Centre universitaire Abdelhafid Boussouf -Mila
Institut de Sciences et Technologies
Département de mathématiques et informatique
$3^{\text{ième}}$ Année mathématiques

Matière: Introduction aux probabilités et statistique descriptive

Chapitre 2:Représentation numérique des données

Présentée par:
AZI Mourad

March 6, 2023

Plan de cours

Introduction

Plan de cours

Plan de cours

Introduction

Une série de données peut être résumée par quelques valeurs numériques appelées caractéristiques des séries statistiques, classées en quatre grandes catégories :

- les caractéristiques de tendance centrale,

Introduction

Une série de données peut être résumée par quelques valeurs numériques appelées caractéristiques des séries statistiques, classées en quatre grandes catégories :

- les caractéristiques de tendance centrale,
- les caractéristiques de dispersion,

Introduction

Une série de données peut être résumée par quelques valeurs numériques appelées caractéristiques des séries statistiques, classées en quatre grandes catégories :

- les caractéristiques de tendance centrale,
- les caractéristiques de dispersion,
- les caractéristiques de forme,

Introduction

Une série de données peut être résumée par quelques valeurs numériques appelées caractéristiques des séries statistiques, classées en quatre grandes catégories :

- les caractéristiques de tendance centrale,
- les caractéristiques de dispersion,
- les caractéristiques de forme,
- les caractéristiques de concentration.

Plan de cours

Introduction

Variable statistique discrète

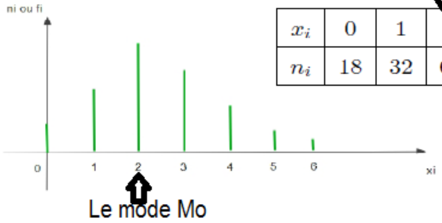
- Paramètres de tendance centrale ou de position
  - Le mode
  - Médiane et Quantiles
  - Moyenne
- Paramètres de dispersion
  - Étendu et écart interquartiles
  - écart-type, variance et coefficient de variation

Variable statistique continue

- Paramètres de tendance centrale ou de position
  - Le mode
  - La médiane et les quartiles
  - Moyenne
- Paramètres de dispersion
  - Étendu et écart interquartiles
  - Variance, écart-type et coefficient de variation

Le mode(ou valeur dominante)

Le mode d’une variable statistique est la modalité associée au plus grand effectif (ou la plus grande fréquence), on le note généralement par M_O.



Le mode Mo

x_i	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

Figure: Le mode

La médiane

La médiane

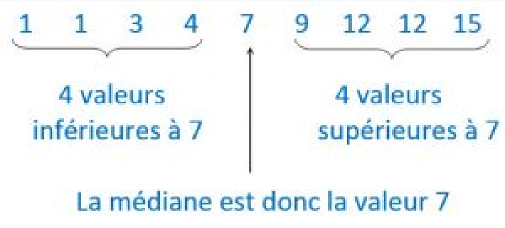
La médiane est la modalité qui partage la série statistique en deux parties égales, dont les éléments de la population étant rongés par ordre croissant ou décroissant. En d'autre terme, la médiane est la valeur statistique qui correspond sur la courbe cumulative, à une ordonnée représentant une fréquence cumulé égale 0,5 $(F_x(M_e) = 0.5)$.

La médiane

La médiane est la modalité qui partage la série statistique en deux parties égales, dont les éléments de la population étant rongés par ordre croissant ou décroissant. En d'autre terme, la médiane est la valeur statistique qui correspond sur la courbe cumulative, à une ordonnée représentant une fréquence cumulé égale 0,5 ($F_x(M_e) = 0.5$).

- Dans le cas où la série comporte un nombre impair n d'observations, la médiane M_e est égale à:

$$M_e = x_{\frac{n+1}{2}}.$$



La médiane

La médiane(n paire)

- Dans le cas où la série comporte un nombre *n* pair d'observations, la médiane est estimée par:

$$M_e = \frac{(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})}{2}.$$

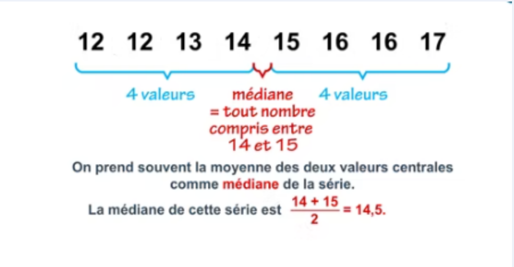


Figure: La mediane

Quantiles

Quantiles

Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.
Pour $k = 4$, les quantiles, appelés quartiles, sont trois nombres Q_1, Q_2, Q_3 tels que :

- Q_1 : c'est la modalité dont 25% des observations lui sont inférieures ou égale;

Quantiles

Quantiles

Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.
Pour $k = 4$, les quantiles, appelés quartiles, sont trois nombres Q_1, Q_2, Q_3 tels que :

- Q_1 : c'est la modalité dont 25% des observations lui sont inférieures ou égale;
- $Q_2 = M_2$: c'est la modalité dont 50% des observations lui sont inférieures ou égale;

Quantiles

Quantiles

Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.
Pour $k = 4$, les quantiles, appelés quartiles, sont trois nombres Q_1, Q_2, Q_3 tels que :

- Q_1 : c'est la modalité dont 25% des observations lui sont inférieures ou égale;
- $Q_2 = M_2$: c'est la modalité dont 50% des observations lui sont inférieures ou égale;
- Q_3 : c'est la modalité dont 75% des observations lui sont inférieures ou égale.

Quartiles

Quartiles(exemple 1)

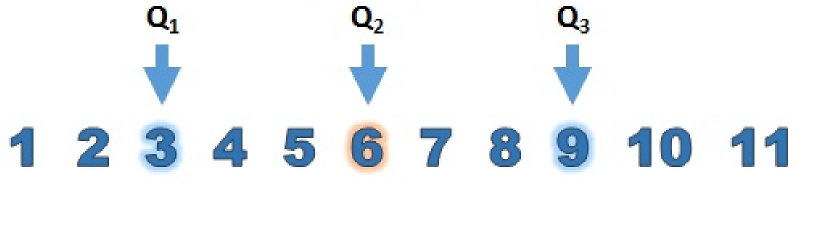


Figure: Quartiles

Quartiles

Quartiles(exemple 2)

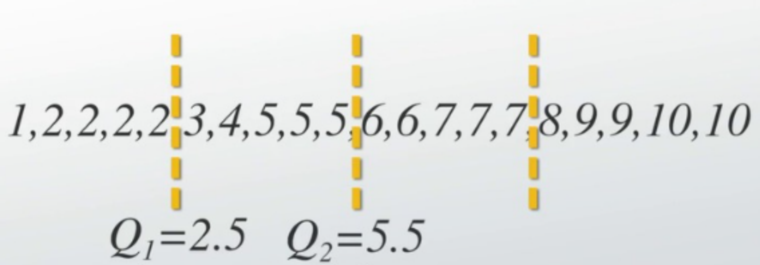


Figure: Quartiles

Quartiles

Diagramme en boîte à moustaches

Le diagramme en boîte à moustaches ou box-plot permet de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quartiles.

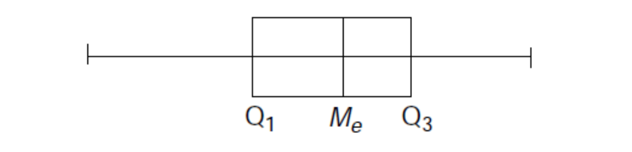


Figure: Le diagramme en boîte à moustaches

La moyenne arithmétique

La moyenne arithmétique, qu'on appelle tout simplement moyenne, est la somme de toutes les données statistiques divisée par le nombre de ces données(n). Elle est donnée par la relation:

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

D'une manière équivalente (série pondérée, effectif total N) elle peut s'écrire:

$$\bar{x} = \frac{\sum_{i=1}^{n} n_i x_i}{N} = \sum_{i=1}^{n} f_i x_i.$$

Plan de cours

Étendu et écart interquartiles

L'étendue est la différence entre la plus grande et la plus petite des valeurs observées :

$$E = x_{max} - x_{min}.$$

Étendu et écart interquartiles

L'étendue est la différence entre la plus grande et la plus petite des valeurs observées :

$$E = x_{max} - x_{min}.$$

L'écart interquartile est la différence entre Q_3 et Q_1:

$$EI = Q_3 - Q_1.$$

écart-type et variance

L'écart-type σ_X d'une variable statistique X est la mesure de dispersion la plus couramment utilisée. Il se définit comme la racine carrée de la variance, et la variance est la moyenne des carrés des écarts à la moyenne x:

$$Var(x) = \sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2.$$

Et

$$\sigma_x = \sqrt{Var(x)}.$$

L'écart-type caractérise la dispersion d'une série de valeurs:

- plus σ_x est petit, plus les données sont regroupées autour de la moyenne et on dit que la population est homogène;

écart-type, variance et coefficient de variation

écart-type et variance

L'écart-type σ_X d'une variable statistique X est la mesure de dispersion la plus couramment utilisée. Il se définit comme la racine carrée de la variance, et la variance est la moyenne des carrés des écarts à la moyenne x:

$$Var(x) = \sum_{i=1}^{n} f_i (x_i - \bar{x})^2 = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2.$$

Et

$$\sigma_x = \sqrt{Var(x)}.$$

L'écart-type caractérise la dispersion d'une série de valeurs:

- plus σ_x est petit, plus les données sont regroupées autour de la moyenne et on dit que la population est homogène;
- plus il est grand, plus les donnée sont dispersés autour de la moyenne (on dit que la série est hétérogène).

écart-type, variance et coefficient de variation

écart-type et variance

L'écart-type σ_X d'une variable statistique X est la mesure de dispersion la plus couramment utilisée. Il se définit comme la racine carrée de la variance, et la variance est la moyenne des carrés des écarts à la moyenne x:

$$Var(x) = \sum_{i=1}^{n} f_i (x_i - \bar{x})^2 = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2.$$

Et

$$\sigma_x = \sqrt{Var(x)}.$$

L'écart-type caractérise la dispersion d'une série de valeurs:

- plus σ_x est petit, plus les données sont regroupées autour de la moyenne et on dit que la population est homogène;
- plus il est grand, plus les donnée sont dispersés autour de la moyenne (on dit que la série est hétérogène).

écart-type et variance

L'écart-type σ_X d'une variable statistique X est la mesure de dispersion la plus couramment utilisée. Il se définit comme la racine carrée de la variance, et la variance est la moyenne des carrés des écarts à la moyenne x:

$$Var(x) = \sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2.$$

Et

$$\sigma_x = \sqrt{Var(x)}.$$

L'écart-type caractérise la dispersion d'une série de valeurs:

- plus σ_x est petit, plus les données sont regroupées autour de la moyenne et on dit que la population est homogène;
- plus il est grand, plus les donnée sont dispersés autour de la moyenne (on dit que la série est hétérogène).

Cependant avant de conclure, il faut faire attention à l'ordre de grandeur des données.

écart-type, variance et coefficient de variation

coefficient de variation

Le coefficient de variation est donnée par la relation:

$$CV = 100\frac{\sigma_x}{\overline{x}}\%.$$

Il permet d'apprécier l'homogénéité de la distribution, une valeur du coefficient de variation inférieure à 15% traduit une bonne homogénéité de la distribution.

Plan de cours

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$Mo = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$

avec:

- L_{i-1}: la borne inférieure de la classe modale;

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$\mathrm{Mo} = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$

avec:

- L_{i-1}: la borne inférieure de la classe modale;
- a_i : le pas de la classe modale;

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$Mo = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$

avec:

- L_{i-1}: la borne inférieure de la classe modale;
- a_i : le pas de la classe modale;
- $\Delta_1 = n_i - n_{i-1}, \Delta_2 = n_i - n_{i+1}$ ou $(\Delta_1 = f_i - f_{i-1},\ \ \Delta_2 = f_i - f_{i-1})$;

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$Mo = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$

avec:

- L_{i-1}: la borne inférieure de la classe modale;
- a_i : le pas de la classe modale;
- $\Delta_1 = n_i - n_{i-1}, \Delta_2 = n_i - n_{i+1}$ ou $(\Delta_1 = f_i - f_{i-1},\ \ \Delta_2 = f_i - f_{i-1})$;
- n_i et f_i sont l'effectif et la fréquence associés à la classe modale;

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$Mo = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$

avec:

- L_{i-1}: la borne inférieure de la classe modale;
- a_i : le pas de la classe modale;
- $\Delta_1 = n_i - n_{i-1}, \Delta_2 = n_i - n_{i+1}$ ou $(\Delta_1 = f_i - f_{i-1},\ \ \Delta_2 = f_i - f_{i-1})$;
- n_i et f_i sont l'effectif et la fréquence associés à la classe modale;
- n_{i-1} est l'effectif de la classe qui précède la classe modale;

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$\text{Mo} = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$

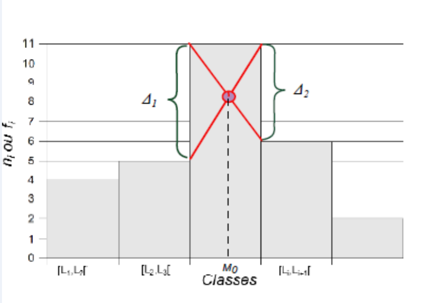
avec:

- L_{i-1}: la borne inférieure de la classe modale;
- a_i : le pas de la classe modale;
- $\Delta_1 = n_i - n_{i-1}, \Delta_2 = n_i - n_{i+1}$ ou $(\Delta_1 = f_i - f_{i-1},\ \ \Delta_2 = f_i - f_{i-1})$;
- n_i et f_i sont l'effectif et la fréquence associés à la classe modale;
- n_{i-1} est l'effectif de la classe qui précède la classe modale;
- n_{i+1} est l'effectif de la classe qui suit la classe modale.

Le mode(ou valeur dominante)

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif (ou la plus grande fréquence partielle). La Le mode est donné par la relation suivante:

$$Mo = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} = L_{i-1} + a_i \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}$$



La médiane

C'est la valeur M_e telle que $F(M_e) = 0.5$.

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{0.5 - F(L_{i-1})}{M_e - L_{i-1}}.$$

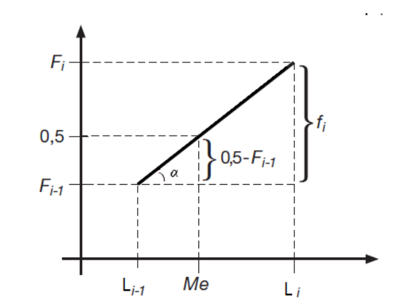
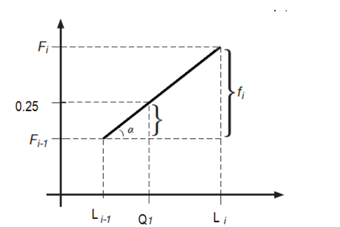


Figure: médiane

Les quartiles

- Q_1 : c'est la modalité dont 25% des observations lui sont inférieures ou égale $(F(Q_1) = 0.25)$:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{0.25 - F(L_{i-1})}{Q_1 - L_{i-1}}.$$



Les quartiles

- Q_3 : c'est la modalité dont $F(Q_3) = 0.75$:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{0.75 - F(L_{i-1})}{Q_3 - L_{i-1}}.$$

Les quartiles

- Q_3 : c'est la modalité dont $F(Q_3) = 0.75$:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{0.75 - F(L_{i-1})}{Q_3 - L_{i-1}}.$$

- Pour la calcule de F(x) on utilise la relation:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{F(x) - F(L_{i-1})}{x - L_{i-1}}.$$

Les quartiles

- Q_3 : c'est la modalité dont $F(Q_3) = 0.75$:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{0.75 - F(L_{i-1})}{Q_3 - L_{i-1}}.$$

- Pour la calcule de F(x) on utilise la relation:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{F(x) - F(L_{i-1})}{x - L_{i-1}}.$$

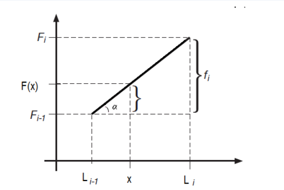
Les quartiles

- Q_3 : c'est la modalité dont $F(Q_3) = 0.75$:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{0.75 - F(L_{i-1})}{Q_3 - L_{i-1}}.$$

- Pour la calcule de F(x) on utilise la relation:

$$\tan(\alpha) = \frac{F(L_i) - F(L_{i-1})}{L_i - L_{i-1}} = \frac{F(x) - F(L_{i-1})}{x - L_{i-1}}.$$



La moyenne arithmétique

La moyenne arithmétique dans le cas d'une variable continue s'écrit:

$$\bar{x} = \frac{\sum_{i=1}^{n} n_i c_i}{N} = \sum_{i=1}^{n} f_i C_i.$$

où: C_i est le centre de la classe i.

Plan de cours

Étendu et écart interquartiles

L'étendue est la différence entre la plus grande et la plus petite des valeurs observées :

$$E = x_{max} - x_{min}.$$

Étendu et écart interquartiles

L'étendue est la différence entre la plus grande et la plus petite des valeurs observées :

$$E = x_{max} - x_{min}.$$

L'écart interquartile est la différence entre Q_3 et Q_1:

$$EI = Q_3 - Q_1.$$

Variance, écart-type et coefficient de variation

la variance est la moyenne des carrés des écarts à la moyenne x:

$$Var(x) = \sum_{i=1}^{n} f_i (C_i - \bar{x})^2 = \sum_{i=1}^{n} f_i C_i^2 - \bar{x}^2.$$

Variance, écart-type et coefficient de variation

la variance est la moyenne des carrés des écarts à la moyenne x:

$$Var(x) = \sum_{i=1}^{n} f_i(C_i - \overline{x})^2 = \sum_{i=1}^{n} f_i C_i^2 - \overline{x}^2.$$

L'écart-type est:

$$\sigma_x = \sqrt{Var(x)}.$$

Le coefficient de variation est donnée par la relation:

$$CV = 100\frac{\sigma_x}{\overline{x}}\%.$$