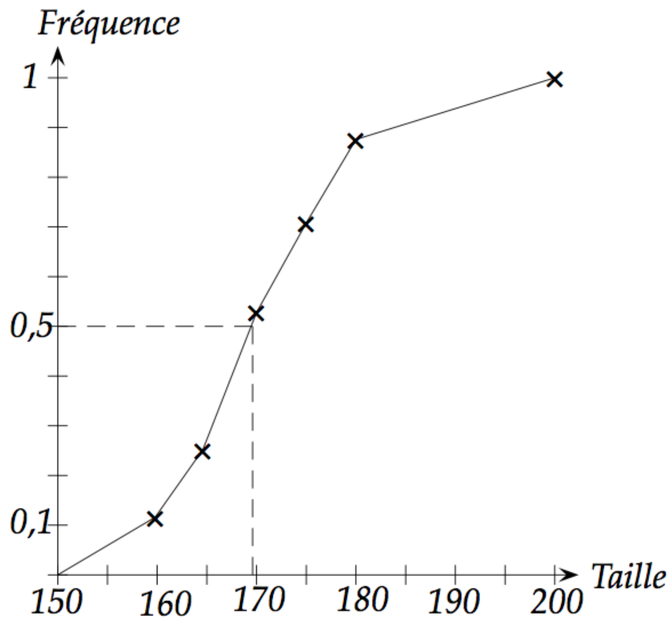
On va calculer, ensemble (oui, je ne vous lâche pas, ne vous inquietez pas) :
- L'étendue,
- La classe modale,
- Le mode,
- La médiane,
- La moyenne.

Alors, pas de temps à perdre, on y va de suite. Je ne rappelle pas à chaque fois les formules pour gagner du temps.

Calcul de l'étendue : $200 - 150 = 50$.
Calcul de la classe modale : $[165;170[$.
Calcul du mode : C'est le centre de la classe modale, soit : $167,5$.
Calcul de la médiane : Rappelons simplement que dans une série statistique à caractère continu, la médiane est la valeur qui correspond à une fréquence de 0,5. Vous avez compris ce que cela veut dire ? On est obligé de calculer les fréquences oui. Allons-y. Je les ai regroupé dans le tableau suivant :

Taille en cm	$[150;160[$	$[160;165[$	$[165;170[$	$[170;175[$	$[175;180[$	$[180;200[$
Effectif	34	42	80	59	50	35
Fréquence	0,11	0,14	0,27	0,20	0,16	0,12
Fréquence cumulée	0,11	0,25	0,52	0,72	0,88	1

Puis on construit la courbe des fréquences cumulées.



Après lecture graphique, on détermine facilement la médiane qui vaut 169cm.
Calcul de la moyenne : on termine par le plus simple :

$$\frac{34 \times 155 + 42 \times 162,5 + 80 \times 167,5 + 59 \times 172,5 + 50 \times 177,5 + 35 \times 190}{34 + 42 + 80 + 59 + 50 + 35} = \frac{51197,5}{300} = 170,66$$

La moyenne est donc de 170,66cm.

II - Propriétés de la moyenne

Deux propriétés à connaître sur la moyenne d'une série statistique.

Propriétés de la moyenne :

- Soit une série statistique $(x_1, x_2, ..., x_n)$ de N éléments, réparties en p sous groupes ayant respectivement $n_1, n_2, ..., n_p$ éléments (tels que $n_1 + n_2 + ... + n_p = N$) et de moyennes respectives $\bar{y_1}$, $\bar{y_2}$, ..., $\bar{y_p}$.
  Alors la série statistique a pour **moyenne** :

$$m = \frac{n_1\bar{y_1} + n_2\bar{y_2} + ... + n_p\bar{y_p}}{N}$$

- Soient a et b deux réels fixés et une série statistique $(x_1, x_2, ..., x_n)$ de N éléments et de moyenne $\bar{x}$.
  Alors la série statistique $(y_1, y_2, ..., y_n)$, avec $y_i = ax_i + b$ pour i variant de 1 à n, a pour **moyenne** :

$$\bar{y} = a\bar{x} + b$$

Oulala, je n'ai rien compris ! Des explications s'il-vous-plaît ?

Je vous explique tout sur un exemple.

Exemple : Dans un lycée, il y a quatre classes de première contenant respectivement 25, 26, 30 et 29 élèves.
A un contrôle commun de mathématiques, les moyennes de ces classes sont respectivement 10,5 ; 11,2 ; 9,8 et 10,3.
On va calculer la moyenne au contrôle pour l'ensemble des classes de première. Vous allez voir, c'est plus simple que cela en a l'air.
On utilise la formule précédent et c'est tout :

$$m = \frac{25 \times 10,5 + 26 \times 11,2 + 30 \times 9,8 + 29 \times 10,3}{25 + 26 + 30 = 29} = \frac{1146,4}{110} = 10,42$$

Donc, la moyenne au contrôle commun de maths est de 10,42. Simple, non ?

III - Variance et écart type

Je vais vous introduire les notions de **variance** et d'**écart type** avec un exemple.

Au sept contrôles de maths du trimestre, un premier élève, nommé Damien (oui, pourquoi pas), a obtenu les notes suivantes : 10 ; 15 ; 12 ; 12 ; 15 ; 14 et 15. Un second élève, Alexandre, a obtenu les notes suivantes : 5 ; 2 ; 8 ; 16 ; 15 ; 10 et 19.

Je vous laisse calculer la moyenne de ces deux élèves... Oui, ils ont la même moyenne de 13,29.

On observe cependant que les notes de Damien sont plus groupées autour de la moyenne que celle d'Alexandre.
Ces deux élèves n'ont pas le même profil. Il est donc nécessaire de pouvoir évaluer la dispersion autour de la moyenne.
Et devinez quoi ? C'est le rôle de la **variance** et de l'**écart type** !

Je vais vous définir ces deux (nouvelles) notions.

Pour les deux définitions qui suivants, on considérera une série statistique $(x_1, x_2, ..., x_n)$ avec n_i les effectifs correspondants à chaque x_i et m la moyenne de cette série.

Définition de la variance : La **variance** de la série statistique est :

$$V = \frac{\sum_{i=1}^{i=n} n_i(x_i - m)^2}{N}$$

Avec :

$$N = \sum_{i=1}^{i=n} n_i$$

Et l'écart type ? C'est encore plus simple quand on connait la variance.

Définition de l'écart type : L'**écart type** de la série statistique est la racine carrée de la variance :

$$S = \sqrt{V}$$

Remarque : L'écart type mesure la dispersion des valeurs d'une série statistique autour de la moyenne. Donc, plus l'écart type est élevé, plus la dispersion est grande.

Je vous laisse donc calculer la variance et l'écart type de l'exemple d'introduction...

IV - Quartile et décile

Deux nouvelles notions que je vous définis tout de suite avant de les expliquer dans un exemple.

Définitions : Soit une série statistique $(x_1, x_2, ..., x_n)$ de taille n.

- Les **quartiles** partagent la série statistique en quatre parties de même effectif.
  Le premier quartile Q_1 est la plus petite valeur de la série telle qu'au moins 25% des données soient inférieures ou égales à Q_1.
  Le troisième quartile Q_3 est la plus petite valeur de la série telle qu'au moins 75% des données soient inférieures ou égales à Q_3.
- L'**intervalle interquartile** est l'intervalle $[Q_1; Q_3]$.
- L'**écart interquartile** est $Q_3 - Q_1$.
- Les **déciles** partagent la série en 10 parties de même effectif.

Remarques : Deux remarques intéressantes.
- La médiane est le deuxième quartile.
- L'écart interquartile mesure la dispersion des valeurs autour de la médiane, c'est-à-dire la dispersion des 50% des valeurs autour de la médiane.

Et on explique tout ça dans un exemple.

Exemple : Dans une classe de 21 élèves, les notes rangées dans l'ordre croissant sont les suivantes :

Rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Note	1	4	4	5	7	7	8	9	10	11	11	11	11	11	13	14	14	15	16	16	19

Quel est le premier et le troisième quartile ?
Pour le premier quartile, il suffit de faire :

$$21 \times 0,25 = 5,25$$

Donc, le premier quartile est la 6^{me} valeur, soit :

$$Q_1 = 7$$

Et pour le troisième quartile :

$$21 \times 0,75 = 15,75$$

Donc, le premier quartile est la 16^{me} valeur, soit :

$$Q_3 = 14$$

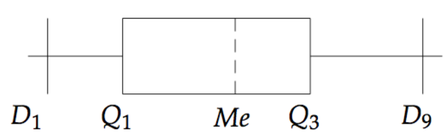
D'autres exemples sont données dans la section suivantes.

V - Diagramme en boîte

La dernière partie de ce chapitre concerne les diagrammes en boîte. C'est tout simplement ce qui permet de représenter une série statistique.

Définition d'un diagramme en boîte : Un **diagramme en boîte** permet de représenter une série statistique au moyenne d'une boite rectangulaire sur laquelle sont indiqués les informations suivantes :
- Le premier quartile Q_1 et le troisième quartile Q_3,
- Le premier décile D_1 et le troisième décile D_3,
- La médiane Me.

Voici sa représentation :



En fait, un diagramme en boîte, c'est une un espèce de dessin qui représente à l'échelle chacun des éléments statistiques de la série statistique.

Prenons un exemple pour illustrer tout cela.

Exemple : Dans deux classes de 30 élèves, les notes obtenues à un devoir commun se répartissent de la comme suit :

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

On calcule rapidement la médiane de chacune de ces deux classes : $M_1 = 11$ et $M_2 = 11$.

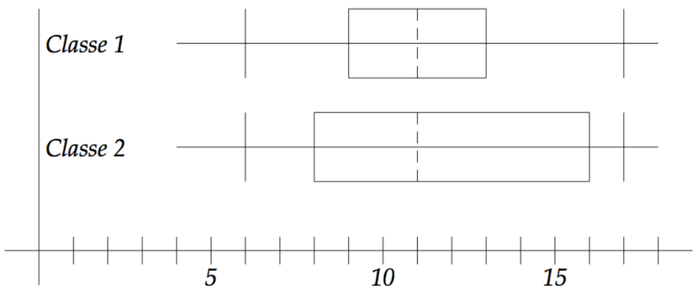
Calculons maintenant, pour les deux classes, D_1, D_9, Q_1 et Q_3.

Calcul de $D1$: $30 \times \frac{10}{100} = 3$, ainsi, D_1 est la note obtenue par le 3^{me} élève.
Calcul de $D9$: $30 \times \frac{90}{100} = 27$, ainsi, D_1 est la note obtenue par le 27^{me} élève.
Calcul de $Q1$: $30 \times \frac{25}{100} = 7,5$, ainsi, D_1 est la note obtenue par le 8^{me} élève.
Calcul de $Q3$: $30 \times \frac{75}{100} = 22,5$, ainsi, D_1 est la note obtenue par le 23^{me} élève.

On rassemble tous ces résultats dans un tableau :

Classe	D_1	D_9	Q_1	Q_3
Classe 1	6	17	9	13
Classe 2	6	17	8	16

On peut à présent construire de le diagramme en boîte pour chacune des classes :



Mais comment lire ce diagramme ? Que représente-t-il ?

Ce diagramme montre que la classe 2 possède des notes beaucoup plus dispersées que la classe 1.
En effet, 25% des notes de la classe 2 sont comprises entre 11 et 16 alors que 25% de celles de la classe 1 sont comprises entre 11 et 13.

Et voilà, nous avons terminé ce chapitre sur les statistiques.
Je vous conseille de vous entraîner sur tout ce qu'on vient d'apprendre sur les exercices.