Statistiques.

Historiquement, l'étude des statistiques doit son développement au besoin de recenser une population. On trouve des données datant d'il y a 4500 ans en Mésopotamie et en Egypte.

Le mot vient du latin *status* (état). Dans la vie courante, le statistiques prennent de plus en plus d'essor et il convient de bien les comprendre afin de s'en faire sa propre interprétation et non celle bien souvent partiale des commanditaires.

"Selon les statistiques, il y a une personne sur cinq qui est déséquilibrée.
S'il y a 4 personnes autour de toi et qu'elles te semblent normales, c'est pas bon."JCVD

I Rappels.

Les statistiques servent à étudier une **population** peuplée d'**individus**. Pour cela, on interroge un **échantillon** sur un ou plusieurs thèmes nommés **caractères** (qualitatif ou quantitatif, continu ou discret). Ces données sont recueillies et regroupées pour former une **série statistique**.

- Les caractères étudiés sont décomptés individuellement ou regroupés en **classe**.
- Le nombre d'individus est exprimé en **effectif** ou en **fréquence** (la nuance majeure venant du fait que, dans le second cas, on ne connaît pas l'effectif de l'échantillon).
- On utilise plusieurs types de représentations : les diagrammes en bâtons, les histogrammes (attention au piège de l'aire) ou les diagrammes circulaires.

Voici la même série statistique exprimée de deux façons différentes. Il s'agit de la longueur des oreilles d'une famille de Takin du Sichuan. La première avec un décompte individuel i variant de 1 à N(=30 ici) et la seconde regroupée en paquets, l'entier j variant de 1 à P(=9 ici)

n° (i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Oreilles (x_i)	0,5	1	4,5	3,5	2,5	2	1	1,5	2	2	4	0,5	1	2,5	2
n°	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Oreilles	0,5	2	3	2,5	3	4	1	2	2	4	1	2	2	3	2,5

Taille (y_j)	0,5	1	1,5	2	2,5	3	3,5	4	4,5
Effectif (n_j)	3	5	1	9	4	3	1	3	1
Fréquence (f_j)	0,1	0,167	0,033	0,3	0,133	0,1	0,033	0,1	0,033

Dans la pratique, il est hors de question de donner toutes les mesures, on attend une interprétation des données, c'est pourquoi on utilise des indicateurs, que l'on choisit en fonction de l'étude demandée.

Moyenne : (Indicateur de **position)** $\overline{x}=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_i=\frac{1}{\mathrm{N}}\sum_{j=1}^{\mathrm{P}} n_j y_j=\sum_{j=1}^{\mathrm{P}} f_j x_j=\overline{y}$ [ici, $\overline{x}$=2,167]

Mode : (Indicateur de position) noté m, c'est la valeur du caractère correspondant au plus grand effectif (ou à la plus grande fréquence). [ici, m=2]

Médiane : (indicateur de **répartition**) Avec une série **ordonnée**, la médiane notée M (=2 ici) partage la population en deux parties de même effectif (i.e. il y a autant de x_i plus petit que M que de x_i plus grand). Cet indicateur est très peu sensible aux variations des valeurs extrêmes.

Si n est impair, alors M=x_k avec $k=\frac{n+1}{2}$. Si n est pair, alors M=$\frac{x_k+x_{k+1}}{2}$ avec $k=\frac{n}{2}$.

Etendue : (indicateur de **dispersion**) On la note e=x_{max}-x_{min} (=4 ici) Cet indicateur est peu utile car très sensible aux variations des valeurs extrêmes.

Exemple : Etudions un nouveau problème : Les accidents de la route : Voici le tableau de la répartition des accidents corporels de la route selon les heures de la journée pour l'année 1992:

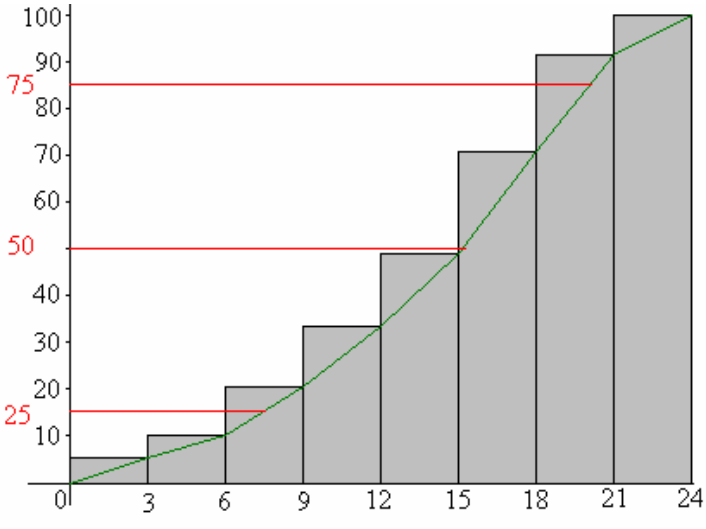
Tranche horaire (en heure)	[0 ;3[	[3 ;6[	[6 ;9[	[9 ;12[	[12 ;15[	[15 ;18[	[18 ;21[	[21 ;24[
Nombre d'accidents	8155	6258	15284	18006	23703	29759	29172	13022

Calculer quand a lieu en moyenne un accident de la route (14h04). Commentez ce résultat (sans intérêt). (N=143359)

On va plutôt s'intéresser à la **répartition** de la série. Pour cela nous avons besoin de la série des fréquences cumulées :

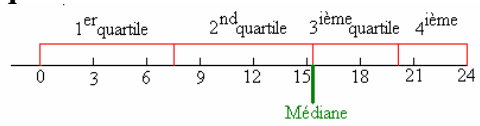
tranche horaire (en heure)	[0 ;3[	[3 ;6[	[6 ;9[	[9 ;12[	[12 ;15[	[15 ;18[	[18 ;21[	[21 ;24[
Fréquences (en %)	5,7	4,4	10,7	12,6	16,5	20,8	20,3	9,1
Fréquences cumulées croissantes (en %)	5,7	10,1	20,8	33,4	49,9	70,7	91	100,1

Tracer l'histogramme des fréquences cumulées croissantes :



En vert, le polynôme statistique de fréquences cumulées croissantes, autrement appelé **courbe de répartition**.

La classe modale [15 ;18[montre que cette période est la plus dangereuse, mais qu'en est-il des autres ? Pour étudier la répartition, on préfère une décomposition en **quartiles** :



Introduction du diagramme de Tukey

Pour les **quartiles** : Q_1 est la plus petite valeur du caractère telle qu'au moins 25% des termes de la série soient inférieurs ou égaux à Q_1. Quant à Q_3, c'est la plus petite valeur du caractère telle qu'au moins 75% des termes de la série soient inférieurs ou égaux à Q_3.

Soulever le problème du caractère continu ou discret.

On définit l'**intervalle interquartile** I=[Q_3,Q_1] c'est un indicateur de dispersion plus précis que l'étendue. On utilise souvent le couple (M,I) ,mais voici mieux encore :

Étudions un cas concret : cinq sportifs ont couru un 1500m et un 5000m. Leurs temps sont donnés dans le tableau suivant :

	Coureur 1	Coureur 2	Coureur 3	Coureur 4	Coureur 5
1500 m	3'58"17	4'05"48	4'12"97	4'08"29	4'00"12
5000 m	14'58"12	14'47"08	15'37"85	13'57"70	14'48"34

Laquelle des deux courses a les temps les plus homogènes ?

Pour le 1500 m : (on convertit tous les temps en secondes pour un calcul plus aisé)

- moyenne : $m = \frac{1}{5}(238{,}17 + 245{,}48 + 252{,}97 + 248{,}29 + 240{,}12) = 245{,}006$ secondes (soit environ 4'05"01)
- variance : $V = \frac{1}{5}(238{,}17^2 + 245{,}48^2 + 252{,}97^2 + 248{,}29^2 + 240{,}12^2) - 245{,}006^2 \simeq 29{,}0$ d'où un écart-type $s \simeq 5{,}39$ secondes
- coefficient de variation : $C_v = \frac{s}{m} \simeq 0{,}022$.

Pour le 5000 m :

- moyenne : $m' = \frac{1}{5}(898{,}12 + 887{,}08 + 937{,}85 + 837{,}70 + 888{,}34) = 889{,}818$ secondes (soit environ 14'49"82)
- variance : $V' = \frac{1}{5}(898{,}12^2 + 887{,}08^2 + 937{,}85^2 + 837{,}70^2 + 888{,}34^2) - 889{,}818^2 \simeq 1020{,}4$ d'où un écart-type $s' \simeq 31{,}94$ secondes
- coefficient de variation : $C_v' = \frac{s'}{m'} \simeq 0{,}036$.

Conclusion : le 1500 m a été plus homogène car $C_v < C_v'$.

On peut également, dans ce type de situation, utiliser l'interquartile relatif.

Pour le 1500 m, on a $\frac{Q_3 - Q_1}{m_e} = \frac{248{,}29 - 240{,}12}{245{,}48} \simeq 0{,}033$.

Pour le 5000 m, on a : $\frac{Q_3' - Q_1'}{m_e'} = \frac{898{,}12 - 887{,}08}{888{,}34} \simeq 0{,}012...$

Conclusion : le 5000 m a été plus homogène que le 1500 m.

Moralité : surtout lorsque les effectifs sont petits, le coefficient de variation et l'interquartile relatif n'aboutissent pas toujours aux mêmes conclusions. (Rappel : l'interquartile ne tient compte que de 50% de la population)

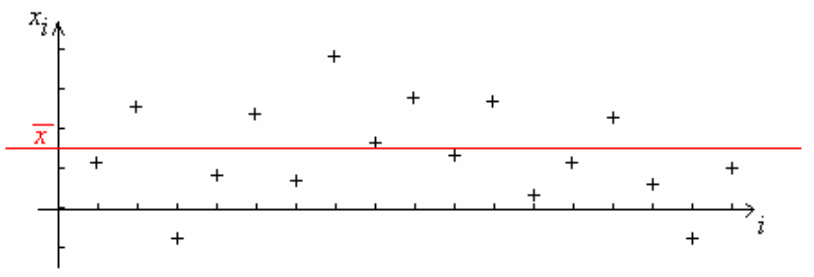
II Variance et écart type.

Définition :.

On recherche un bon indicateur de dispersion. La première idée est de regarder l'**écart à la moyenne**.

$$e_m = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|$$

qui est complètement inadapté à l'analyse du fait de la présence des valeurs absolues. Aussi, on lui préfère la **variance**



$$V=\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2=\frac{1}{N}\sum_{j=1}^{P}n_j(y_j-\overline{y})^2=\sum_{j=1}^{P}f_j(y_j-\overline{y})^2.$$

Or cet invariant a un vrai défaut, son unité. En effet, il ne s'agit pas de la même unité que le caractère (si le caractère est une taille en mètre, alors la variance est en m^2). Aussi, on lui préfère l'**écart type**

$$\boxed{s=\sqrt{V}}.$$

Propriété (8.A)(Formule dite de Koenig-Huygens) : $V=\frac{1}{N}\left(\sum_{j=1}^{P}n_jy_j\right)^2-\overline{y}^2$.

(La moyenne des carrés moins le carré de la moyenne)

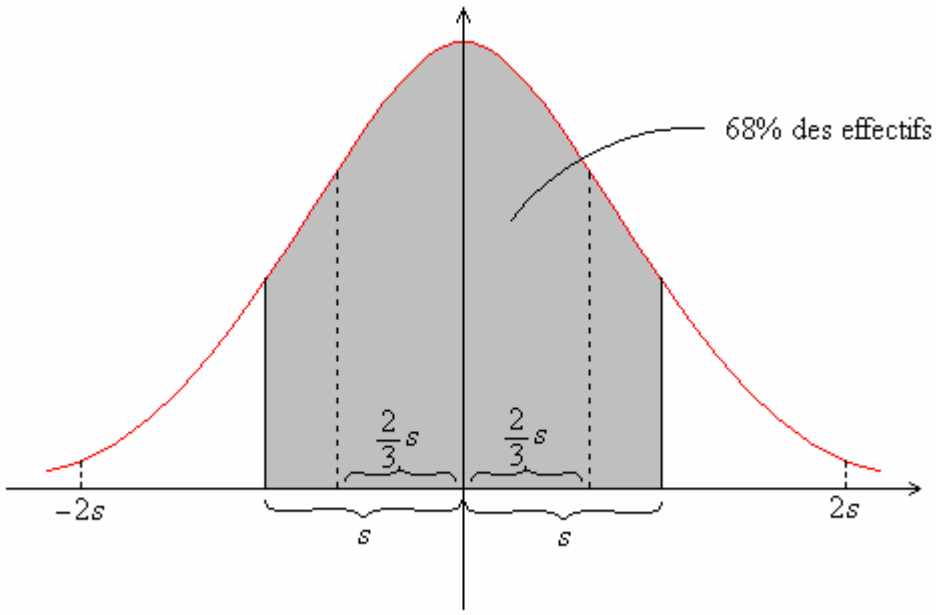
Démonstration : On développe $(y_j-\overline{y})^2$ dans la définition.

Il reste à légitimer ce choix de V :

Proposition (8.B) : La fonction $g : a\rightarrow\frac{1}{N}\sum_{j=1}^{P}n_j(y_j-a)^2$ possède un minimum V atteinte en $a=\overline{y}$.

Démonstration : il suffit de dériver et de déterminer le sommet de la parabole.

Dans une série répartie de façon normale (loi de Gauss, anecdote de l'usure de la marche en pierre), on a les chiffres suivants : 50% des effectifs sont dans l'intervalle $[\overline{y}-2/3s\ ;\ \overline{y}+2/3s]$,
68% des effectifs sont dans l'intervalle $[\overline{y}-s\ ;\ \overline{y}+s]$,
95% des effectifs sont dans l'intervalle $[\overline{y}-2s\ ;\ \overline{y}+2s]$



III Centrer, réduire.

Dans le but de comparer plusieurs séries entre elles, on fait subir une tranformation affine à la série, ce qui est le moins traumatisant pour les données.

Proposition (8.C) :Soit $(y_j\ ;n_j)$ une série statistique, M sa médiane, Q_1, Q_3, ses quartiles, V sa variance et s son écart type. On pose $(z_j\ ;n_j)$ la série statistique telle que $z_j=ay_j+b$ avec a et b deux réels (a non nul). On note M' sa médiane, Q'_1, Q'_3, ses quartiles,

V' sa variance et *s'* son écart type.
On a alors M'=*a*M+*b*, V'=a^2V, *s'*=|*a*|*s*, et si *a*>0, $Q'_1=aQ_1+b$ et $Q'_3=aQ_3+b$.

Qu'en est-il de la moyenne ? du mode ?

Centrer, c'est se ramener à $\overline{z}=0$.
Réduire, c'est se ramener à *s*=1.
Demander aux élèves ce qu'il faut prendre pour *a* et *b*. [*a*=1/*s* et *b*= $-\overline{y}/s$]

Ainsi, avec y_j une série statistique quelconque, la série $\boxed{z=\frac{y-\overline{y}}{s}}$ est toujours centrée réduite.

Intérêt sur un exemple:
Avez-vous déjà réfléchi à la manière dont on peut comparer les décathloniens entre eux. Ces champions doivent courir (100m, 400m, 1500m), sauter (longueur, hauteur, perche), combiner les deux (110m haies) et lancer (disque, javelot, poids). Et tout cela en deux jours!
Certains sont plus doués pour certaines épreuves. Comment comparer un athlète meilleur qu'un autre au lancer du poids, mais moins bon au saut en longueur?. Voilà le problème auquel la fédération internationale d'athlétisme est confrontée depuis les débuts de l'existence de cette épreuve.

En fait à chaque performance est attribué un certain nombre de points que l'on trouve dans une table (actuellement la table "hongroise", anciennement la table "finlandaise").
La question est de savoir comment établir ces tables, comment comparer un saut de 7m60 en longueur avec un jet de 18m50 au poids. C'est ici que les statistiques vont intervenir.

On sait que, sous des conditions fort générales, un ensemble de données obéit à la loi normale. On peut supposer qu'il en va ainsi pour les résultats des athlètes pour un sport donné. A ce moment il est possible de calculer la moyenne et l'écart-type des scores obtenus. Si *m* désigne cette moyenne, et σ l'écart-type, la distribution *F*(*x*) de *x* obéit à la loi:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

On peut ***normaliser*** en effectuant un changement d'origine de telle sorte que la moyenne vaille 0 et un changement d'unité pour que l'écart vaille 1.

$$F(x) = \lambda\, e^{-\frac{x^2}{2}}$$

Chacune des épreuves est alors représentée par la même loi *F*(*x*) et on peut comparer celles-ci; il ne reste plus qu'à faire choix d'une échelle.