



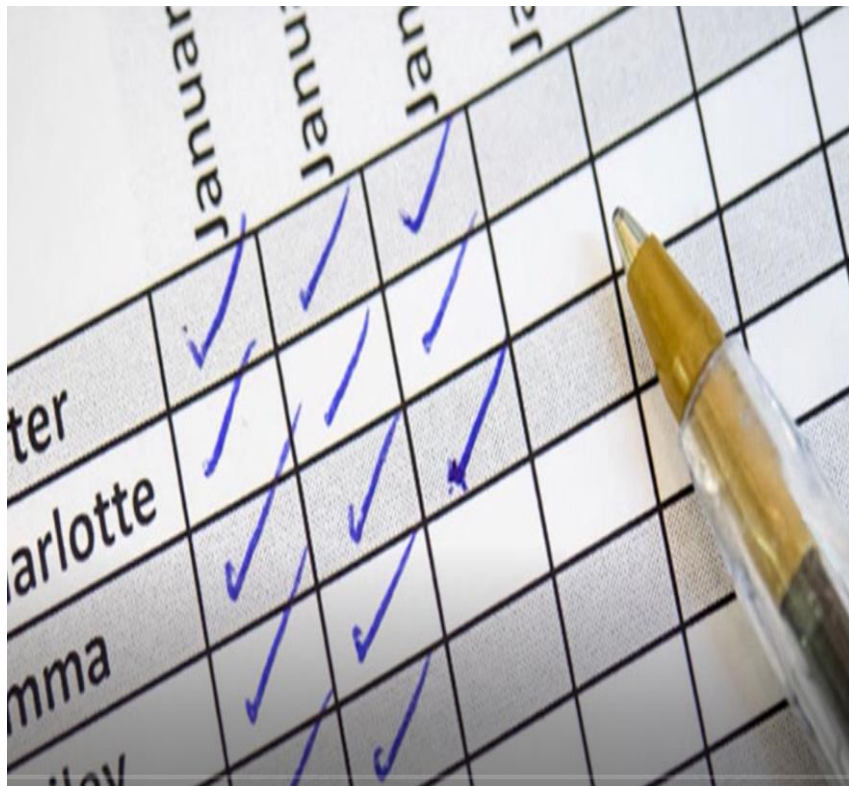
Cours

Statistiques Uni-variées

K. BELATTAR,

Département Informatique - Université d'Alger 1

Introduction aux données



| Result of first semester | | | | | | | |
|--------------------------|---------------|---------|-------|---------|-------|------------|------------|
| Roll no. | Students Name | English | Maths | Science | total | percentage | Class rank |
| 1 | Leena | 67 | 71 | 65 | 203 | 84.58 | 8 |
| 2 | Ritika | 65 | 78 | 70 | 213 | 88.75 | 3 |
| 3 | Pradhnya | 64 | 76 | 68 | 208 | 86.67 | 6 |
| 4 | Smita | 69 | 73 | 64 | 206 | 85.83 | 7 |
| 5 | Rashmi | 64 | 68 | 58 | 190 | 79.17 | 19 |
| 6 | Prachi | 68 | 75 | 59 | 202 | 84.17 | 10 |
| 7 | Sheetal | 67 | 69 | 64 | 200 | 83.33 | 11 |
| 8 | Garima | 64 | 70 | 65 | 199 | 82.92 | 13 |
| 9 | Angel | 68 | 73 | 76 | 217 | 90.42 | 1 |
| 10 | Bindiya | 69 | 73 | 70 | 212 | 88.33 | 4 |
| 11 | Vinita | 65 | 79 | 59 | 203 | 84.58 | 8 |
| 12 | Jyoti | 68 | 78 | 70 | 216 | 90.00 | 2 |
| 13 | Dani | 64 | 68 | 59 | 191 | 79.58 | 18 |
| 14 | Vijay | 61 | 65 | 59 | 185 | 77.08 | 22 |
| 15 | Sagar | 62 | 67 | 63 | 192 | 80.00 | 17 |
| 16 | Jyotiram | 63 | 66 | 60 | 189 | 78.75 | 20 |
| 17 | Ajmal | 67 | 65 | 54 | 186 | 77.50 | 21 |
| 18 | vaibhav | 68 | 75 | 57 | 200 | 83.33 | 11 |
| 19 | Ravi | 68 | 71 | 59 | 198 | 82.50 | 15 |
| 20 | Rahul | 69 | 79 | 63 | 211 | 87.92 | 5 |
| 21 | Suraj | 64 | 65 | 64 | 193 | 80.42 | 16 |
| 22 | Abhijit | 68 | 64 | 67 | 199 | 82.92 | 13 |
| 23 | Danial | 63 | 60 | 61 | 184 | 76.67 | 23 |

Introduction aux données



Introduction aux données

Données

10 \$

« 10 dollars »



Vente: 3.5 million dollars



Vente: 3 million dollars

Information

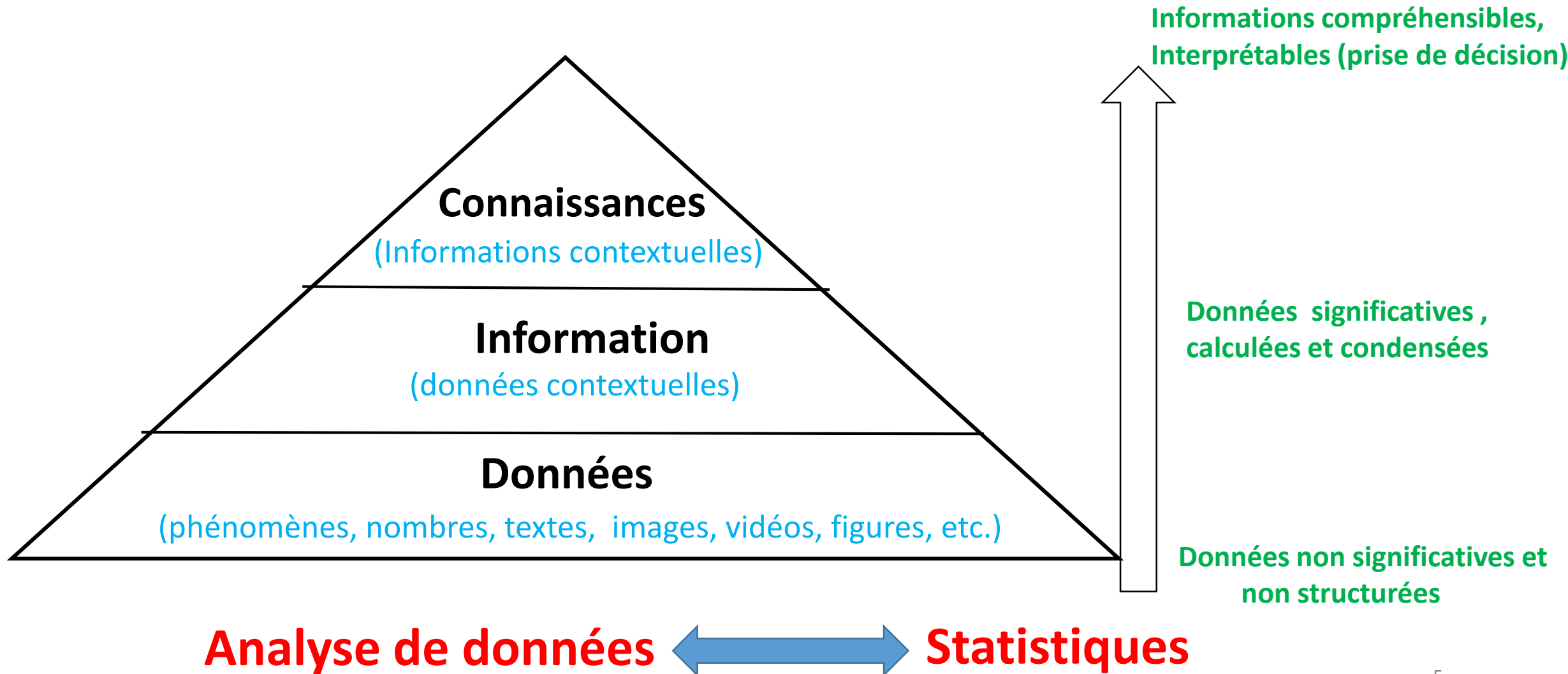
- 5 dollars par unité
- prendre 50 dollars (gain)
- Gain par voiture: 5000 dollars

La voiture jaune est plus chère.

Connaissance

Acheter la voiture rouge (moins chère)

Introduction aux données



Introduction aux données

- ❖ Office Nationale des Statistiques : l'Institution Centrale des Statistiques de l'Algérie chargée de la collecte, du traitement et de la diffusion de l'information statistique socio-économique de la population Algérienne.



<https://www.ons.dz/spip.php?rubrique38>

- ❖ Données relatives aux attributs de la population Algérienne et sur les entreprises.

Manipuler les données

Taux d'alphabétisation de la population → manipuler ces données qui consiste à :

1) Collecter des données

- Nom
- Age
- Genre
- Date et lieu de naissance,
- Etat civil,
- Occupation,
- Education,
- etc.

Manipuler les données

Manipuler ces données consiste à :

2) Organiser des données

| | Taille de population | Chiffre |
|----------------|---|------------|
| Sexe | Taille de la population (sexe male) | 1210854877 |
| | Taille de la population (sexe femelle) | 6232770258 |
| Région | Taille de la population (régions rurale) | 587584719 |
| | Taille de la population (régions urbaine) | 8337734885 |
| Sexe et Région | Taille de la population (régions rurale, sexe male) | 3777106125 |
| | Taille de la population (régions rural, sexe femelle) | 427781058 |
| | Taille de la population (régions urbaine, sexe male) | 195489200 |

Manipuler les données

Manipuler ces données consiste à :

3) Analyser les données

| Algérie | Analyse (2020) |
|--------------------------------------|------------------------------|
| Densité de la population | 382 personnes par Km carrées |
| Ratio sexe (femelles par 1000 males) | 943 |
| Taux d'alphabétisation | 74,04% |



Résultat d'analyse

Manipuler les données

Manipuler ces données consiste à :

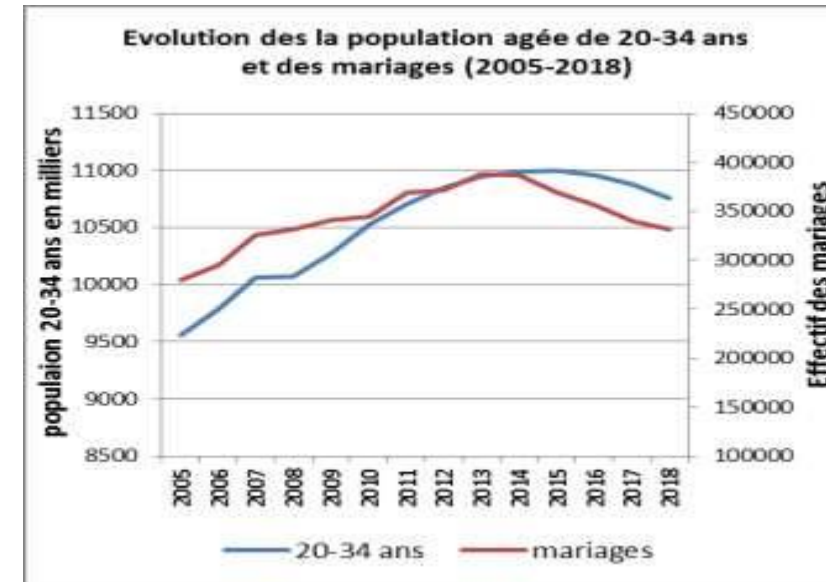
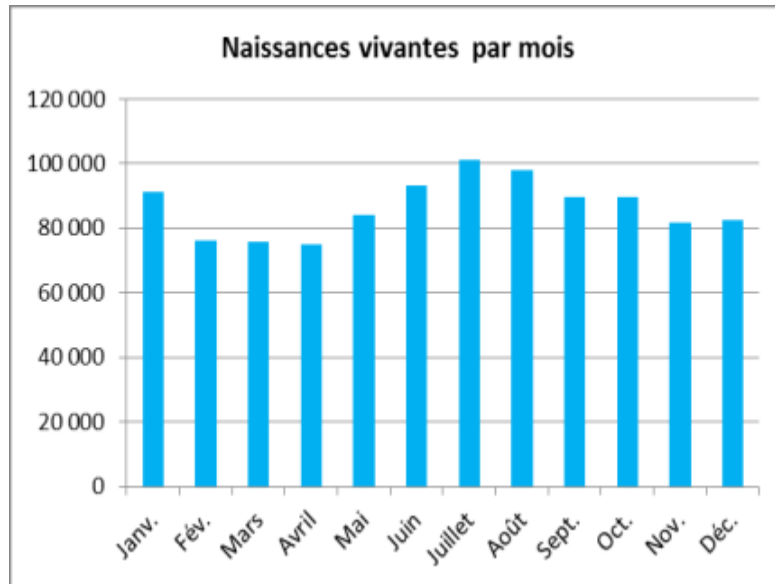
4) Interpréter les données

| Algérie | Analyse (2020) | Analyse (2018) |
|--------------------------------------|---------------------------|----------------|
| Densité de la population | 382 personnes par carrées | 324 |
| Ratio sexe (femelles par 1000 males) | 943 | 933 |
| Taux d'alphabétisation | 74,04% | 65.38% |

Manipuler les données

Manipuler ces données consiste à :

5) Représenter les données



Statistiques



(2) Organisation de données



(1) Collection de données



(3) Analyse de données



Statistiques

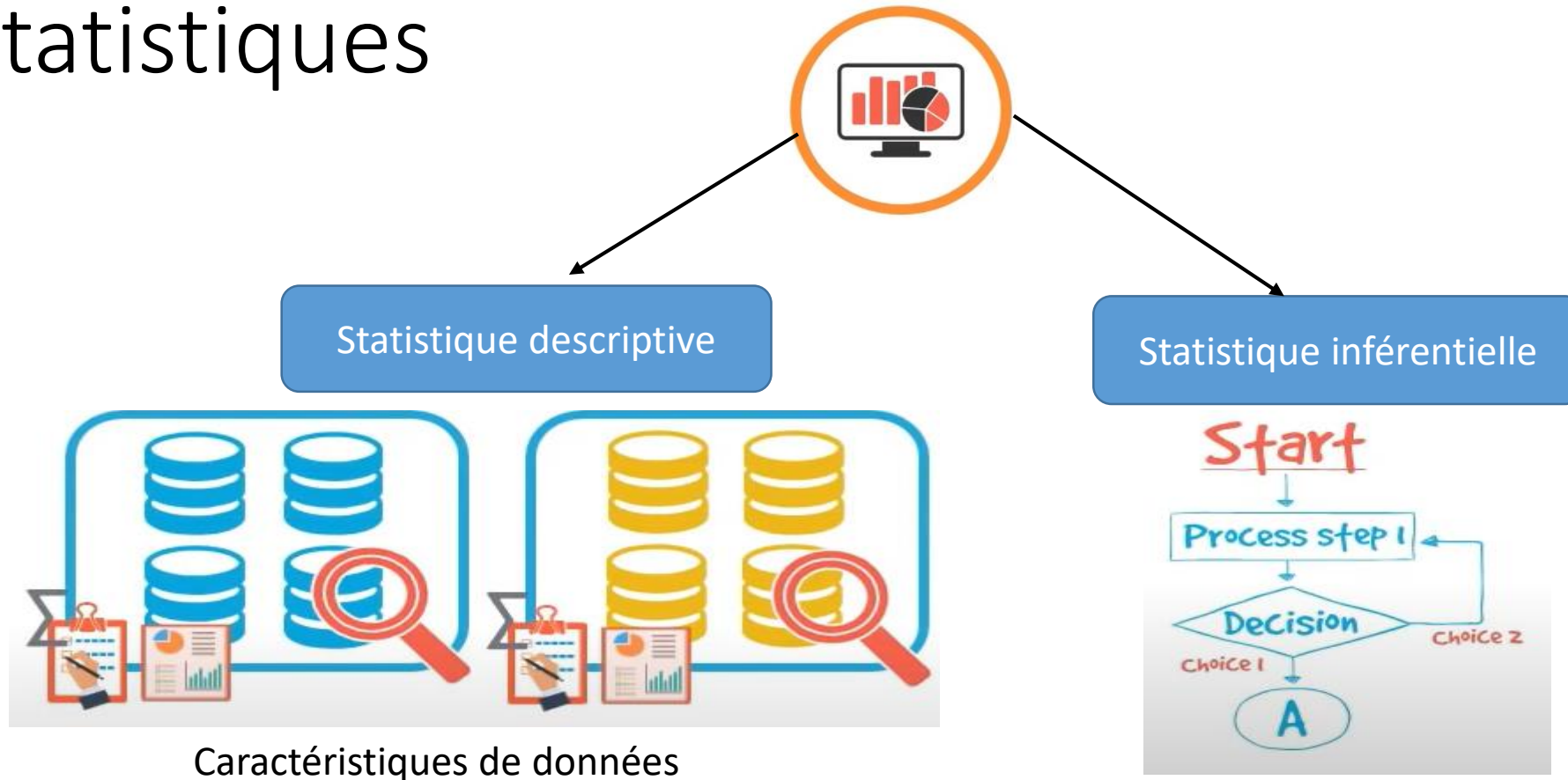


(5) Représentation de données



(4) Interprétation de données

Statistiques



Moyenne

Ecart type

Mode

Corrélation

etc.

→ Mesures statistiques (uni-variées, bi-variées)

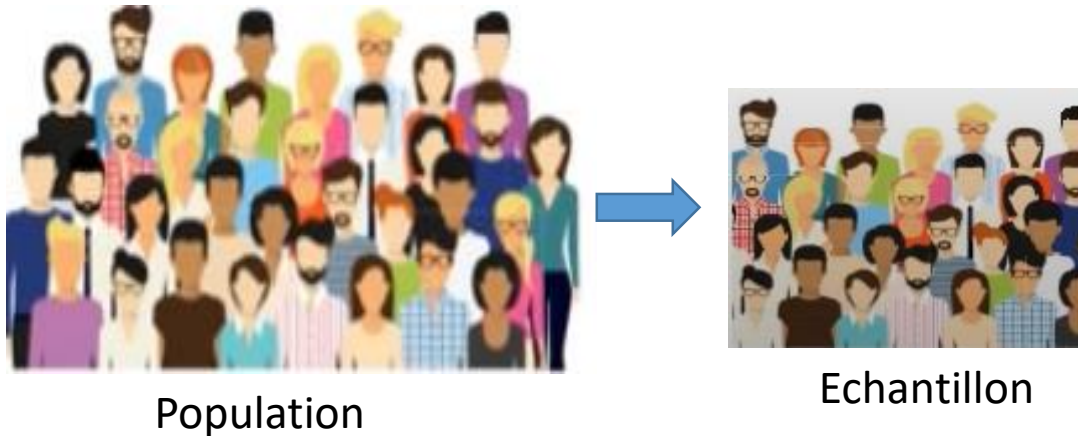
→ Graphiques de visualisation

Inférer une conclusion

→ **Théorie des probabilités**

Concepts statistiques

- ❖ **Population:** ensemble sur lequel les données sont collectées.
- ❖ **Individu :** tout élément de la population.
- ❖ **Echantillon:** sous ensemble de la population à analyser.



Concepts statistiques

- ❖ **Variable statistique**: est une caractéristique d'un individu de la population décrivant sa qualité et quantité.
- ❖ **Variables quantitatives** (**rapport**, **intervalle**)
- ❖ **Variables qualitatives** (**nominales**, **ordinales**)
- ❖ **Variables discrètes et continues**

Langage

Région

Age



Genre

Poids

Taille



Poids d'une personne



Nombre des personnes



Couleur



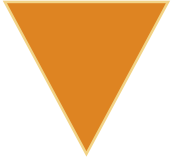
Degré dommage accidentel

Types des mesures statistiques uni-variées

Mesures de
fréquence

Mesures de tendance
centrale

Mesures de
dispersion

- 
- Variables quantitatives
 - Variables qualitatives

- **Effectif , effectif cumulé croissant / décroissant**
- **Fréquence, fréquence cumulée croissante / décroissante**
- **Pourcentage, pourcentage cumulé**

Mesures de fréquence

Soit les valeurs $\{x_1, x_2, \dots, x_N\}$ un ensemble de N observations (série statistique à une variable) de l'attribut X .

| Mesure de fréquence | Formule mathématique |
|------------------------------|--|
| Effectif (fréquence absolue) | n_i : nombre d'apparition d'une valeur x_i |
| Effectif cumulé croissant | $N_i^{\uparrow} = \sum_{k=1}^i n_k$ |
| Effectif cumulé décroissant | $N_i^{\downarrow} = N - \sum_{k=1}^i n_k$ N : somme totale des effectifs ($n_1 + \dots + n_N$). |

Mesures de fréquence

| Mesure de fréquence | Formule mathématique |
|--------------------------------|---|
| Fréquence (relative) | $f_i = \frac{n_i}{N}$ |
| Fréquence cumulée croissante | $F_i^{\uparrow} = \frac{\sum_{k=1}^i n_k}{N}$ |
| Fréquence cumulée décroissante | $F_i^{\downarrow} = 1 - \frac{\sum_{k=1}^i n_k}{N}$ |
| Pourcentage | $P = f_i * 100\%$ |

Mesures de fréquence

Exemple 1:

On s'intéresse à la variable qualitative « état-civil » pour 20 personnes.

célibataire (C), marié(e) (M), veuf(ve) (V), divorcée (D).

Considérons la série statistique suivante : M M V V C M C C C M C M V
C M D D C C M.

Calculer l'effectif, la fréquence, l'effectif cumulé et la fréquence cumulé de chaque valeur de la variable.

Mesures de fréquence

Exemple 2:

On s'intéresse à la variable quantitative continue « nombre d'enfant allergique à la coste ».

| Nombre d'enfants | [11-21[| [21-31[| [31-41[| [41-51[|
|------------------|---------|---------|---------|---------|
| n_i (Effectif) | 16 | 40 | 34 | 10 |

Calculer la fréquence, l'effectif cumulé et la fréquence cumulé de chaque valeur de la variable.

Types des mesures statistiques uni-variées

Mesures de
fréquence

Mesures de
tendance centrale

Mesures de
dispersion

Variables quantitatives

Les valeurs de données sont accumulés
dans le centre de la distribution.

- **Moyenne**
- **Médiane**
- **Mode** (**et** variables qualitatives)

| Mesure de tendance centrale | Formule mathématique | Caractéristiques |
|--|--|--|
| Moyenne arithmétique d'une variable discrète | $\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i$ | <ul style="list-style-type: none"> - Considérer toutes les valeurs de la série - Sensible aux outliers |
| Moyenne arithmétique d'une variable continue (classes) | $\bar{X} = \frac{\sum_{i=1}^M n_i c_i}{N}$ <p> c_i : centre de chaque classe ($\frac{e_{i-1} + e_i}{2}$) M : nombre de classes N : somme des effectifs pour chaque classe ($N = \sum_{i=1}^M n_i$) </p> | <ul style="list-style-type: none"> - Considérer toutes les valeurs de la série - Sensible aux outliers |
| Moyenne arithmétique pondérée | $\bar{X} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$ | <ul style="list-style-type: none"> - Considérer toutes les valeurs de la série - Sensible aux outliers |
| Moyenne Harmonique | $H = \frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}$ | <ul style="list-style-type: none"> - Considérer toutes les valeurs de la série - Sensible aux outliers |
| Moyenne géométrique | $G = \frac{1}{N} \sum_{i=1}^N n_i \log x_i$ | <ul style="list-style-type: none"> - Considérer toutes les valeurs de la série - Sensible aux outliers |
| Moyenne quadratique | $Q = \sqrt{\frac{1}{N} \sum_{i=1}^N n_i x^2}$ | <ul style="list-style-type: none"> - Considérer toutes les valeurs de la série - Sensible aux outliers |

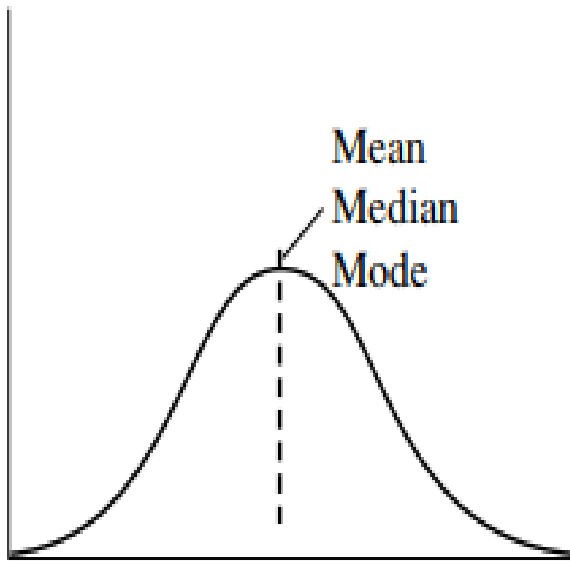
Mesures de tendance centrale

| Mesure de tendance centrale | Formule mathématique | Caractéristiques |
|---|---|------------------|
| Médiane d'une variable discrète | Valeur médiane d'un ensemble ordonnée de valeur | / |
| Médiane d'une variable continue (classes) | $medianex = e_{i-1} + a_i \left(\frac{N/2 - N_i}{n_{mediane}} \right)$ <p> <i>e_{i-1}</i>: extrémité inférieure de la classe médiane [<i>e_{i-1}</i>, <i>e_i</i>[. <i>N_i</i>: effectif cumulé de classe qui précède la classe médiane. <i>n_{mediane}</i>: effectif de la classe médiane <i>a_i</i>: amplitude de la classe médiane (<i>e_i</i> - <i>e_{i-1}</i>). </p> | / |

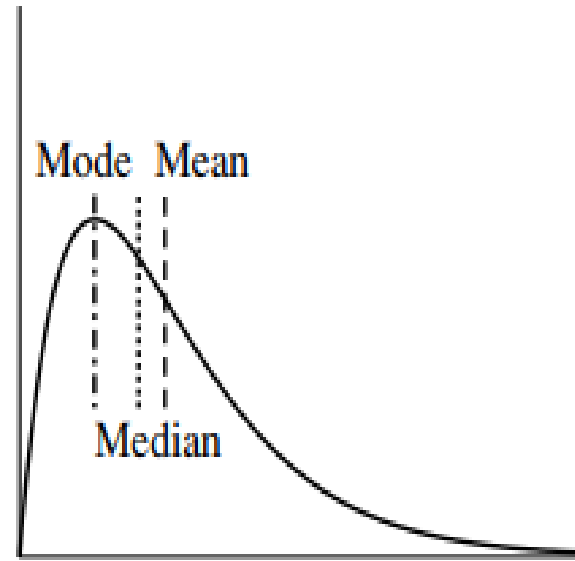
Mesures de tendance centrale

| Mesure de tendance centrale | Formule mathématique | Caractéristiques |
|--|---|---|
| Mode d'une variable discrète | La valeur qui a la plus grande fréquence. | Un (unimodale) ou plusieurs (bimodale, trimodale, etc.) modes |
| Mode d'une variable continue (classes) | $Mode = e_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2}$ <p> e_{i-1}: extrémité inférieure de la classe modale $[e_{i-1}, e_i[$. Δ_1 : différence entre l'effectif ou fréquence de la classe précédente et celui ou celle de la classe modale. Δ_2 : différence entre l'effectif ou fréquence de la classe modale et celui ou celle de la classe suivante. a_i : amplitude (longueur) de la classe modale $(e_i - e_{i-1})$. </p> | / |

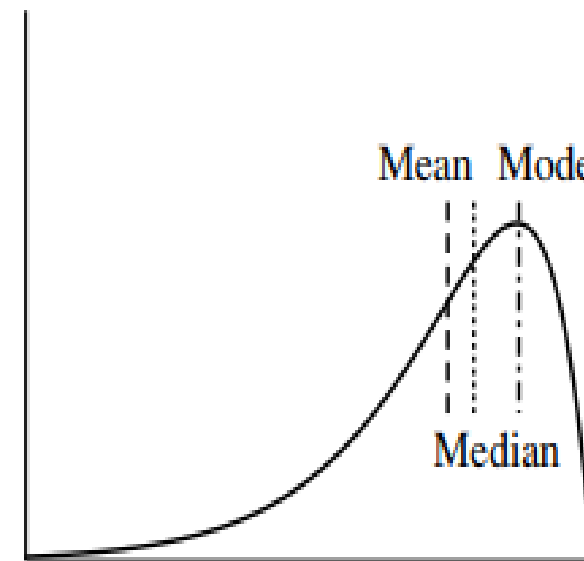
Mesures de tendance centrale



(a) Symmetric data



(b) Positively skewed data



(c) Negatively skewed data

Mesures de tendance centrale

Exemple 1:

Soit le score total des étudiants dans 5 communications[/100 pour chacune]:

60 70 74 78 78

Calculer la moyenne , la médiane et le mode de cette série statistique.

Moyenne=72

Mode= 78

Médiane= 74

Types des mesures statistiques uni-variées

Mesures de
fréquence

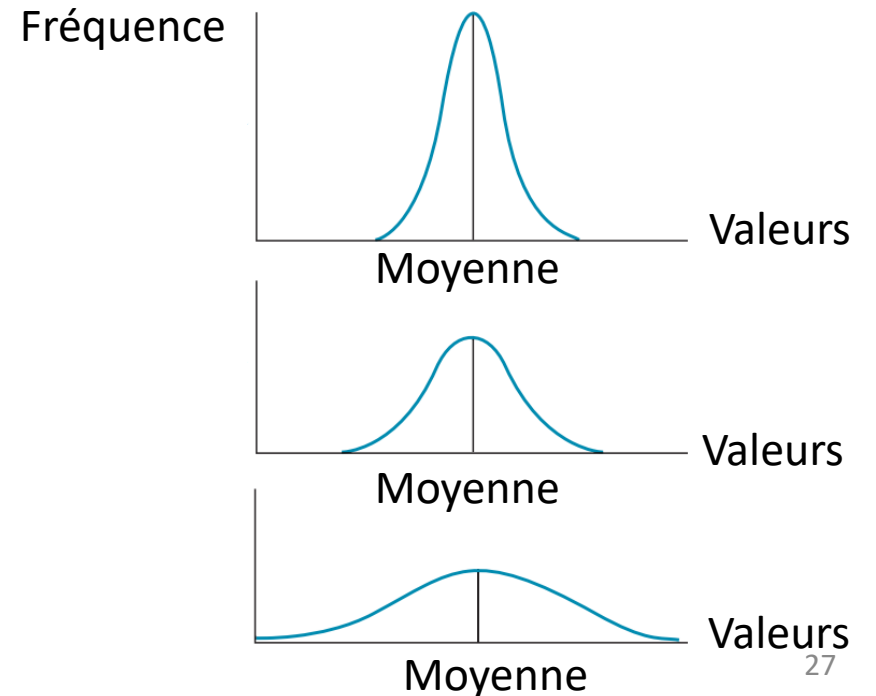
Mesures de tendance
centrale

Mesures de
dispersion

Décrire la variabilité d'une distribution
par rapport à une variable particulière.

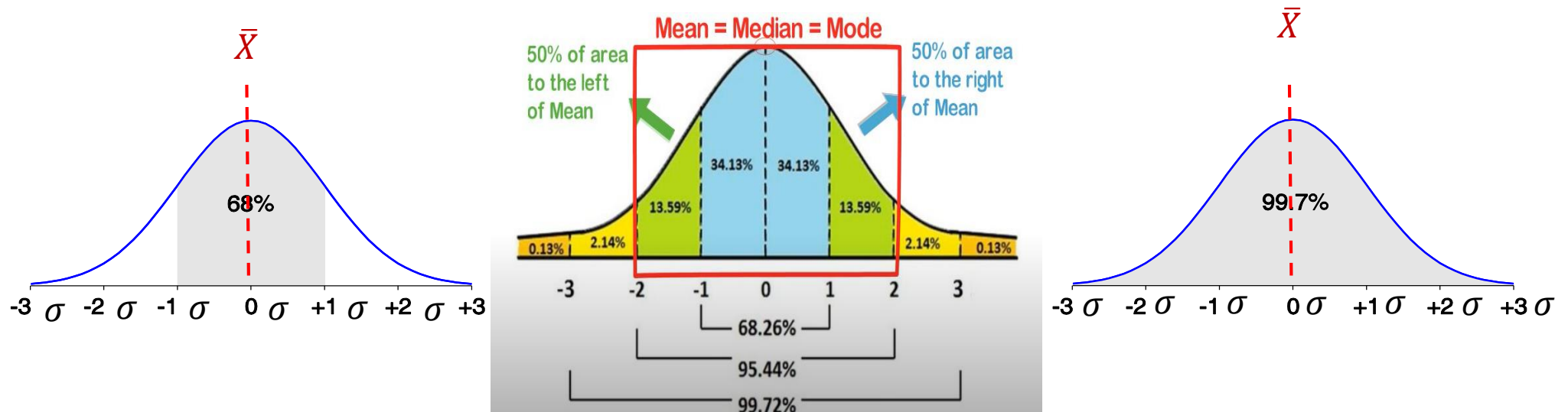
- **Etendue**
- **Quantiles**
- **Ecart inter-quantiles**
- **Variance**
- **Ecart-type**
- **etc.**

- Variables quantitatives



Types des mesures statistiques uni-variées

- La courbe de **la distribution normale**:
 - $[\bar{X} \pm \sigma]$: contient approximativement 68% des valeurs.
 - $[\bar{X} \pm 2\sigma]$: contient approximativement 95% des valeurs
 - $[\bar{X} \pm 3\sigma]$: contient approximativement 99.7 % des valeurs



Mesures de dispersion

Quantiles

| Mesure de dispersion | Formule mathématique | |
|--|--|---|
| Etendue d'une variable discrète | $E = x_{max} - x_{min}$ | |
| Etendue d'une variable continue (classes) | $E = e_{max} - e_{min}$ | |
| Quartiles d'une variable discrète (4 quantiles, $Q_{k=1}, Q_{k=2}, Q_{k=3}$) | $F(Q_1)=25\%$ $F(Q_2)=50\%$ $F(Q_3)=75\%$ | $N(Q_1)=\frac{1}{4}N$ $N(Q_2)=\frac{2}{4}N$ $N(Q_3)=\frac{3}{4}N$ |
| Déciles d'une variable discrète (10 quantiles, $D_{k=1}, D_{k=2}, \dots, D_{k=9}$) | $F(D_1)=10\%$ $F(D_2)=20\%$ $F(D_9)=90\%$ | $N(D_1)=\frac{1}{10}N$ $N(D_2)=\frac{2}{10}N$ $N(D_9)=\frac{9}{10}N$ |
| Centiles d'une variable discrète (100 quantiles, $C_{k=1}, C_{k=2}, \dots, C_{k=99}$) | $F(C_1)=1\%$ $F(C_2)=2\%$ $F(C_{99})=99\%$ | $N(C_1)=\frac{1}{100}N$ $N(C_2)=\frac{2}{100}N$ $N(C_{99})=\frac{99}{100}N$ |

Mesures de dispersion

| Mesure de dispersion | Formule mathématique |
|---|---|
| Quantiles d'une variable continue (classes) | $Quantile_q = e_{i-1} + a_i \frac{\frac{k}{q} * N - N_i}{n_i}$ <p> k: numéro de quantile ($0 < k < q$) q: nombre total des quantiles Quartiles ($0 < k < 4$), Déciles ($0 < k < 10$), Centiles ($0 < k < 100$) e_{i-1}: extrémité inférieure d'une classe quantile $[e_{i-1}, e_i[$ a_i: amplitude (longueur) de la classe quantile ($e_i - e_{i-1}$) N_i: effectif cumulé de classe qui précède la classe quantile $[e_{i-1}, e_i[$ n_i: effectif de la classe quantile $[e_{i-1}, e_i[$ (qui contient le quantile à calculer). </p> |
| Ecart inter-quantiles | Ecart inter-quartiles: $I_Q = Q_3 - Q_1$ (contenant 50% des observations) Ecart interdéciles: $I_D = D_9 - D_1$ (contenant 80% des observations) Ecart intercentiles: $I_C = C_{99} - C_1$ (contenant 98% des observations) |

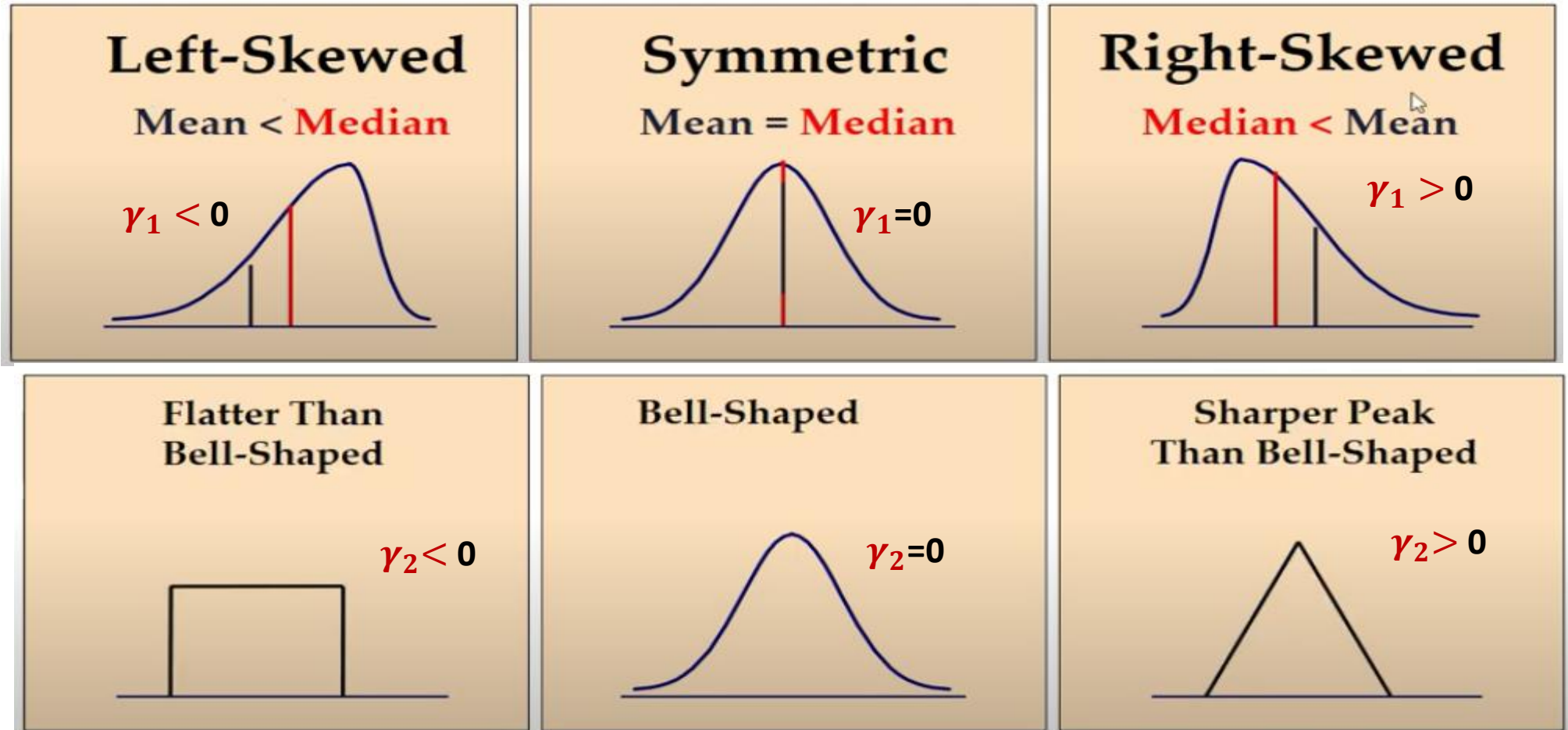
Mesures de dispersion

| Mesure de dispersion | Formule mathématique |
|----------------------------------|---|
| Variance d'une variable discrète | $\sigma^2 = \frac{\sum_{i=1}^N n_i * (x_i - \bar{X})^2}{N} = (\frac{1}{N} \sum_{i=1}^N n_i x_i^2) - \bar{X}^2$ $= \frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N} = (\frac{1}{N} \sum_{i=1}^N x_i^2) - \bar{X}^2$ |
| Variance d'une variable continue | $\sigma^2 = \frac{\sum_{i=1}^M (C_i - \bar{X})^2}{N}, N = \sum_{i=1}^M n_i$ $\sigma^2 = \frac{\sum_{i=1}^M n_i * (C_i - \bar{X})^2}{N}$ <p>C_i : Centre de chaque classe M : nombre de classes</p> |
| Ecart type | $\sigma_X = \sqrt{\sigma^2}$ |
| Coefficient de variation | $CV = \frac{\sigma_X}{ \bar{X} }$ |

Mesures de dispersion

| Mesure de dispersion | Formule mathématique | Caractéristiques |
|--|--|---|
| Coefficient d'asymétrie (skewness) d'une variable discrète | <p>Pour un échantillon:</p> $\gamma_1 = \frac{\mu_3}{\sigma^3} \text{ avec } \mu_3 = \frac{1}{(N-1)(N-2)} \sum_{i=1}^N n_i (x_i - \bar{X})^3$ <p>Pour une population:</p> $\gamma_1 = \frac{\mu_3}{\sigma^3} \text{ avec } \mu_3 = \frac{1}{N} \sum_{i=1}^N n_i (x_i - \bar{X})^3$ | <ul style="list-style-type: none"> • $\gamma_1=0 \rightarrow$ distribution symétrique • $\gamma_1<0 \rightarrow$ distribution étalée vers la gauche • $\gamma_1>0 \rightarrow$ distribution étalée vers la droite |
| Coefficient d'aplatissement (kurtosis) d'une variable discrète | <p>Pour un échantillon:</p> $\gamma_2 = \frac{\mu_4}{\sigma^4} \text{ avec } \mu_4 = \frac{1}{(N-1)(N-2)} \sum_{i=1}^N n_i (x_i - \bar{X})^4$ <p>Pour une population:</p> $\gamma_2 = \frac{\mu_4}{\sigma^4} \text{ avec } \mu_4 = \frac{1}{N} \sum_{i=1}^N n_i (x_i - \bar{X})^4$ | <ul style="list-style-type: none"> • $\gamma_2 = 0 (= 3) \rightarrow$ distribution normalisée (non normalisée) normale • $\gamma_2 < 0 (< 3) \rightarrow$ distribution normalisée (non normalisée) plus aplatie • $\gamma_2 > 0 (> 3) \rightarrow$ distribution normalisée (non normalisée) moins aplatie ou pointue |

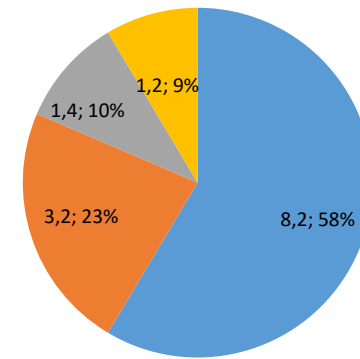
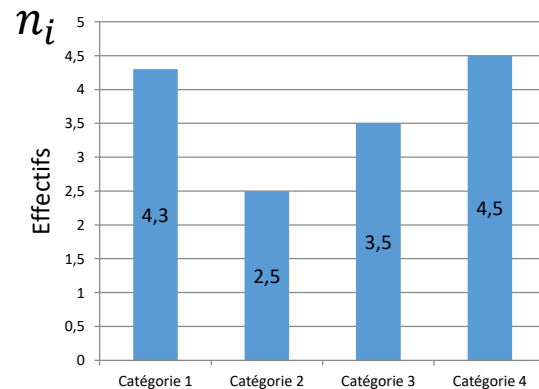
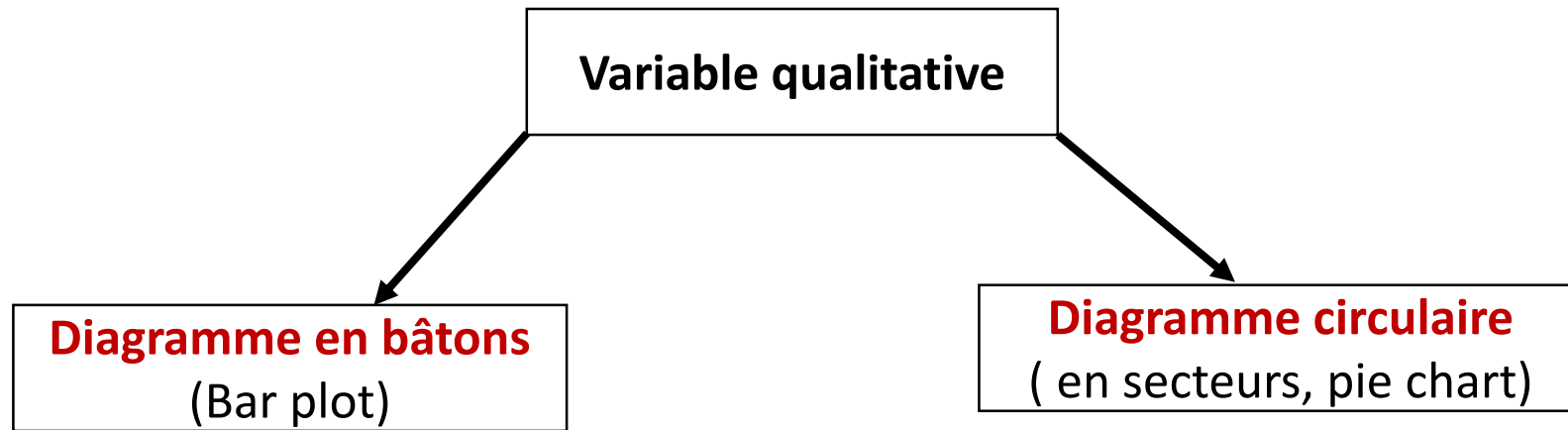
Mesures de dispersion



Visualisation des descriptions statistiques

- **Graphiques:** l'une des techniques de l'exploration de données.
- Convertir les données en formats graphiques (visuel) ou tabulaire tels que les points, les lignes, etc.
 - Représenter les données (individus) par des points. Les coordonnées de points sont définies par les valeurs de variables.
 - Visualiser et analyser les relations entre :
 - 1) les données (si les points forment un cluster ou c'est un point outlier ou encore un autre pattern)
 - 2) les attributs (corrélation, indépendance, causalité).

Visualisation des descriptions statistiques uni-variées

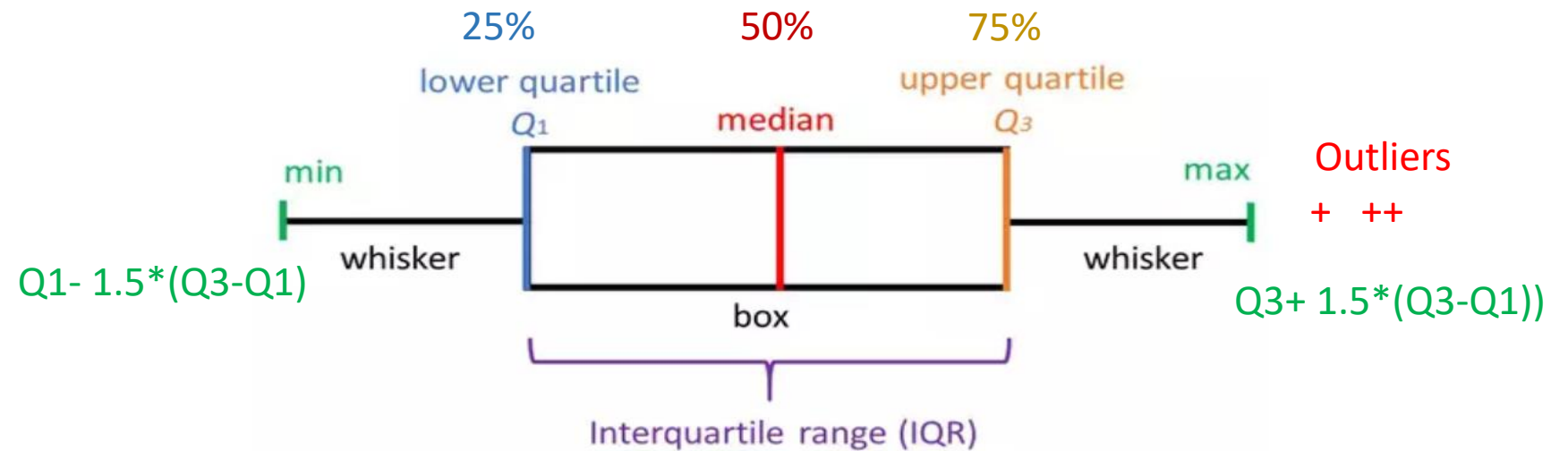
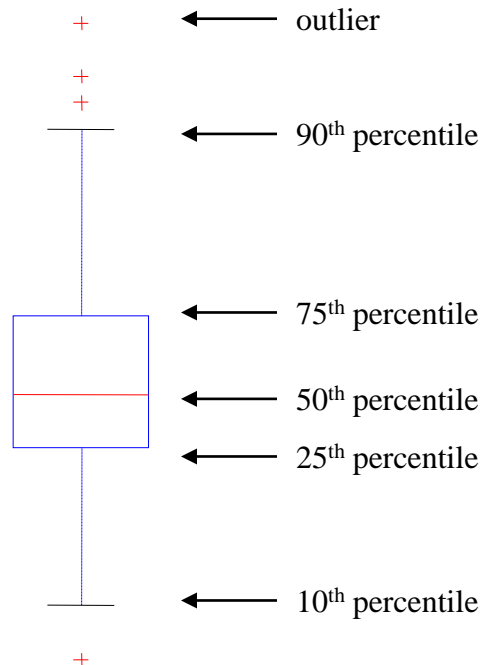


$$Angle_i = 360^\circ * \frac{n_i}{N} = 360^\circ * f_i$$
$$\sum Angle_i = 360^\circ$$

Visualisation des descriptions statistiques uni-variées

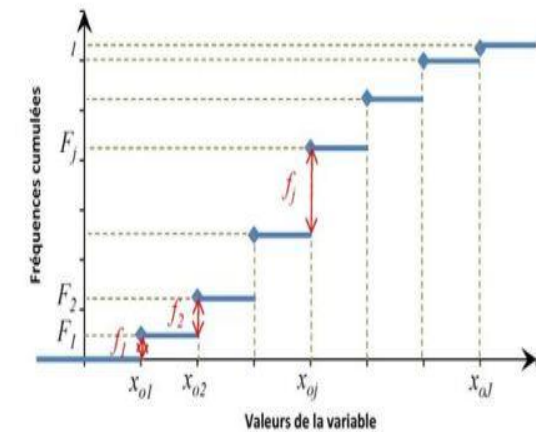
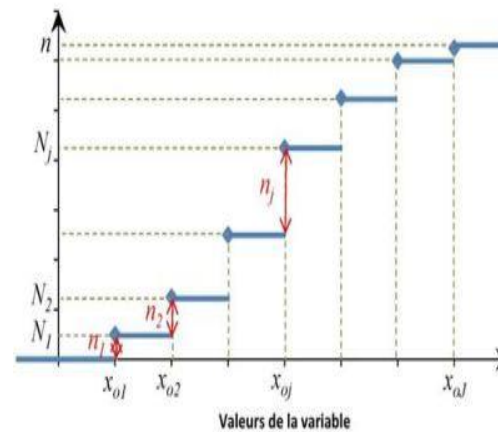
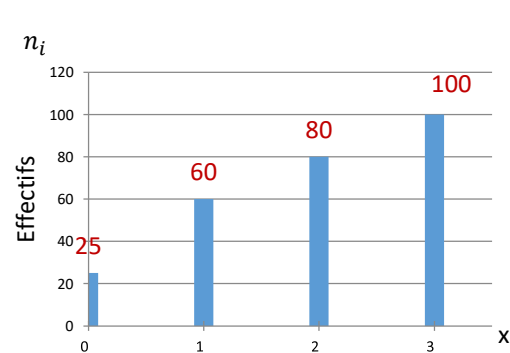
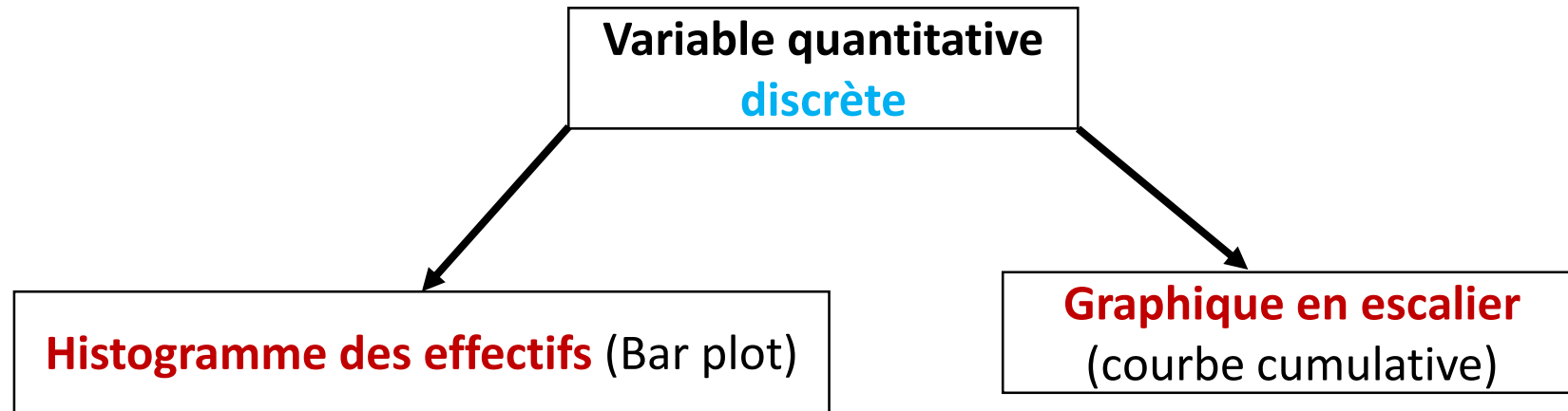
Variable quantitative (**discrète, continue**)

Boîtes à moustaches (Boxplots)

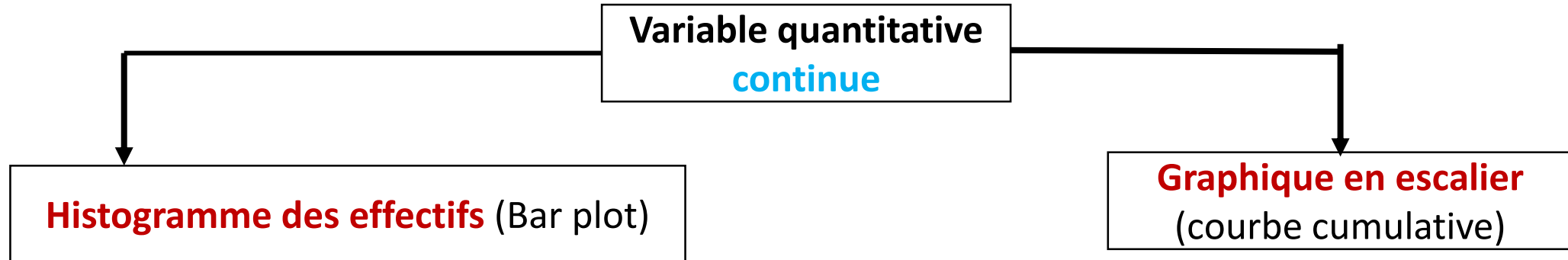


Outlier est à l'extérieur de l'intervalle $[Q1 - 1.5 * (Q3 - Q1), Q3 + 1.5 * (Q3 - Q1)]$

Visualisation des descriptions statistiques uni-variées



Visualisation des descriptions statistiques uni-variées

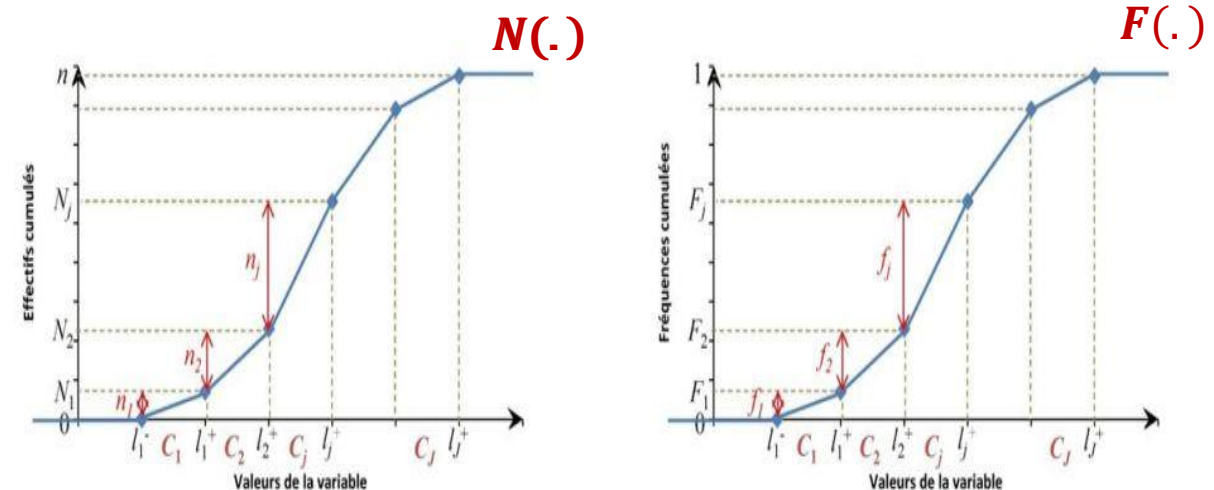


Note : Classes (bin) n'ont pas la même amplitude (a_i):

- 1) Choisir amplitude de base (a_{base})
- 2) Corriger les effectifs de chaque classe:

$$n_{Corrigé\ i} = n_i * \frac{a_{base}}{a_i}$$

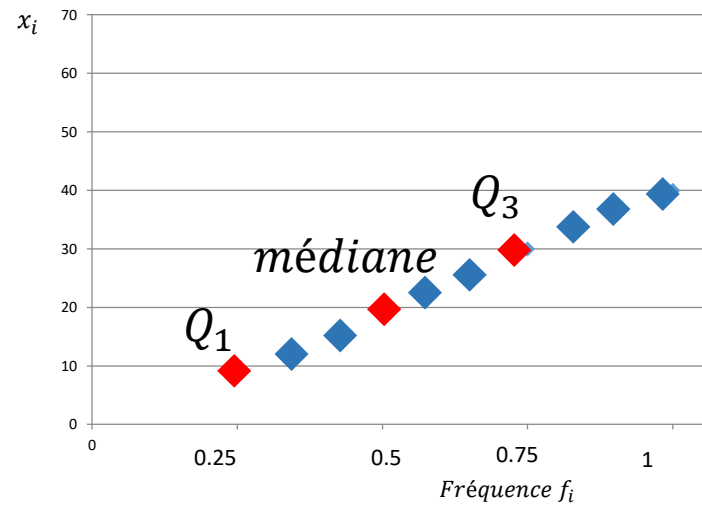
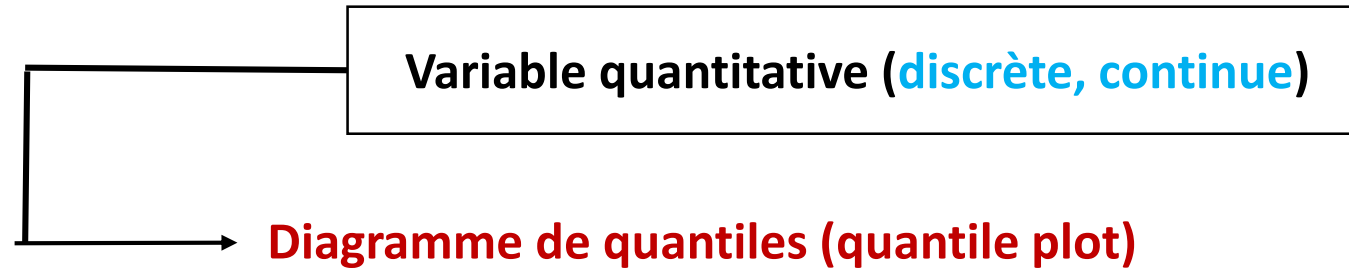
- 3) Dessiner l'histogramme à partir des effectifs corrigés



$N(.)$: fonction cumulative des effectifs

$F(.)$: fonction cumulative des fréquences

Visualisation des descriptions statistiques uni-variées

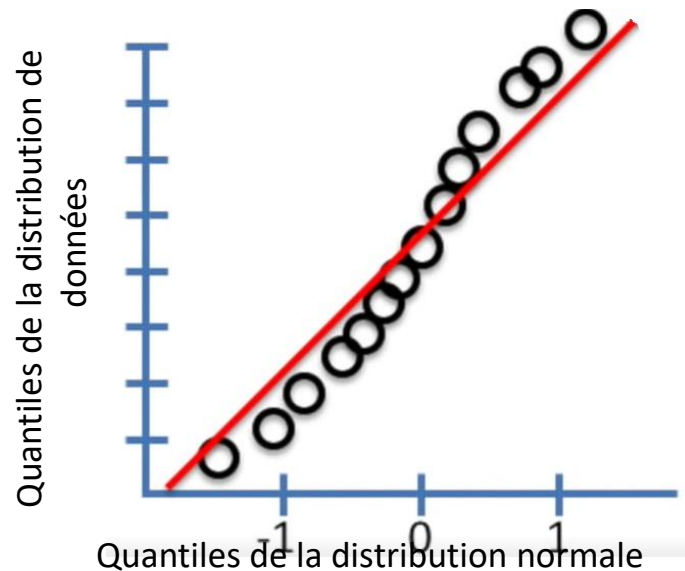


Visualisation des descriptions statistiques uni-variées

Variable quantitative (**discrète, continue**)

Diagramme de quantiles-quantiles (quantile plot)

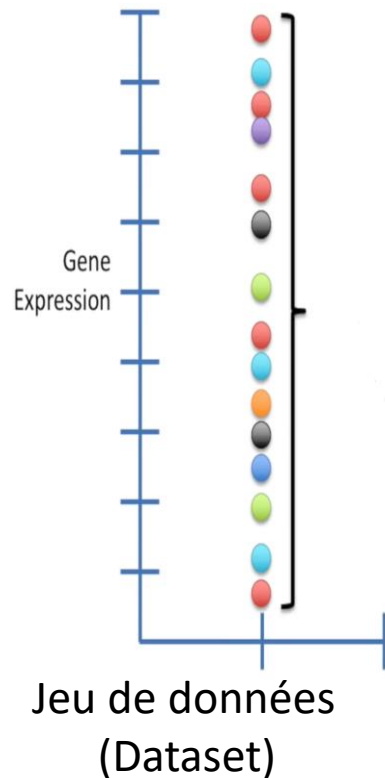
- Déterminer la distribution qui correspond aux donnés.
- Comparer deux jeux de données (distribution similaire/différente)



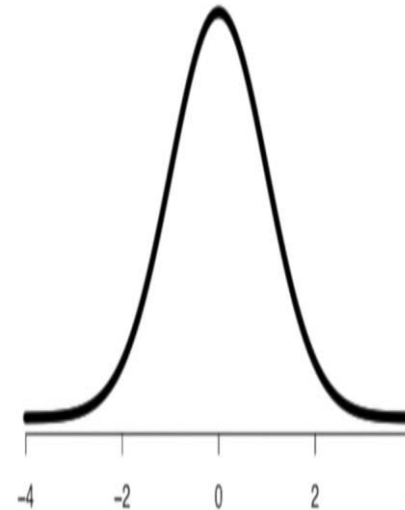
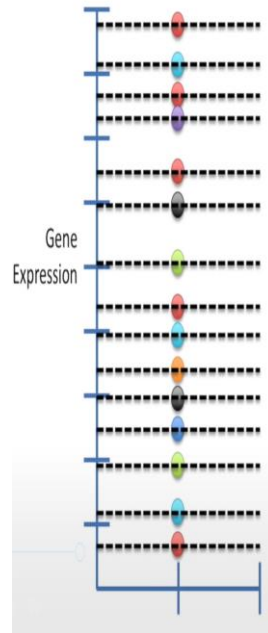
1. Calculer pour chaque point du dataset son quantile.
2. Calculer la courbe d'une distribution choisie
3. Ajouter les quantiles à la distribution choisie (même nombre que ceux du dataset)
4. Dessiner le graphe Q-Q (intersection des quantiles de du jeu de données et la distribution choisie)
5. Aligner les intersections de quantiles sur une droite

Visualisation des descriptions statistiques uni-variées

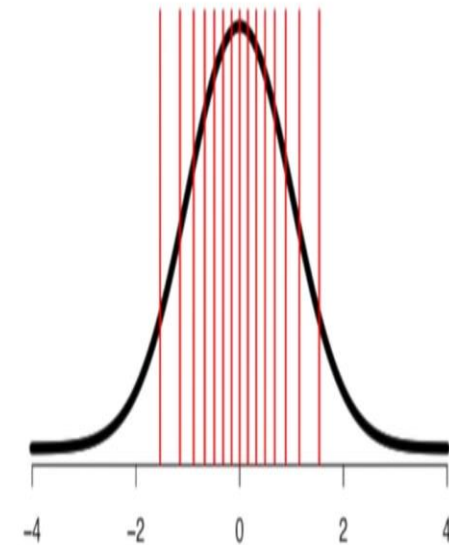
Exemple: Vérifier la distribution (normale) des 15 expressions génétiques.



(1) 15 Quantiles de taille égale (dataset)

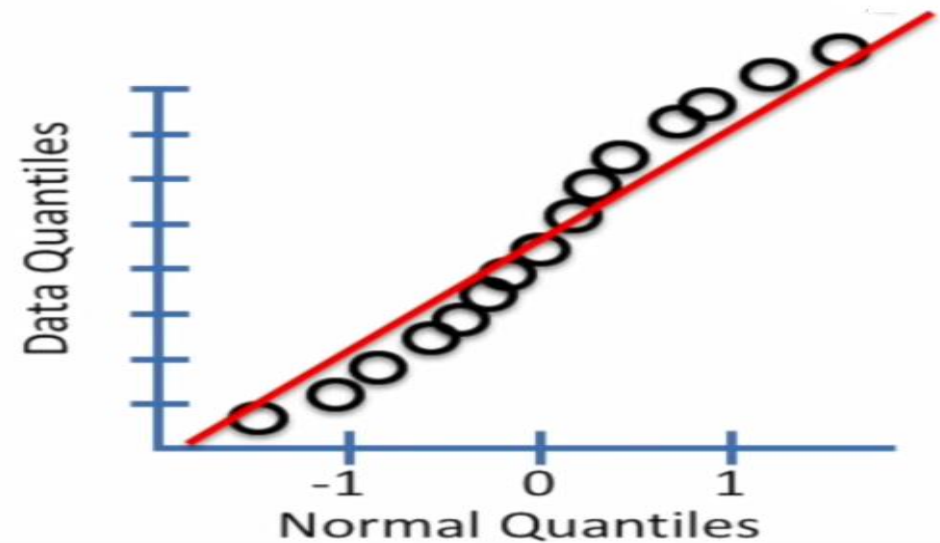
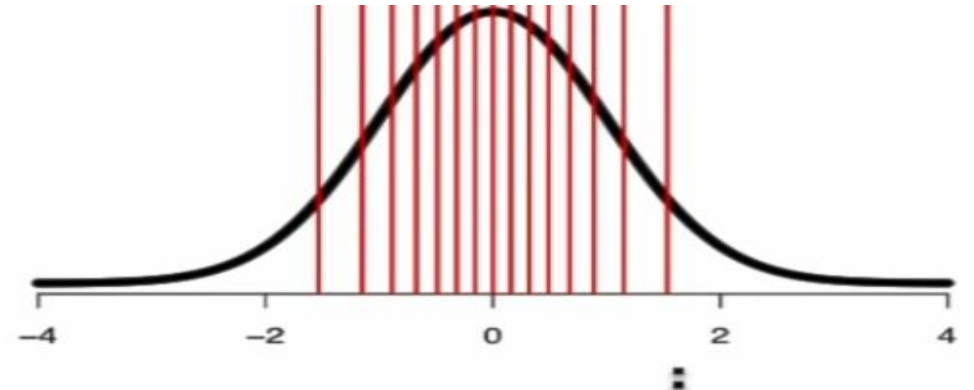
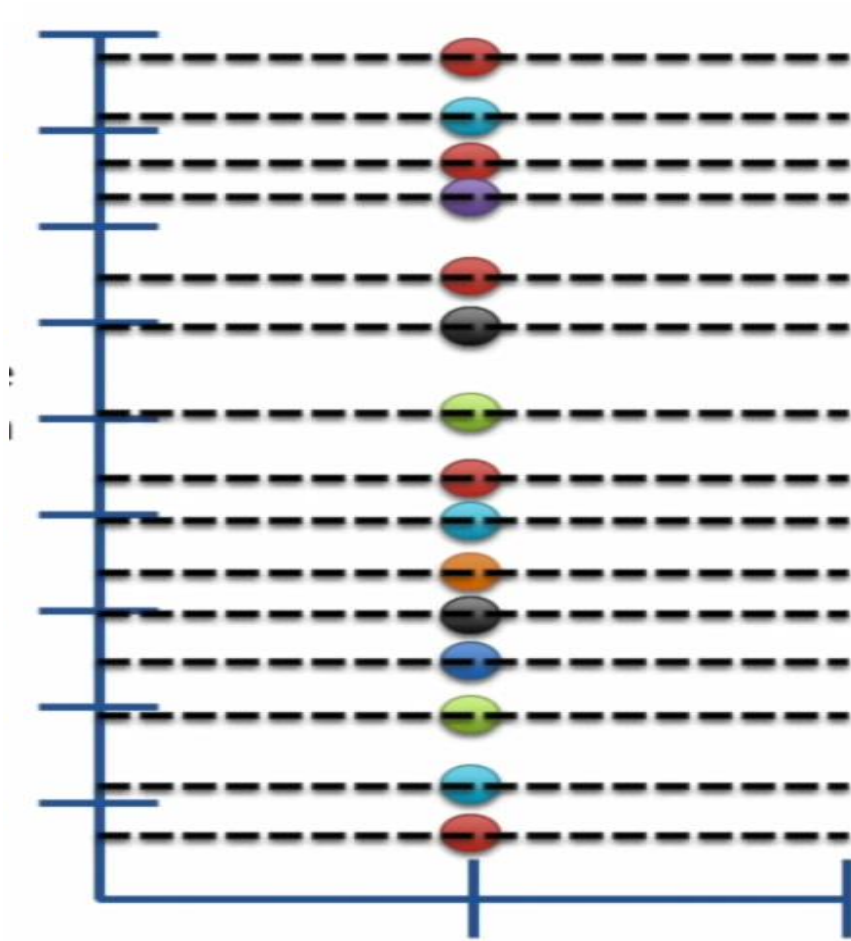


(2) Distribution normale

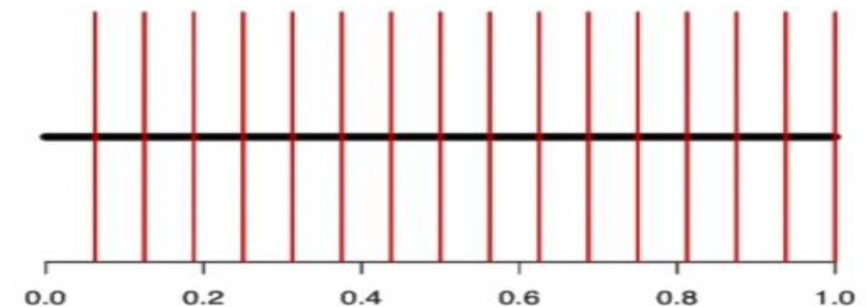
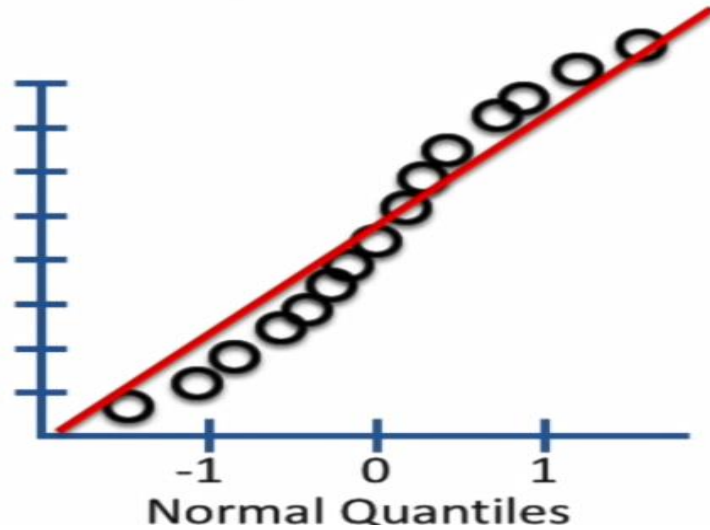
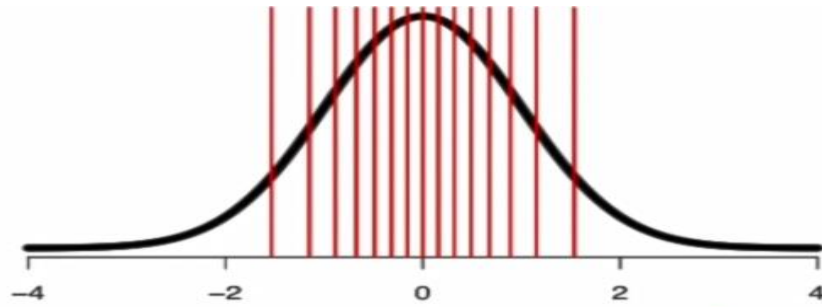


(3) 15 quantiles de taille égale (distribution normale)

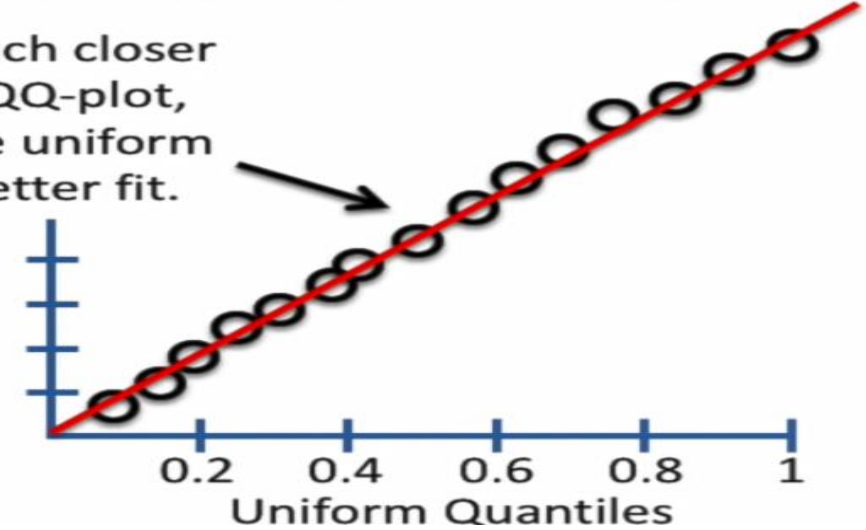
Visualisation des descriptions statistiques uni-variées



Visualisation des descriptions statistiques uni-variées

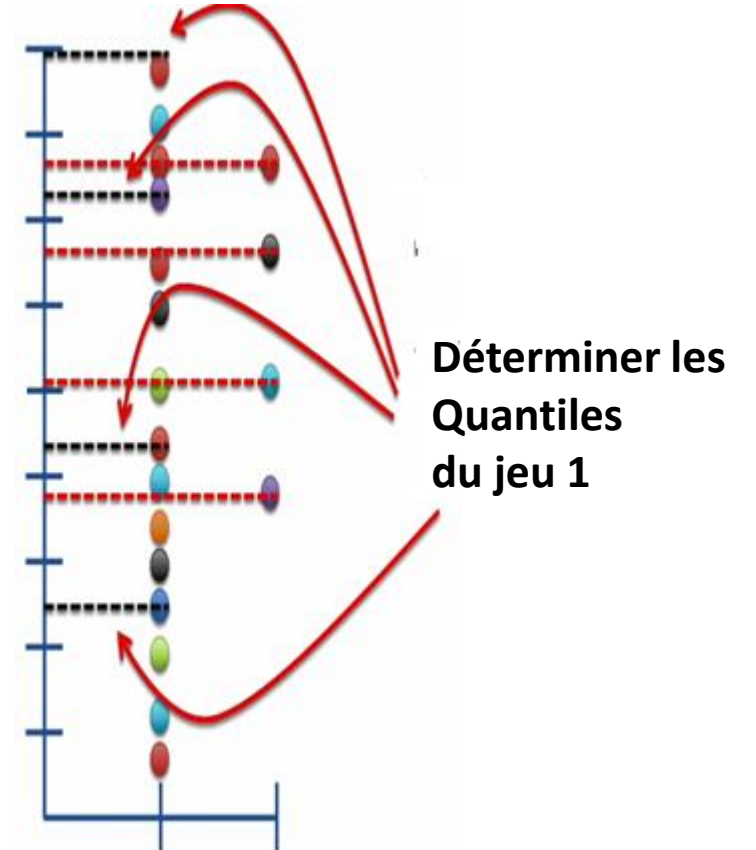
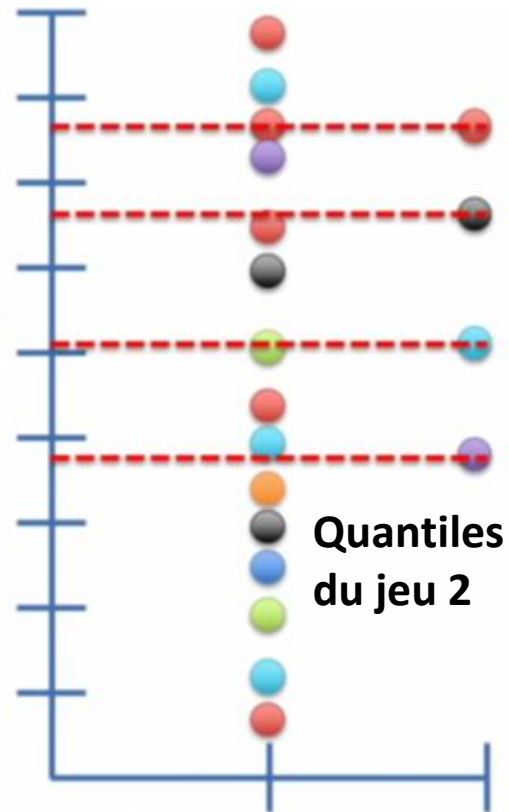


The points are much closer to the line in the QQ-plot, indicating that the uniform distribution is a better fit.

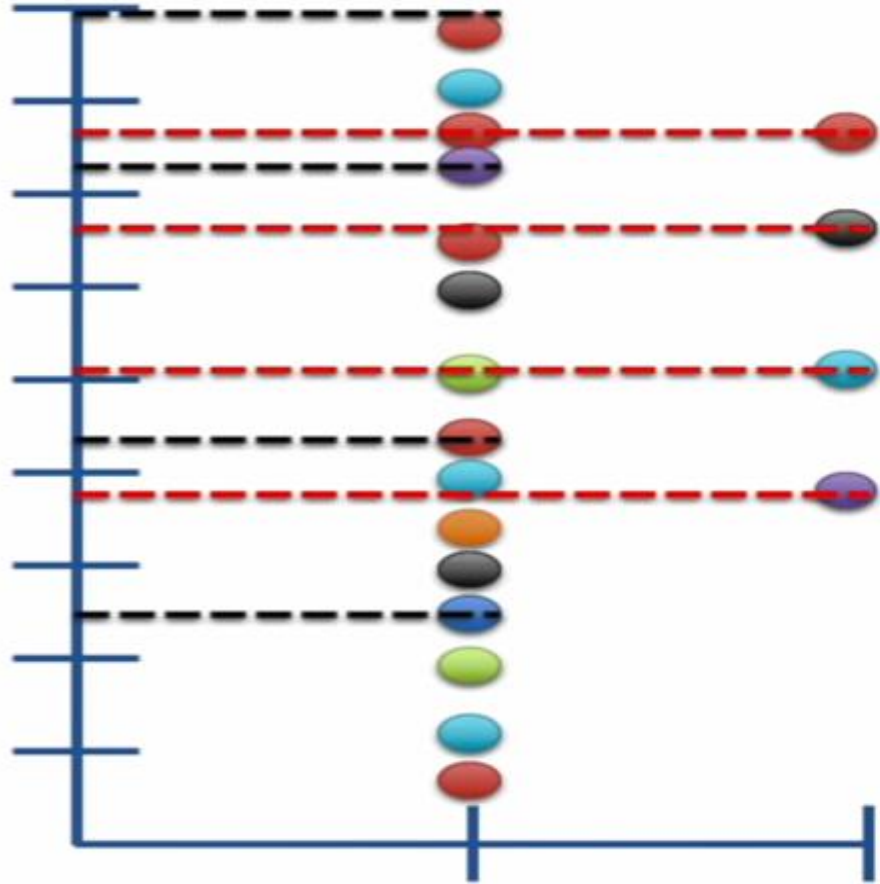


Visualisation des descriptions statistiques uni-variées

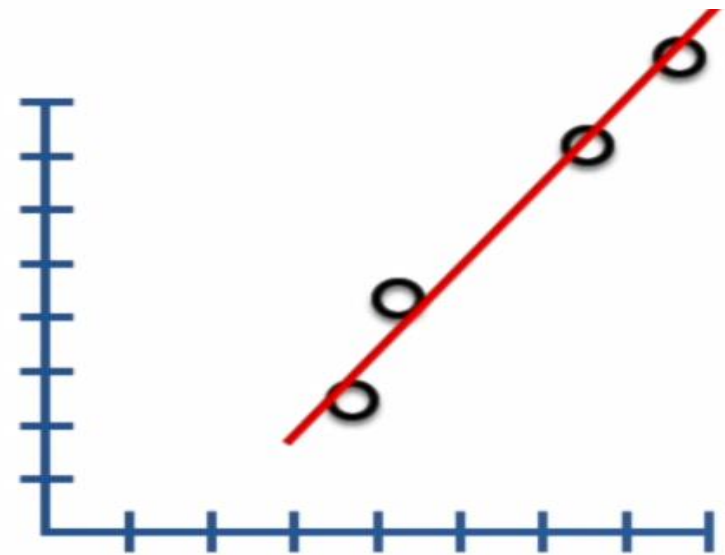
Exemple: Comparer le deux jeux de données.



Visualisation des descriptions statistiques uni-variées



Original Dataset Quantiles



New Dataset Quantiles