Statistiques descriptives

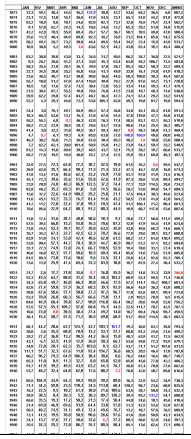
Guillaume Wisniewski
guillaume.wisniewski@limsi.fr

LIMSI — Université Paris Sud

13 septembre 2014

Rappel/problématique

Ceci est un corpus



Conséquence

En l'état c'est inutilisable !
On a besoin

- avoir un aperçu / résumé
- identifier certaines tendances / répondre à certaines questions

Vocabulaire :

- uni varié : on s'intéresse à 1 valeurs
- multi varié : plusieurs valeurs + leurs « interactions »

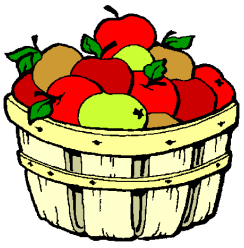
Running example

Quelle est la plus grosse variété de pomme ?

La démarche statistique

1. collecte des données
2. description des données (résumé), identifier les caractéristiques pertinentes
3. test d'hypothèse différencier les observations significatives des autres
4. estimation : extraire de l'information pour prédire les caractéristiques de nouveaux exemples

Dans notre cas...



Pour chaque variété de pomme, collecter les tailles d'un échantillon

Difficulté

échantillon $\Rightarrow$ propriété générale

Problématique locale

Comment résumer un ensemble de valeur ?

- de manière quantitative
- de manière qualitative

1er descripteur : la moyenne

Notations :

- population de n individus/exemples/échantillons
- chaque individu x_i est décrit par un ensemble de valeur
- techniquement : $x_i \in \mathbb{R}^d$

Moyenne :

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

- décrit le comportement « moyen »
- critère de position
- central tendancy

Exemple

Pour un exemple univarié

- exemple : note à un examen, taille d'une pomme
- $\mathcal{D} = \{12, 14, 8, 18, 3\}$
- $\mu =$

Pour un exemple multivarié

- exemple : note à l'examen d'info et à l'examen de physique
- $\mathcal{D} = \left\{ \begin{pmatrix} 12 \\ 16 \end{pmatrix}, \begin{pmatrix} 14 \\ 4 \end{pmatrix}, \begin{pmatrix} 8 \\ 8 \end{pmatrix}, \begin{pmatrix} 18 \\ 19 \end{pmatrix}, \begin{pmatrix} 3 \\ 8 \end{pmatrix} \right\}$
- $\mu =$

Exemple

Pour un exemple univarié

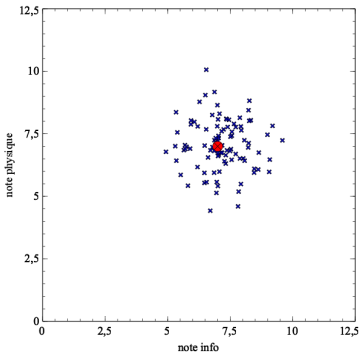
- exemple : note à un examen, taille d'une pomme
- $\mathcal{D} = \{12, 14, 8, 18, 3\}$
- $\mu = 11$

Pour un exemple multivarié

- exemple : note à l'examen d'info et à l'examen de physique
- $\mathcal{D} = \left\{ \begin{pmatrix} 12 \\ 16 \end{pmatrix}, \begin{pmatrix} 14 \\ 4 \end{pmatrix}, \begin{pmatrix} 8 \\ 8 \end{pmatrix}, \begin{pmatrix} 18 \\ 19 \end{pmatrix}, \begin{pmatrix} 3 \\ 8 \end{pmatrix} \right\}$
- $\mu = \begin{pmatrix} 11 \\ 11 \end{pmatrix}$

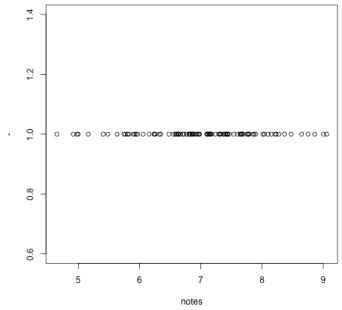
Interprétation

- population = nuage de points
- moyenne = « centre » du nuage (position)



Et en 1-D ?

Comment représenter des données en 1-D ?
La mauvaise solution :

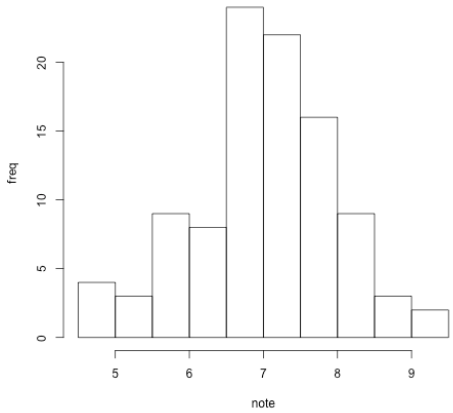


En pratique

- représenter la fréquence d'une valeur
- = histogramme
- nécessite de discrétiser les valeurs

⇒ distribution des données

Résultat



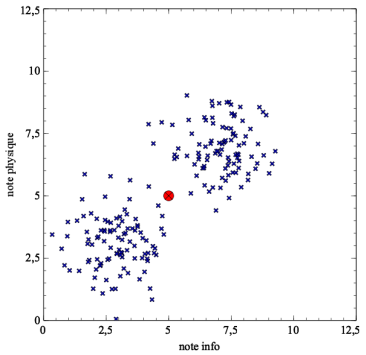
Le problème de la moyenne

La moyenne = bon descripteur d'éléments qui « se ressemble »

Exemple

- j'ai trois pommes qui font 600g ⇒ chaque pomme fait 200g
- vrai tant que toutes les pommes ont, à peu près, la même taille

En image

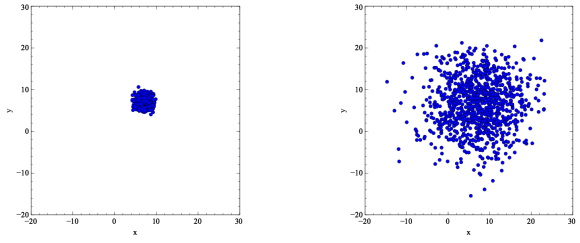


2e descripteur : variance

$$\sigma^2 = \frac{1}{n}\sum_{i=}^{n}(x_i - \mu)^2$$

- σ^2 = variance, σ = déviation standard
- $x_i - \mu$ = déviation de la moyenne = position par rapport à la moyenne
- mesure la dispersion des données autour de la moyenne

En image



Bilan

- dans beaucoup de cas, moyenne et variance sont suffisant pour décrire des données (avoir une idée de leur tête)
- essentiellement quand les données sont proches
- note pour plus tard : proche = distribuée selon une normale

3e descripteur : médiane

1. trier les exemples d'une population (ordre indifférent)
2. médiane = élément à la position $\frac{n}{2}$

- description de la position (comme la moyenne)
- moins sensible aux outliers / mieux adapté aux données peux similaires

Exemples

Considérons la population :

$$\mathcal{D} = \left\{ \underbrace{10, ..., 10}_{5 \text{ fois}}, 20 \right\}$$

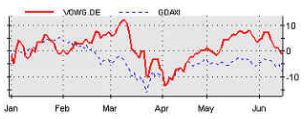
Deux descriptions de la centralité :

$$\mu = 11.6$$
$$\text{médiane} = 10$$

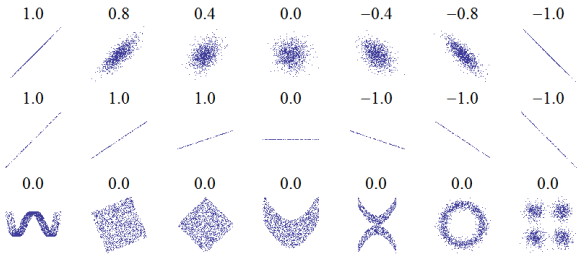
$\Rightarrow$ meilleur description de la note majoritaire

4e descripteur : corrélation

- 2 variables aléatoires X et Y (p.ex. une mesurant la largeur de la tête, l'autre sa hauteur)
- y a-t-il une « relation » entre ces deux variables ?
- deux caractéristiques de la relation :
  - « sens » (positif si X augmente quand Y augmente)
  - « force » (impact de la variation d'une des variables)



En image



Quantitativement

- hypothèse sur la « forme » de la relation entre X et Y (p.ex. linéaire)
- coefficient de Pearson (corrélation linéaire) d'une série de valeur $(x_i, y_i)_{i=1}^n$

$$\rho_{X,Y} = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot \sigma_x \cdot \sigma_y}$$

- attention au problème d'instabilité numérique lors du calcul
- indique la « force » d'une corrélation linéaire entre les données

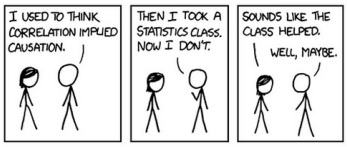
Autres mesures

- le coefficient de Pearson ne permet de détecter que certaines dépendances très particulières
- il existe des mesures plus générales permettant de détecter quasiment toutes les dépendances fonctionnelles :
  - information mutuelle
  - corrélation polychorique
  - copule
  - ...

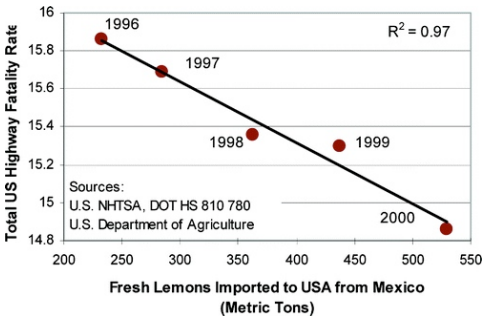
Corrélation et causalité

Il n'y a pas de lien entre corrélation et causalité

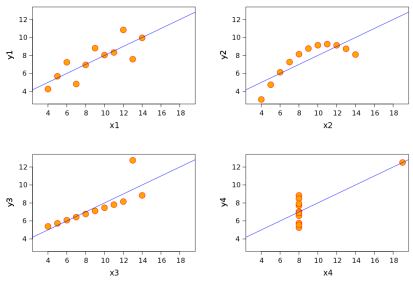
- corrélations : certaines valeurs « évoluent » dans le même sens
- causalité : lien « physique »
- quantification d'observations statistiques $\neq$ modélisation d'un système



Exemple



Conclusion : le danger des stat. descriptives



Toutes les propriétés statistiques de ces populations sont identiques (moyenne, variance, corrélation)