Université Claude Bernard Lyon 1
Année 2024-2025

Statistique pour l'Informatique

Notes du cours

version du 18 septembre 2024

Johannes Kellendonk

Table des matières

1 Statistique déscriptive **5**
1.1 Notions de base 5
1.2 Description des données avec une seule caractéristique 5
1.2.1 Notions 5
1.2.2 Représentation graphique des échantillons 6
1.2.3 Notions pour les variables numériques 6
1.3 Description des données avec deux caractéristiques 7
1.3.1 Fréquences 7
1.3.2 Covariance et coefficient de corrélation 8
1.3.3 Régression linéaire 9

Chapitre 1

Statistique déscriptive

Les notions traduites en anglais sont donnés en *italique*.

1.1 Notions de base

Une **population** est une collection d'individus (ou d'objets) avec certaines caractéristiques.

Un **échantillon** *(sample)* est une partie d'une population.

En statistique, une caractéristique des individus est aussi appellée une **variable**. Si I est l'ensemble des individus, on note x_i la caractéristique de l'individu i.

Une variable (caractéristique) peut être

- qualitative : par exemple une couleur
- quantitative (ou numérique) : un nombre (réel)
- discrète : finie ou dénombrable
- continue : si elle prend ses valeurs dans un intervalle de $\mathbb{R}$ (plus une condition de continuité que j'énoncerai plutard)

Un **paramètre** est un nombre inconnu mais fixe (une variable est variable).

L'effectif (ou taille) d'un échantillon I est le nombre de ses individus, c.à.d. le nombre d'éléments dans I, ce qu'on note $\#I$.

1.2 Description des données avec une seule caractéristique

Dans cette section on considère des échantillons avec individus qui ont une seule caractéristique. On note un tel échantillon $\{x_i : i \in I\}$ ou $\{x_1, \cdots, x_n\}$, où la caractéristique de l'individu $i \in I$ est notée x_i.

1.2.1 Notions

Fréquence absolue (ou l'effectif) de la caractéristique c de l'échantillon est le nombre d'individus avec caractéristique c, noté

$$\#\{i \in I : x_i = c\}.$$

Un **mode** est une caractéristique qui a la fréquence maximale. La **distribution de l'échantillon** est le tableau des fréquences de l'échantillon. La **distribution de la population** est le tableau des fréquences de la population.

La **fréquence** (ou fréquence relative) de la caractéristique c de l'échantillon est la fréquence absolue divisée par la taille de l'échantillon.

$$f(c) = \frac{\#\{i \in I : x_i = c\}}{\#I}.$$

Si la variable est numérique on définit la **fréquence cumulée** (ou fonction de répartition empirique) de la caractéristique c de l'échantillon par

$$F(c) = \frac{\#\{i \in I : x_i \leq c\}}{\#I}.$$

1.2.2 Représentation graphique des échantillons

Diagramme des colonnes *(bar chart)* Dans un diagramme des colonnes on représente les fréquences absolues (ou relatives) d'une caractéristique discrète par la hauteur d'une colonne.

Diagramme circulaire (Camembert) *(Pie chart)* Dans un diagramme circulaire (camembert) on représente les fréquences relatives d'une caractéristique discrète par la surface d'un segment dans un disque.

Histogramme d'un échantillon numérique Dans un histogramme on représente la fréquence relative que la valeur d'une caractéristique appartient à un certain intervalle par la surface d'un rectangle. Pour ceci la caractéristique doit être numérique (continue ou discréte) et on doit partitionner $\mathbb{R}$ en des intervalles. Cela veut dire qu'on choisit des nombres $r_0 < r_1 < \cdots < r_N$ *(break points)* donnant les intervalles $[r_{n-1}, r_n]$. (On choisit r_0 plus petit et r_N plus grand que les valeurs de la variable). On calcule les fréquences relatives que la caractéristique tombe dans les intervalles $[r_{n-1}, r_n[$,

$$f_n = \frac{\#\{i \in I : r_{n-1} \leq x_i < r_n\}}{\#I}.$$

et on trace, au dessus de tout intervalle $[r_{n-1}, r_n]$ un rectangle avec surface égale à f_n (et donc avec une hauteur égale à $h_n = \frac{f_n}{r_n - r_{n-1}}$). L'histogramme dépend de la partition choisie (ce qui peut être trompeur).

1.2.3 Notions pour les variables numériques

On considère un échantillon $x = \{x_1, \cdots, x_n\}$ (de taille n) avec $x_i \in \mathbb{R}$.

La **moyenne de l' échantillon** (moyenne empirique) *(mean)* x est

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

On note aussi $\bar{x} = m(x)$.

La **médiane de l'échantillon** *(median)* est la valeur qui partage les valeurs de x en deux parties : les petites et les grandes valeurs. Pour ceci il faut ordonner les valeurs. Soit x_k^* la k-ième plus petite valeur de l'échantillon (comptée avec multiplicité [1]). La médiane est alors définie comme

$$mediane(x) = \begin{cases} x^*_{\frac{n+1}{2}} & \text{si } n \text{ est impair} \\ \frac{1}{2}(x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}) & \text{si } n \text{ est pair.} \end{cases}$$

1. par exemple, pour l'échantillon 1 2 4 3 2 on obtient $x_1^* = 1$, $x_2^* = 2$, $x_3^* = 2$, $x_4^* = 3$, $x_5^* = 4$

Les **quartiles de l'échantillon** partagent les valeurs de x en quatre parties selon leur grandeur, Q_1, Q_2, Q_3 sont les valeurs au milieu de ces quatres parties. Comme ceci n'est pas toujours bien défini on prend

$$Q_1 = x^*_{\lfloor \frac{n}{4} \rfloor},\ Q_2 = x^*_{\lfloor \frac{2n}{4} \rfloor},\ Q_3 = x^*_{\lfloor \frac{3n}{4} \rfloor}$$

ou $\lfloor r \rfloor$ est le plus grand entier plus petit ou égal à que r.

La **variance de l'échantillon** *(variance)* x est

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

L'écart type de l'échantillon (déviation standard) *(standard deviation)* x est

$$\sigma(x) = \sqrt{V(x)}.$$

La **variance sans biais de l'échantillon** *(variance)* x est

$$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

L'écart type sans biais de l'échantillon (déviation standard) *(standard deviation)* x est

$$s(x) = \sqrt{var(x)}.$$

Il y a juste un facteur $\frac{n}{n-1}$ de difference entre la variance et la variance sans biais, ce qui est negligible si n est grand. Mais la commande en python travaille avec la variance sans biais.

La somme de deux échantillons $x = \{x_1, \cdots, x_n\}$ et $y = \{y_1, \cdots, y_n\}$ est $x + y = \{x_1 + y_1, \cdots, x_n + y_n\}$ et si $r \in \mathbb{R}$ on pose $rx = \{rx_1, \cdots, rx_n\}$.

Si la caractéristique d'un échantillon a une unité, par exemple cm, alors la moyenne, la médiane, les quartiles et l'èkart type ont tous la même unité, pendant que la variance a l'unité au carrée (par exmple cm^2). C'est pour cela que c'est l'écart type qui nous donne l'information de l'ampleur de la déviation des valeurs de la moyenne (statistiquement).

Proposition 1 *Le moyenne est additive $m(x+y) = m(x)+m(y)$ et $m(rx) = rm(x)$ pour $r \in \mathbb{R}$. La variance et l'écart type ne sont pas additives mais $V(rx) = r^2V(x)$ et $\sigma(rx) = |r|\sigma(x)$, pour $r \in \mathbb{R}$.*

1.3 Description des données avec deux caractéristiques

On considère maintenant des échantillons qui ont des individus avec deux caractéristiques. Autrement dit, les individus ont un couple de caractéristiques. On note $(x, y) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$.

1.3.1 Fréquences

La **fréquence absolue** de la caractéristique (a, b) de l'échantillon (x, y) est $\#\{i \in I : x_i = a, y_i = b\}$.

La **fréquence relative** de la caractéristique (a, b) de l'échantillon est

$$f(a, b) = \frac{\#\{i \in I : x_i = a, y_i = b\}}{\#I}.$$

On a

$$f(a) = \frac{\#\{i \in I : x_i = a\}}{\#I} = \sum_b \frac{\#\{i \in I : x_i = a, y_i = b\}}{\#I},$$

et

$$f(b) = \frac{\#\{i \in I : y_i = b\}}{\#I} = \sum_a \frac{\#\{i \in I : x_i = a, y_i = b\}}{\#I},$$

La **table de contingence** est la table des fréquences (absolues ou relatives) des couples des caractéristiques.

On peut visualiser un échantillon avec deux caractéristiques numériques dans un diagramme de nuage (scatterplot). C'est l'ensemble des points dans $\mathbb{R}^2$ qui corresponds aux couples (x_i, y_i) de l'échantillon.

1.3.2 Covariance et coefficient de corrélation

La moyenne de (x, y) est

$$m(x, y) = (m(x), m(y))$$

donc c'est un point de $\mathbb{R}^2$. La **covariance** de (x, y) est une matrice 2×2

$$\begin{pmatrix} V(x,x) & V(x,y) \\ V(y,x) & V(y,y) \end{pmatrix}$$

ou

$$V(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

en particular, $V(x, x)$ est la variance de x et la matrice est symétrique.

Comme pour la variance il y a aussi une version non-biaisée de la covariance, elle est donnée par

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Proposition 2 *La covariance est additive dans la premiére variable et $V(rx, y) = rV(x, y)$ pour $r \in \mathbb{R}$. Par symétrie ceci s'applique aussi à la deuxième variable.*

Le **coefficient de corrélation** des deux variables x et y est

$$cor(x, y) = \frac{V(x, y)}{\sqrt{V(x, x)}\sqrt{V(y, y)}},$$

pourvu que les variances de x et de y ne sont pas nulles. On note que si les deux caractéristiques ont des unités distinctes, par example cm et kg, alors les unités de la matrice de covariances sont mixtes. Par contre le coefficient de corrélation n'a pas d'unité. C'est pour cela que c'est ce coefficient qui nous donne une information sur la corrélation.

Théorème 1 *On suppose que $V(x, x) \neq 0$ et $V(y, y) \neq 0$. On a*

$$-1 \leq cor(x, y) \leq 1.$$

De plus,

$cor(x, y) = 1$ si et seulement s'il existe $a > 0$, $b \in \mathbb{R}$ t.q. $\forall i \in I : y_i = ax_i + b$
$cor(x, y) = -1$ si et seulement s'il existe $a < 0$, $b \in \mathbb{R}$ t.q. $\forall i \in I : y_i = ax_i + b$

On note que si $cor(x, y) = \pm 1$ alors y_i est déterminé par x_i et vice versa, donc il y a une correlation parfaite entre les deux characteristiques. Par contre, si $cor(x, y)$ est proche de 0 les données ne sont pas corrélées.

1.3.3 Régression linéaire

Etant donné un échantillon à deux variables numériques $(x, y) = \{(x_1, y_1), \cdots (x_n, y_n)\}$, on cherche à déterminer les paramétres $a \in \mathbb{R}^*, b \in \mathbb{R}$ de la droite

$$y = ax + b$$

qui approche le mieux les données (x_i, y_i). Ceci a un sens seulement si le coefficient de corrélation $cor(x, y)$ est proche de 1 ou de -1 et on peut donc soupconner un corrélation linéaire entre les deux variables. L'erreur entre la valeur y_i de l'échantillon et la valeur donnée par l'équation de la droite est qunatifié par la fonction

$$\mathcal{E}(a, b) := \frac{1}{n} \sum_{i=1}^{n} (y_i - (ax_i + b))^2.$$

$\mathcal{E}(a, b)$ est minimal si

$$a = \frac{V(x, y)}{V(x, x)}$$
$$b = \overline{y} - a\overline{x}$$

et donc la droite cherchée est donnée par

$$y = \frac{V(x, y)}{V(x, x)}(x - \overline{x}) + \overline{y}.$$