RAPPELS DE STATISTIQUES

VALEURS CENTRALES ET DISPERSION

4A1CD | MODULE 3

Objets d'étude

- **Valeurs centrales :** mesures qui permettent de dégager un comportement moyen, ou une tendance générale observée dans la population

  Exemples : moyenne - médiane - mode

- **Dispersion :** mesures permettant de mesurer l'éloignement ou la dispersion des valeurs autour de la valeur centrale.

  Exemples : variance - étendue

- Ces critères servent à étudier le comportement d'une population à l'égard d'une caractéristique donnée

Objets d’étude

- Nous verrons que certaines mesures présentées ne sont pas adaptées à tous les types de variables
- Il est important de comprendre quelle mesure a un sens selon la nature des variables.

Exemple : il serait risible de vouloir calculer une valeur moyenne pour une variable nominale. Que serait une valeur moyenne entre « Lyon » et « Paris » ?

VALEURS CENTRALES

MODE - MOYENNE - MÉDIANE - QUANTILES

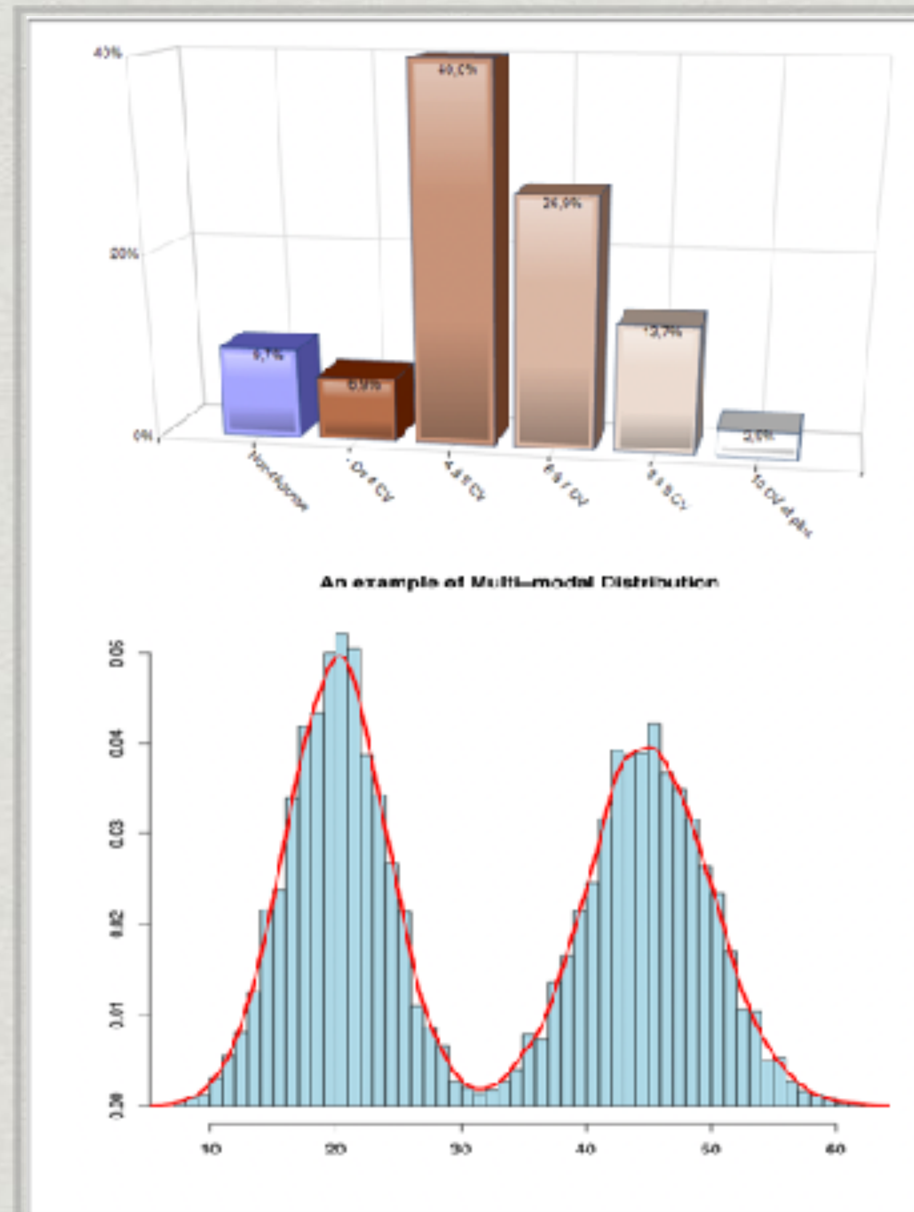
Le Mode

- Grandeur statistique qui caractérise la valeur ou la caractéristique la plus représentée au sein d'une population/échantillon.
  C'est donc **la valeur la plus fréquente** dans un échantillon.
- Elle peut s'utiliser pour n'importe quel **type de variable : quantitative et qualitative.**

Exemples : La destination préférée des français ou encore la note la plus fréquemment obtenue par des étudiants lors d'un examen.

Le Mode

- Une distribution qui ne comporte qu'un seul mode est dite **uni-modale.**
- Si la distribution comporte plusieurs modes, elle est dite **multi-modale.**
- Ces caractéristiques sont valables quelque soit la nature de la variable



La Moyenne

- Grandeur statistique qui permet de caractériser le comportement moyen d'une population concernant **un attribut (ou une variable) quantitative.** La moyenne est souvent utilisée pour synthétiser un ensemble d'informations.
- Cette grandeur permet également de donner **une tendance** concernant notre échantillon.

Exemples :

- la taille moyenne d'une femme de 20 ans en Europe est de 165cm
- Le nombre moyen d'enfants par foyer est de 1.87

La moyenne n'a pas toujours un sens physique réel, comme le montre le dernier exemple. Il s'agit d'une grandeur abstraite dans la plupart des cas.

La Moyenne

- Il s'agit donc d'un **nombre réel** $\bar{x}$ que l'on peut calculer aisément lorsque l'on dispose d'**une variable quantitative (discrète ou continue).**

- Si l'on dispose d'un échantillon retraçant les valeurs prises par des individus dans une population $(x_i)_{i=1}^n$.
La valeur moyenne est définie par :

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

C'est donc la somme de la caractéristique sur chaque individu, divisé par le nombre total d'individus

La Moyenne

- Grandeur très souvent utilisée par les organismes statistiques (INSEE) ou encore par les instituts de sondage (IFOP)
- Elle présente cependant un inconvénient majeur … elle est sensible aux valeurs extrêmes présentent dans l’échantillon

  Exemple : se renseigner sur le salaire moyen des français en interrogeant 100 personnes parmi lesquelles se trouvent un millionnaire. Notre moyenne sera alors **biaisée et anormalement élevée.**

Il existe cependant une autre grandeur statistique permettant d’étudier le comportement moyen d’une population et qui n’est pas sensible à ses valeurs extrêmes.

La Médiane

- Elle permet également de traduire un comportement moyen d'une population vis-à-vis d'une **caractéristique quantitative** donnée.
- La médiane a la particularité d'être une valeur séparant notre échantillon **en deux ensembles de même taille.** Ainsi 50% des individus de l'échantillon ont une valeur inférieure à la valeur médiane et 50% ont une valeur supérieure à la valeur médiane.

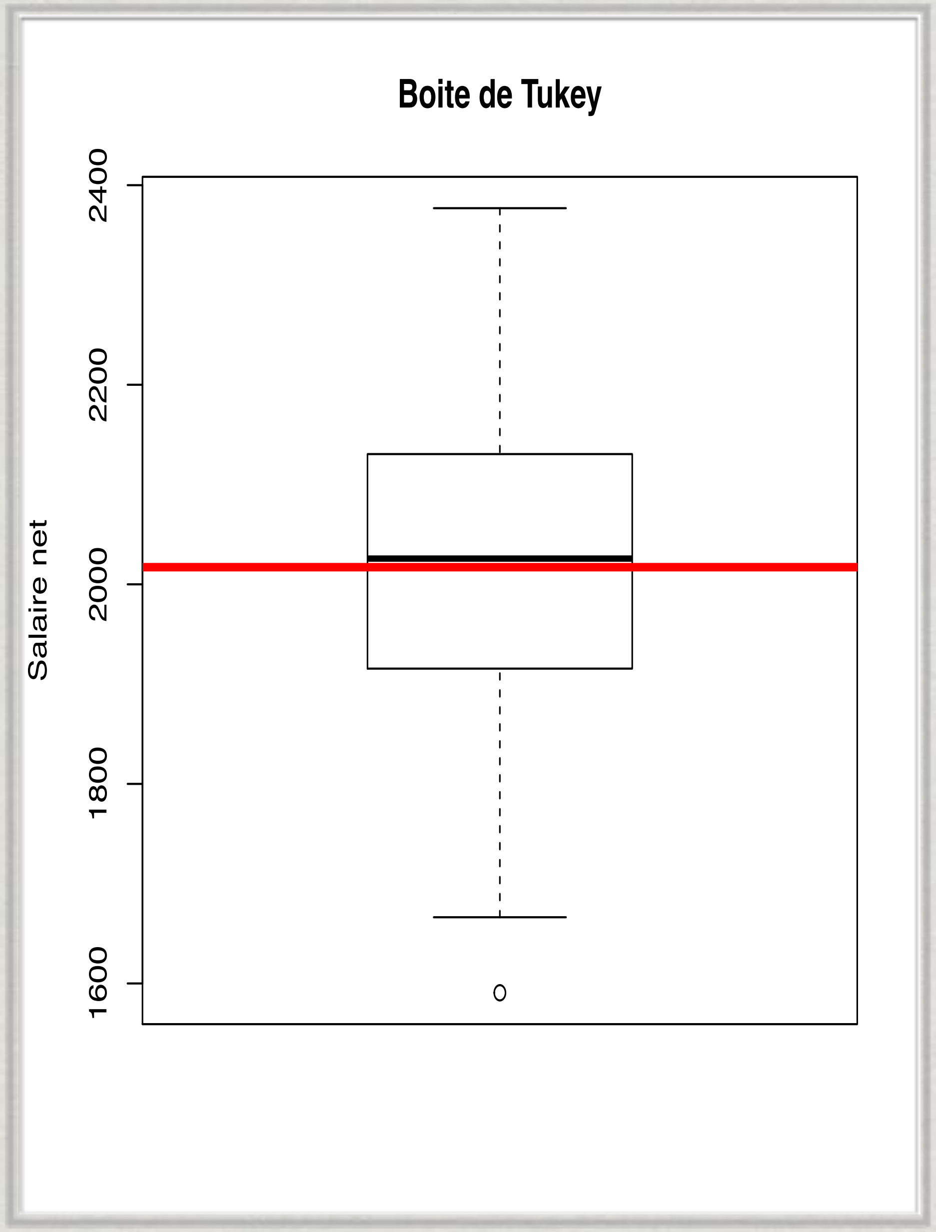
  Exemple : en 2018 le salaire médian des français était de 1710 euros net par mois alors que le salaire net moyen des français était de 2250 euros net par mois.
- Moyenne et médiane ne sont pas nécessairement égales. Dans cet exemple, on voit même que notre moyenne est biaisée par les très grands salaires alors qu'une grande majorité de la population a un salaire inférieur à la moyenne.

La Médiane

- Cette valeur n’est que très rarement utilisée toute seule.
- Pour se faire une meilleure idée de la distribution, il est d’usage de considérer les différents quantiles de la distribution (de notre échantillon).
- Pour rappel, le quantile d’ordre α, noté q_α, relativement à un échantillon, est la valeur de l’échantillon pour laquelle une proportion α des individus ont une valeur inférieure à q_α.
- Une représentation très utilisée pour représenter quelques quantiles important est la « boîte de Tukey » ou la « boite à moustaches ».

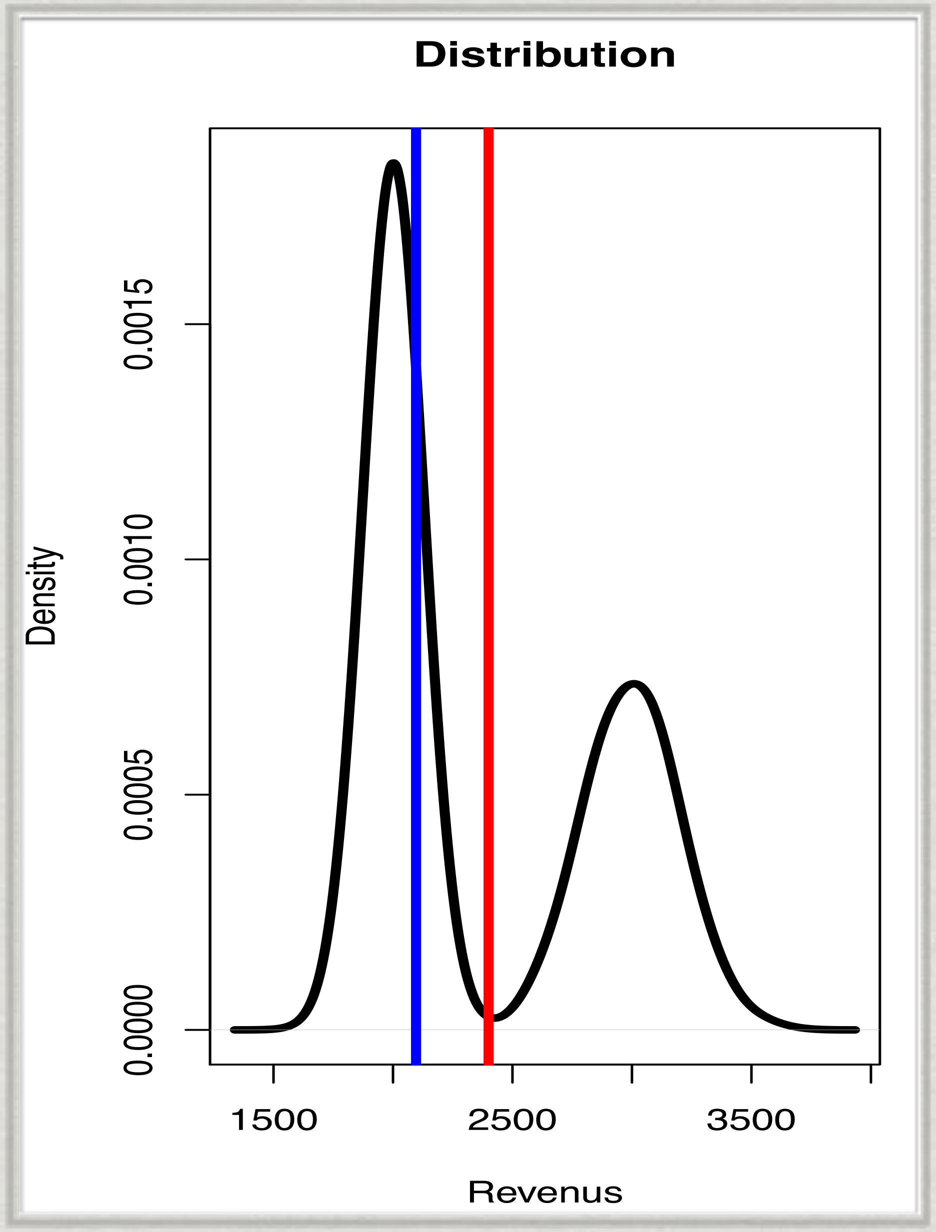
Boite de Tukey

- La valeur moyenne ici de 2010
- La valeur médiane (en noire) est 2030, donc très légèrement supérieure à la moyenne (en rouge)
- Les limites de la boîte représente les 1er et 3ème quartiles (25% et 75% des valeurs prises pas l'échantillon se trouvent dans cette boîte)
- Les moustaches ont une longueur maximale égale 1.5 l'intervalle inter-quartiles
- Le point se trouvant au delà de la moustache est un **point extrême ou une valeur aberrante car très éloignée de la distribution**



Quelques remarques

- Le mode est une valeur centrale ayant une signification différente de la moyenne (en rouge) dans le cas de l'étude d'une **variable quantitative.**
- De même que la médiane (en bleue), si ces deux valeurs sont proches en générales, on verra un critère reposant sur **l'écart-type** qui vous indiquera quelle mesure il est préférable d'utiliser.
- **Rappels :**
  - mode - moyenne - médiane se calculent pour une **variable quantitative**
  - Seul le mode est utilisé comme valeur centrale pour une **variable qualitative**



Exemple d'application

- La table suivante nous renseigne sur le chiffre d'affaire généré sur une période d'une semaine par une enseigne de la mode.

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
C.A.	1245	1379	1980	1150	1670	2689	2354

1. Calculer le chiffre d'affaire moyen sur une semaine
2. Déterminer le mode de cette distribution, *i.e.* le jour de la semaine où l'enseigne a généré son chiffre d'affaire le plus important
3. Déterminer le chiffre d'affaire médian

Exemple d'application

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
C.A.	1245	1379	1980	1150	1670	2689	2354

1. Le chiffre d'affaire moyen $\bar{x}$ de l'enseigne généré par jour est égal à

$$\bar{x} = \frac{1245 + 1379 + 1980 + 1150 + 1670 + 2689 + 2354}{7} = 1781$$

2. Le mode de cette distribution est le samedi et correspond au jour où le chiffre d'affaire généré est le plus important.

3. Pour déterminer le chiffre d'affaire médian, on commence par ordonner les différents de C.A. par ordre croissant, puis on détermine la valeur de notre échantillon qui le sépare en deux échantillons de taille égale.
   Le chiffre d'affaire médian est égal à 1670.

MESURES DE DISPERSION

ECART-TYPE - ÉTENDUE - FORME D’UNE DISTRIBUTION

Ecart-type

- La variance mesure **l'écart quadratique moyen des valeurs prises par l'échantillon à la valeur moyenne de l'échantillon.**
- Il s'agit donc d'une mesure de **dispersion des valeurs autour de la valeur moyenne** $\bar{x}$.
- L'écart type non biaisé s d'un échantillon est alors défini comme la racine carré de la variance

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

  où $\bar{x}$ désigne la moyenne évaluée sur notre échantillon.
- Tout comme la moyenne ou la médiane, l'écart-type n'a de sens que pour des **variables quantitatives !**

Ecart-type

- Un **écart-type important** indique que les valeurs prises par l'échantillon sont **très dispersées et éloignées de la valeur moyenne.**
- Cela est cependant à nuancer selon les valeurs prises par l'échantillon, il est alors d'usage de considérer le coefficient de variation qui se définie comme **le rapport entre l'écart-type et la moyenne**

$$c_v = \frac{s}{\bar{x}}$$

- Si la valeur du coefficient de variation c_v est plus petite que 1, une bonne valeur centrale est alors **la moyenne**. Dans le cas contraire, on considèrera l**a médiane** qui sera plus représentative de la population.

Etendue

- L’étendue e est une mesure de dispersion qui consiste à regarder l’écart entre la plus grande valeur et la plus petite prisent par notre échantillon.

$$e = \max(x_i) - \min(x_i)$$

- Bien que moins utilisée, elle permet cependant d’avoir une idée de l’étalement des valeurs.

- Si elle est utilisable pour une variable quantitative, elle trouve cependant tout son sens dans l’étude d’une **variable qualitative numérique.**

Exemple : on peut mesurer l’étendue des notes ou d’un jugement de valeurs concernant l’évaluation d’un produit par des clients

Exemple d’application

- Reprenons l’exemple d’ étude du C.A. d’une enseigne.

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
C.A.	1245	1379	1980	1150	1670	2689	2354

1. Déterminer la variance de chiffre d’affaire de la semaine pour cette enseigne

2. Déterminer l’étendue associée à cet échantillon

Exemple d'application

1. On commence par calculer la variance s^2 comme la sommes des carrés des écarts à la moyenne $\bar{x}$

$$s^2 = \frac{1}{6}\left((1245 - 1781)^2 + (1379 - 1781)^2 + \ldots + (2354 - 1781)^2\right) = 34\ 197.7$$

2. Pour le calcul de l'étendue e , on commence par chercher les valeurs maximales et minimales de notre échantillon

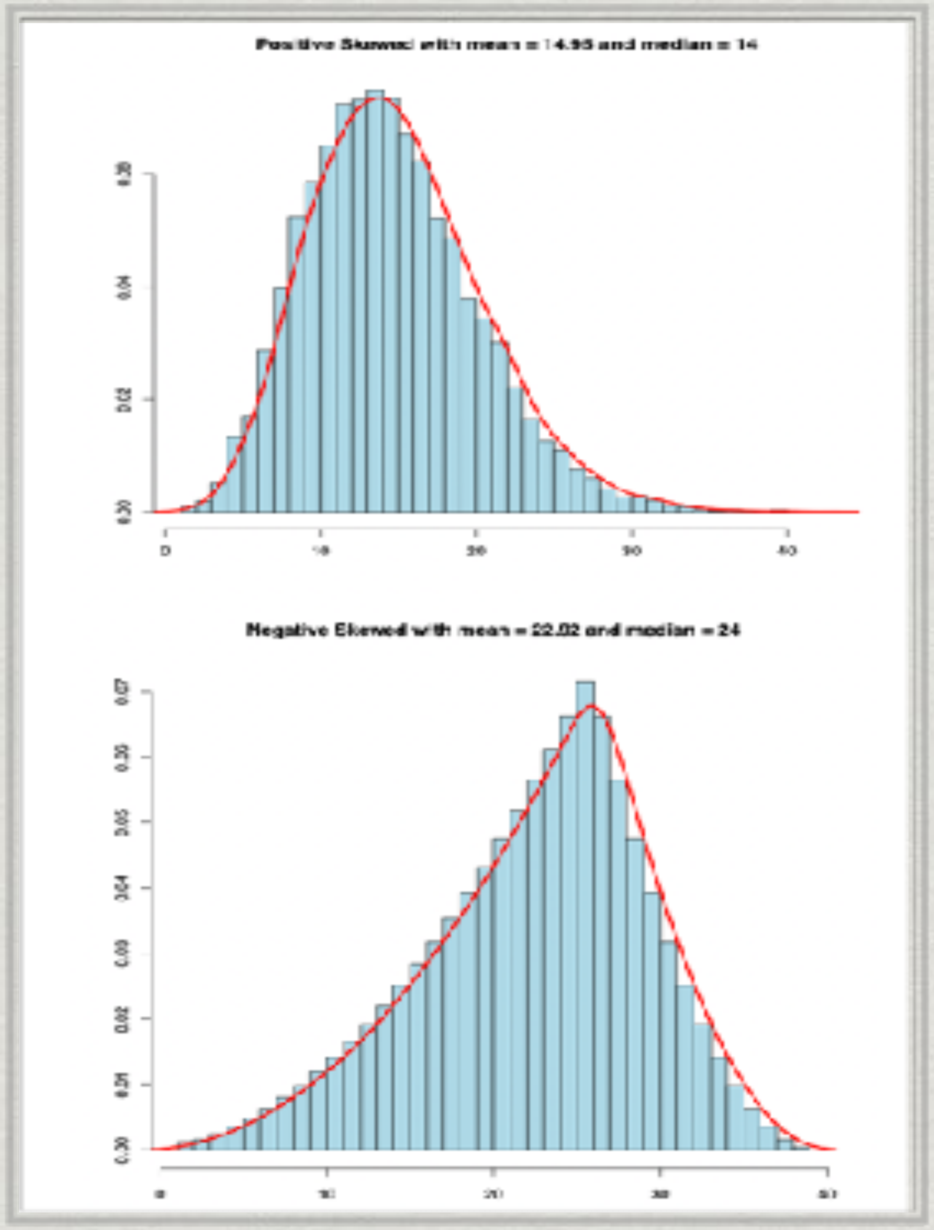
$$e = 2689 - 1150 = 1539$$

Forme : Asymétrie

- On peut juger de la symétrie de la distribution en calculant le moment centré réduit d'ordre 3 sur notre échantillon

$$\gamma = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

- On peut également apprécier l'asymétrie de la distribution en comparant moyenne et médiane
  - si moyenne > médiane
    distribution étalée à droite ou **asymétrie positive**
  - si moyenne < médiane :
    distribution étalée à gauche ou **asymétrie négative**

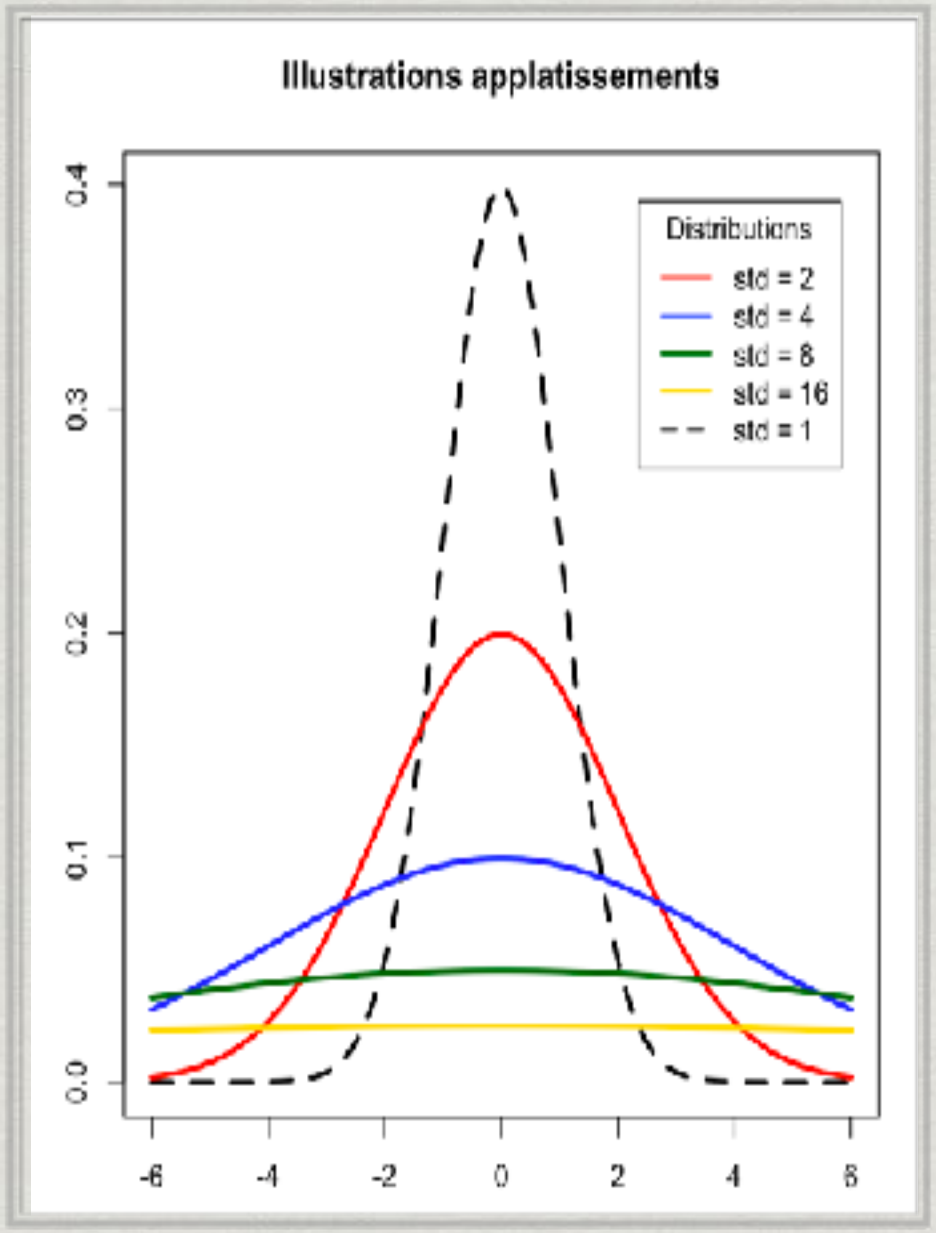


Forme : Aplatissement

- L'aplatissement ou le kurtosis κ mesure la répartition des masses autour de la valeur centrale.

- Il se détermine en calculant le moment centré réduit d'ordre 4 sur notre échantillon

$$\kappa = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4}$$



	Effectifs	% Obs.
Pas du tout	5	2,9%
Plutôt non	11	6,3%
moyennement	85	48,6%
Plutôt oui	69	39,4%
Tout à fait	5	2,9%
Total	**175**	**100%**

Réponses effectives : 175
Taux de réponse : 100%

Non-réponse(s) : 0
Modalités les plus citées : moyennement; Plutôt oui; Plutôt non

MISE EN PRATIQUE

MANIPULATION SUR SPHINX CAMPUS :
RIEN À FAIRE, CES VALEURS SONT CALCULÉES AUTOMATIQUEMENT