Cours d'Eléments de Statistique

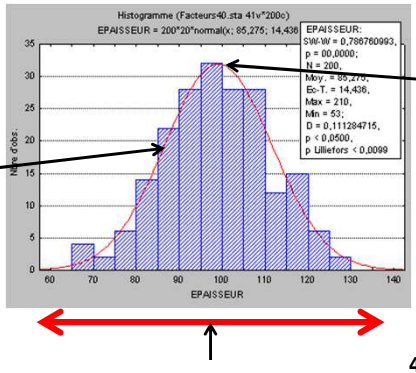
Jean Christophe Meunier

Module 3
Étude des séries statistiques : Caractérisation des données

1ère Bac, Commerce Extérieur
Année académique 2015-2016

EPFC

Indicateurs de la courbe de distribution



3. Forme :
Asymétrie et aplatissement

1. Position :
Valeur qui occupe une position précise (ex. tendance centrale) ou qui découpe l'échantillon selon une proportion spécifiée

2. Dispersion :
Dans l'absolu ou par rapport à la tendance centrale

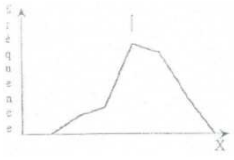
4. Concentration :
Ressource/argent : répartition des ressources au sein de population

2

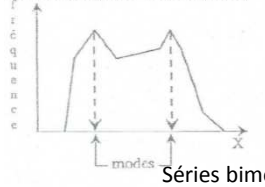
A. Indicateurs de position

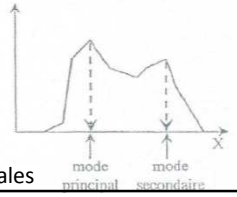
I. Tendance centrale

1. Mode (x_0)
   - Valeur ou classe (x_i) de la série statistique dont l'effectif (n_i) est le plus élevé
   - Déterminé via table/graphe des effectifs
     - Repérer x_i dont n_i est le plus élevé
     - Données groupées : on parle de 'classe modale'
   - Un vs. Plusieurs modes
     - Unimodale : un seul 'pic' d'effectifs
     - Bimodale : deux 'pics' d'effectifs
       - Peut-être un indice que 2 populations ≠ sont considérées sur le même graphe (ex. taille homme et femme)



Série unimodale





Séries bimodales

4

I. Tendance centrale

2. Médiane ($\tilde{x}$)

- Valeur ou classe (x_i) qui 'coupe' l'échantillon en deux parties égales
  - Les effectifs des valeurs > et < à la valeur médiane sont égaux (à N/2 : 50%< et 50%>)
  - Si N est impair (2p + 1)
    - Une seule valeur se situe exactement à la moitié de l'échantillon : Médiane = $(p+1)^{ème}$ valeur
    - Ex série statistique impaire (N=9) :
      - » 1 1 2 2 **2** 3 4 5 5 → Médiane = 2 ($(p+1)^{ème}$ valeur))
  - Si N est pair (2p)
    - Deux valeurs se situent 'à cheval' sur la moitié de l'échantillon : Médiane = moyenne de $p^{ème}$ valeur et de $(p+1)^{ème}$ valeur
    - Ex série statistique paire (N=10) :
      - » 1 1 2 2 **2 3** 3 4 5 5 → Médiane = (2+3)/2 = 2,5

5

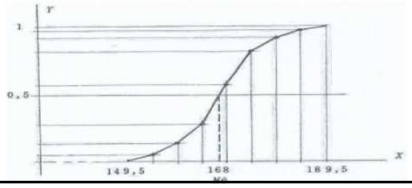
I. Tendance centrale

2. Médiane ($\tilde{x}$)

- Comment retrouver la médiane :
  - Par série statistique brute et ordonnée (cf. supra)
    - Ex série statistique impaire (N=9) : 1 1 2 2 **2** 3 4 5 5 → Médiane = 2 ($(p+1)^{ème}$ valeur))
  - Par table des effectifs cumulés
    - Repérer valeur ou classe (x_i) qui comprend la $(p+1)^{ème}$ valeur (N impair) ou la $p^{ème}$ et $(p+1)^{ème}$ valeur (N pair)

Réponses x_i de la variable x	45	55	60	75	80	85	90
Effectifs n_i	1	2	3	5	2	1	1
Effectifs cumulés n_i	1	3	6	11	13	14	15

- Par graphe des effectifs cumulés



6

I. Tendance centrale

3. Moyenne (μ ou $\bar{x}$) *

- Somme de toutes les observations divisée par nombre d'observations
- Soit,

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

Via données brutes — Via table effectif

Si données groupées par classes, le centre de classe peut être considéré comme estimation de x_i - Sous l'h° d'équirépartition au sein des classes

- Ex série (N=13): 1 1 1 1 1 2 2 2 2 3 3 3 3

$$\bar{x} = \frac{1+1+1+1+\cdots+3+3+3}{13}$$ Via données brutes

$$\bar{x} = \frac{(5*1)+(4*2)+(4*3)}{13}$$ Via table effectif

* $\bar{x}$ quand échantillon ; μ quand population

7

I. Tendance centrale

3. Moyenne (μ ou $\bar{x}$) : propriétés

- Uniquement pour variables quantitatives
- Unique : une seule moyenne pour toute série statistique
- Somme des écarts entre x_i et la moyenne est nulle
  - Les différences positives et négatives s'annulent

$$\sum_{i=1}^{i=n}(x_i - \bar{x}) = 0$$

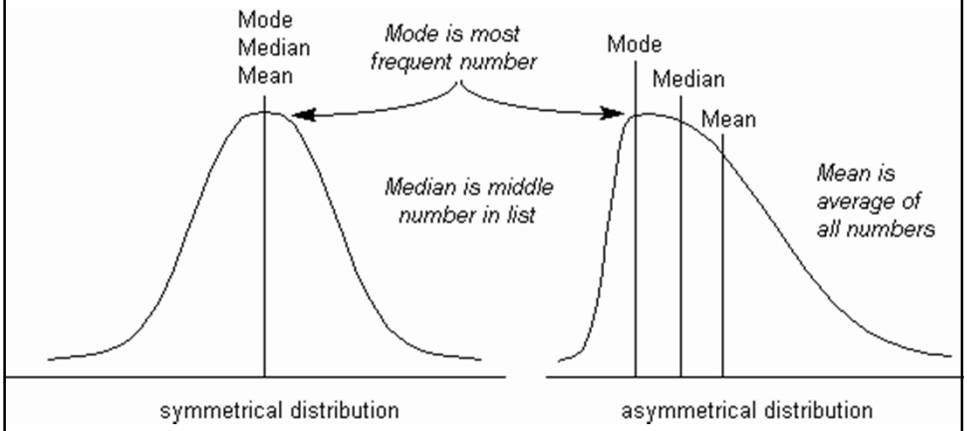
- Moyenne de deux séries statistiques (pour une même variable)
  - Ex moyenne taille homme (série a) et taille femme (série b)

$$\bar{x} = \frac{\sum_{i=0}^{i=n} x_i(série\ a) + \sum_{i=0}^{i=n} x_i(série\ b)}{N_a + N_b} = \frac{\sum_{i=1}^{i=c} n_i x_i(série\ a) + \sum_{i=1}^{i=c} n_i x_i(série\ b)}{N_a + N_b}$$

8

I. Tendance centrale

- Mode, médiane & moyenne



9

II. Autres indicateurs de position

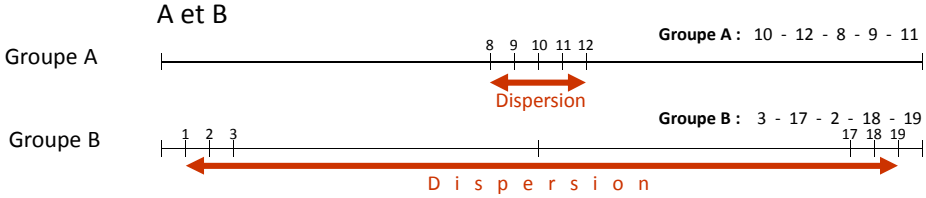
1. Percentile
   - Valeur des observations qui 'découpe' l'échantillon selon une proportion précise exprimé en pourcentage
     - Ex : percentile 12 → 12% de l'échantillon inférieur à cette valeur et 88% supérieur
2. Décile
   - Idem mais échantillon segmenté par tranche de 10%
     - Ex : décile 3 → 30% inférieur à cette valeur et 70% supérieur
3. Quartile
   - Idem mais échantillon segmenté par tranche de 25%
     - Q_1 → 25% de l'échantillon inférieur à cette valeur et 75% supérieur
     - Q_2 (= médiane) → 50% inférieur et 50% supérieur
     - Q_3 → 75% de l'échantillon inférieur à cette valeur et 25% supérieur

10

B. Indicateurs de dispersion

I. Variance et écart-type

- Compléments aux indices de position
  - Indices de position ne disent rien sur la dispersion
    - Ex : pour une même moyenne (10), scores sur 20 de deux groupes A et B



  - Notion de 'Moment' : écart moyen des x_i à la moyenne

$$m = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - \bar{x})$$

- Donne une indication de la dispersion des valeurs autour de la moyenne mais 'pas utilisable' comme tel
  - Les différences positives et négatives s'annulent
  - Valeur absolue des différences ou les élever à la puissance 2 → moment d'ordre 2 = variance (cf. dia suivante)

12

I. Variance et écart-type

1. Variance (σ^2 ou s^2)*
   - Moyenne des carrés des écarts des valeurs x_i à la moyenne

Ecarts à la moyenne élevés au carré

$$\sigma^2 = \frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \left(X_3 - \overline{X}\right)^2 + \ldots\ldots\ldots\ldots \left(X_n - \overline{X}\right)^2}{n}$$

   - ou, plus simplement

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{i=c} n_i(x_i - \bar{x})^2$$

Via données brutes — Via table effectif

Si données groupées par classes, le centre de classe peut être considéré comme estimation de x_i
- Sous l'hypothèse d'équirépartition au sein des classes

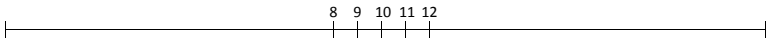
13

* s^2 quand échantillon ; σ^2 quand population

I. Variance et écart-type

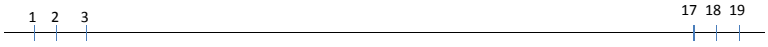
1. Variance (σ^2 ou s^2)

Groupe A



$$Variance = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} = \frac{10}{5} = 2$$

Groupe B



$$Variance = \frac{(1-10)^2 + (2-10)^2 + (3-10)^2 + (17-10)^2 + (18-10)^2 + (19-10)^2}{6} = \frac{388}{6} = 64{,}66$$

14

I. Variance et écart-type

2. Ecart-type (σ ou s)*
   - Racine carrée de la variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{i=n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{i=c} n_i(x_i - \bar{x})^2}$$

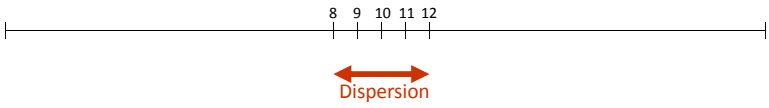
   - Indice similaire à la variance mais plus facilement interprétable
     - moyenne des écarts et non moyenne des carrés des écarts
     - La valeur d'écart-type peut s'exprimer selon la même métrique que la série statistique dont il est issu

* *s* quand échantillon ; σ quand population

15

I. Variance et écart-type

2. Ecart-type (σ ou s)



Groupe A : Variance (σ^2)= 2 Ecart type (σ) = $\sqrt{2}$=1,41

Groupe B : Variance (σ^2)= 64,66 Ecart type (σ) =$\sqrt{64,66} = 8,04$

16

II. Autres indicateurs de dispersion

1. Ecart absolu moyen par rapport à la médiane
   - 1. Valeur absolue des écarts à la médiane
   - 2. Somme de toute ces valeurs
   - 3. Divisé par n, nombre d'observations
2. Ecart absolu moyen par rapport à la moyenne
   - Idem mais par rapport à la moyenne
3. Etendue
   - Différence entre la plus grande et la plus petite valeur
4. Espace inter-quartile (Q_1 - Q_3)
   - Intervalle (entre Q_1 et Q_3) qui contient les 50% de la population se situant au centre de la distribution

17

C. Indicateurs de forme

I. Asymétrie

- Coefficient d'asymétrie : Skewness
  - Obtenu à partir du moment centré d'ordre 3

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

  - Formule Skweness (γ_1)

$$\gamma_1 = \frac{m_3}{s^3}$$

0 si parfaitement symétrique
- si asymétrique à droite
+ si asymétrique à gauche

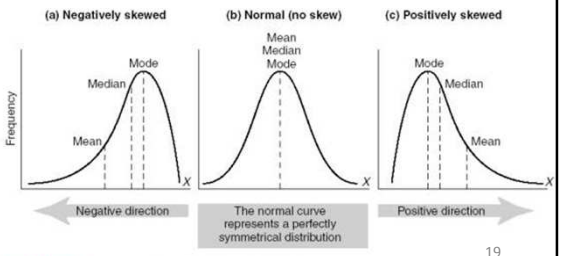


FIGURE 15.6 Examples of normal and skewed distributions

19

II. Aplatissement

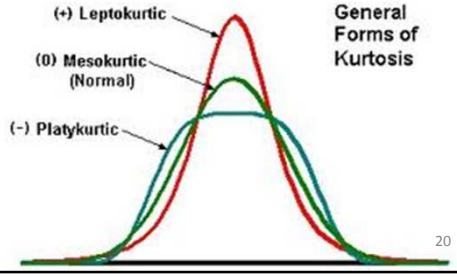
- Coefficient d'aplatissement : Kurtosis
  - Obtenu à partir du moment centré d'ordre 4

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

  - Formule Kurtosis (γ_2)

$$\gamma_2 = \frac{m_4}{s^4} - 3$$

0 si suit parfaitement la loi normale
- si plus aplatis
+ si plus pointus

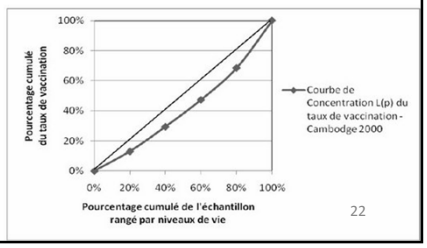


20

D. Indicateurs de concentration

I. Courbe de Lorenz

- Variables
  - Ressources ou unités monétaires
  - Permet de voir la répartition équitable ou non des ressources au sein de l'échantillon
- Sur le graphe
  - Axe des X :
    - Fréquence cumulée : %^{age} cumulé de l'échantillon/des effectifs
  - Axe des Y :
    - %^{age} cumulé - par valeur de X – de l'enveloppe totale



22

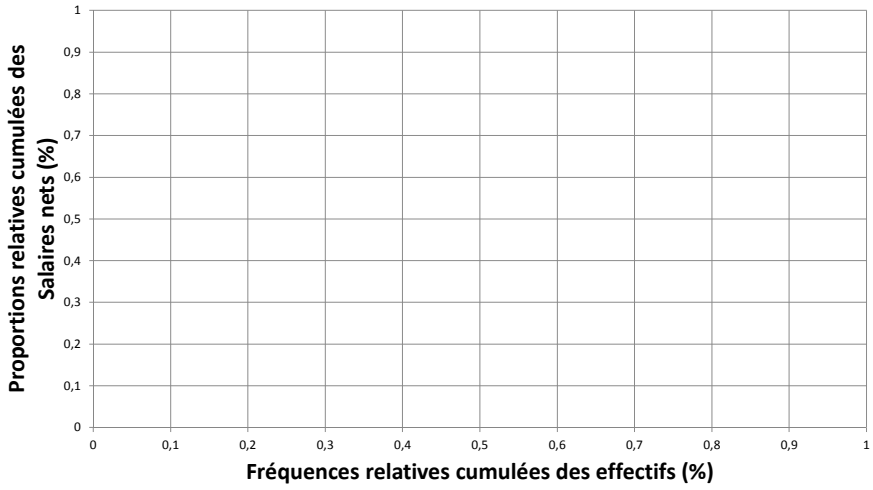
I. Courbe de Lorenz

- Questions : « Comment le montant total se répartit au sein de l'échantillon ? »
- Calcul

Réponses x_i de la variable	30	40	50	60	70	80	90	100
Effectifs (n_i)	2	5	4	9	12	4	5	1
Effectifs cumulés	2	7	11	20	32	36	41	42
Fréquences relatives (%)	4.7%	11.9%	9.5%	21.4%	28.6%	9.5%	11.9%	2.4%
Fréquences relatives cumulées (%) --> Axe X de la courbe	4.7%	16.6%	26.2%	47.6%	76.2%	85.7%	97.6%	100%
Valeurs globales au sein de l'effectif (x_i X n_i)	60	200	200	540	840	320	450	100
Valeurs globales cumulées	60	260	460	1000	1840	2160	2610	2710
Proportions relatives cumulées (%) --> Axe Y de la courbe	2.2%	9.6%	17.0%	36.9%	67.9%	79.7%	96.3%	100%

23

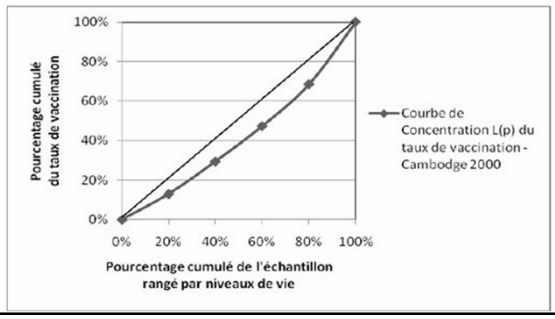
Indices de concentration : courbe de Lorenz



24

II. Indicateur de Gini

- Indicateur de Gini = (aire de concentration)/(aire sous la diagonale)
  - Si IG = 1 → répartition égale (équirépartition)
  - Si IG < 1 → répartition inégale (plus la valeur dévie de 1, plus inégalité est grande)



25