statistique descriptive

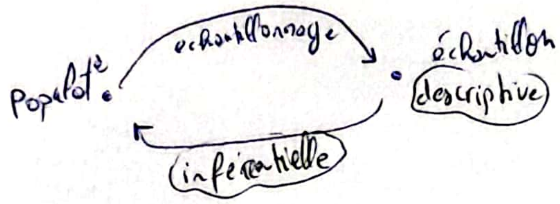
. statistiques: données numériques
. statistique: méthode d'analyse des données numérique.

(statistique descriptive)

. décrit les caractéristiques d'un échantillon

(statistique inférentielle)

"généraliser à partir d'un échantillon"



. Etude statistique (steps):
① problématique → ② collecte des données → ③ préparat° données
④ Analyse statistique (déductive ; inductive) → ⑤ product° résultat
"graphe" "généraliser"

collecte de Données:

populat° <
- recensement (tous) "enquête exhaustive"
- échantillonnage (no tous) "enquête partielle (sondage)"

méthode d'échantillonnage:

. Echantillonnage →
- aléatoire simple (avec / sans remise)
  "sans remise est plus fréquent et plus précis"
- stratifié ou par groupes
  . pour populat° hétérogène.
- ~~à 2 ou plus~~ niveaux
  . tirage au sort des familles
- méthode des quotas (quota)
  . sondage d'opinion

Précision de l'enquête
dépend ↗ <
- taille de l'échantillon (valeur, %) ↗
- degré d'homogénéité (variabilité)

définit° des observat°s
- modalité — non-numeric (° qualitative (catégorielle) échelle <
  - Nominale (catégorie) ← possibilité
  - ordinale (ordre de catégorie)
- possibilité ~ (mesures / valeurs — numeric (° quantitative <
  - continue (intervalle)
  - discrète (entier)
  } 'Rang

①

Chap 1 Représentation graphique des données :

Chap 2 Caractéristiques de position, de dispersion et de forme :

→ variables
- qualitatives
- quantitatives discrètes
- quantitatives continues.

- graphe !
- qualitative
  - nominale → sex / eye color
  - ordinale → score / satisfaction
  - → - Diagramme sectoriel - Bâton
  - → Mode (IQR variance)
  - → Mode + Médiane, Rang
- quantitative (Echelle)
  - Discrète → Nbre d'enfant → · diagramme en bâton · cumulatif
  - continue → Revenu, âge, taille → · Histogramme
  - → Mode, Médiane, moyenne, Rang, Ecart-type (IQR + variance)

effectifs cumulés

fréquences cumulées

- indicateurs !
- caractéristiques de :
  - tendances centrales : moyenne arithmétique, médiane, mode
  - Dispersion : Etendu, Ecart-type, indice de variabilité
  - Forme : quantiles, indice symétrie, indice aplatissement

N : # observations totale = "effectif cumulé"

- médiane (cas) valeur de variable
  - valeurs distincts (N valeurs)
    - N impair : $Q_2 = x_{(k)}$ tq $k = \frac{(N+1)}{2}$
    - N pair : $Q_2 = \frac{x_{(k)} + x_{(k+1)}}{2}$ tq $k = N/2$
  - valeurs répétées
    - effectifs cumulés croissants
    - $\frac{\max(N_i)}{2}$ f
    - Q_2 correspond
  - variable continue
    - ① · go classe médiane
    - ② · $Q_2 = B_{inf} + \left[\left(\frac{N}{2} - F\right) / f\right] \cdot E$

- quartiles Q_1 / Q_3 (1er 3ème quartile)
  Q_1 25% ; Q_3 75%

→ Q_1 … $N_i < N/4$ … $N_i < \frac{3N}{4}$ … Q_3

→ rang $Q_1 = \frac{N}{4}$ → rang $Q_3 = \frac{3N}{4}$

Mode : valeur plus fréquente

moyenne arithm.

discret (valeur) : $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

continue (centre de classe) : $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} n_i x_i = \sum f_i x_i$

moyenne géométrique : taux d'accroissement annuel ; **harmonique** : moyenne de % ; **quadratique** : moyenne d'écart

Etendue : différence (max − min)

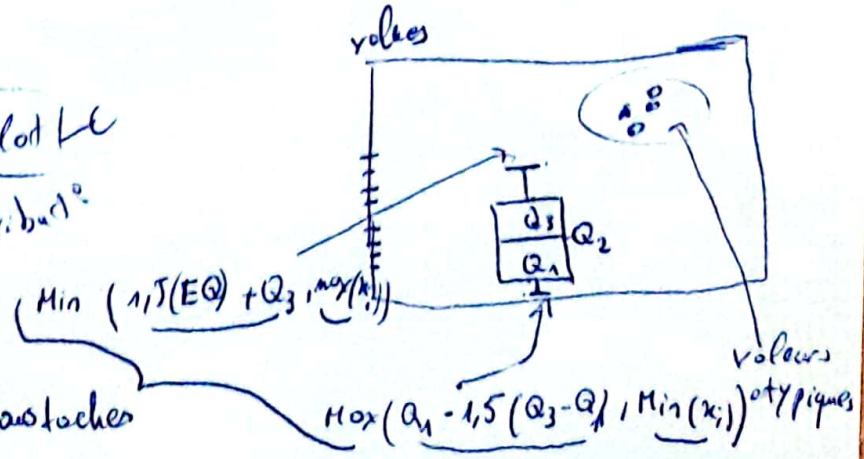
Variance : $s^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{1}^{n} f_i (x_i - \bar{x})^2$ (écart moyen sous-carré comme normal)

x_i : valeur ou centre de classe

Ecart interquartile : $EQ = Q_3 - Q_1 \equiv IQR$ (Range) $\equiv Iq$

(pour distributions dissymétriques)

boîte à moustache / **box-plot** : représentation graphique d'une distribution



Etendue et Variance / Ecart-type sont des mesures de dispersion absolue

* coefficient de variat° ≡ indice de variabilité
" dispersion relative pour comparer 2 distribut°s "

$$CV = \frac{S}{\bar{X}}(100) \equiv \frac{\text{écart-type}}{\text{moyenne}} \times 100$$

→ dispersion relative plus grande que celle de la distribut°
≡ ↗ ⇒ moins homogène ..
→

Mesures de forme :

* coefficient d'asymétrie (SKWENESS)

$$AS = \frac{m_3}{s^3} \; ; \; m_3 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^3 = E(x - E(x))^3$$

AS = 0
distribut° symétrique

AS > 0
distribut° présente une asymétrie à gauche

AS < 0
distribut° présente une asymétrie à droite

* coefficient d'aplatissement : (KURTOSIS)

$$AP = \frac{m_4}{s^4} \; ; \; m_4 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^4$$

AP = 3
Distribut° a un coef. AP similaire à la distribut° normale :
Distribut° mésocurtique

AP > 3
Distribut° sera plus tassée que la distribut° normale avec des queues épaisses :
Distribut° Leptocurtique

AP < 3
Distribut° présente des queues + fines que celle loi normale :
Distribut° Platicurtique

- Distribut° Normale
- chap 3 : Relation entre les variables

1. Distribut° Normale : (Gauss)

" forme de cloche "
" distribut° de référence "

→ propriétés : • unimodale et symétrique (moyenne = mode = médiane)

50% > moyenne
50% < moyenne

. écart type et moyenne

$(\mu - \sigma, \mu + \sigma)$ → 68,26% popul°

$(\mu - 2\sigma, \mu + 2\sigma)$ → 95,44% popul°

$(\mu - 3\sigma, \mu + 3\sigma)$ → 99,74%

→ vérificat° de la normalité d'une variable :
. cloche
. unimodale
. AS ≈ 0
. AP ≈ 3
. Diagrammes < PP Probability-Probability plot
QQ Quantile-Quantile plot

Normale + unimodale ⇓ moyenne = mode = médiane

1. Scores standardisés :

Score Z : score centré et réduit à partir d'un score individuel.
donne une indicat° précise de la posit° du score de l'individu au sein de la distribut°.

$$Z = \frac{X - \mu}{s}$$

distribut° normal centré réduite (système de mesure) où :
→ mode = moyenne = médiane = 0
↳ s = 1 (toujours) "écart type"

Chap 3

① Liaison entre 2 variable qualitative
- coefficient χ^2
- Représentat° graphique

② Entre 2 quantitative
- coefficient de corrélat°
cov
- représentat° graphique

③ Entre 1 qualit. et 1 quantité.
- représentat° graphique

① - tableau de contingence :

$X \backslash Y$	y_1	y_j	y_L	total
x_1				
x_i		n_{ij}		n_i
x_K				
total		n_j		n

lignes totals des individus

$n_i = \sum_j n_{ij}$

$n = \sum_{ij} n_{ij} = \sum_i n_i = \sum_j n_j$

colonnes totales

$$\chi^2 = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\left(\frac{n_i n_j}{n}\right)}$$

(valeur Théorique si indépendance)

$\chi^2 = 0$: indépendance (profils identiques)

$\chi^2 \neq 0$, ↗ si profils différent ↗

- Diagrammes en bâtons superposés :

② - tableau de données (forme de n couples $(x_i; y_i)$ $i = 1, \dots, n$

- Représentat° graphique : Nuage de points M_i

- coefficient de corrélat°, $r \in [-1, 1]$
= indicateur de variat° de 2 variables quantitatives continues

$r > 0$: sens de corrélat° positive

$r < 0$: sens contraire & corrélat° négative

$$r_{xy} = \frac{cov(x,y)}{s_x s_y} \quad ; \quad cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

r : r → sans unité, indépendant de l'origine choisie.

si il est proche de -1 ou 1 : il y a une relation presque affine entre x et y

(9)

si x qualitative à k modalités, l'ensemble des n individus peut être subdivisé en k groupes sur lesquels x est constante

si y est quantitative, les k groupes peuvent être représentés par une boîte à moustache de façon à pouvoir les comparer.

? (girl, boy) ?