Probability and Statistics for Data Science

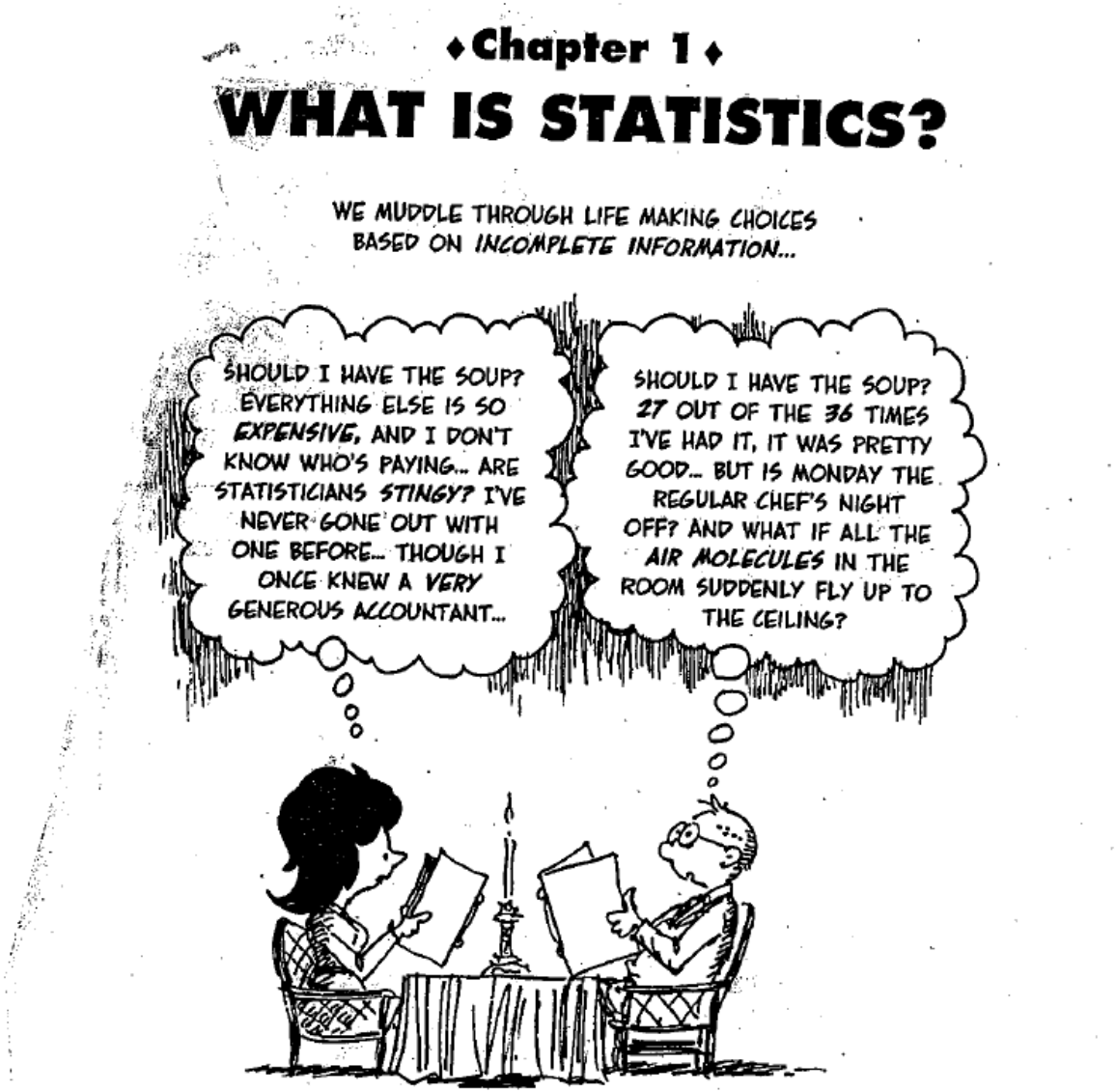
S. Gadat - Toulouse School of Economics

Lecture 1 - Introduction to Statistics and Statistical models

2023

Outline

Aims


♦Chapter 1♦
WHAT IS STATISTICS?
WE MUDDLE THROUGH LIFE MAKING CHOICES BASED ON INCOMPLETE INFORMATION...
SHOULD I HAVE THE SOUP? EVERYTHING ELSE IS SO EXPENSIVE, AND I DON'T KNOW WHO'S PAYING... ARE STATISTICIANS STINGY? I'VE NEVER GONE OUT WITH ONE BEFORE... THOUGH I ONCE KNEW A VERY GENEROUS ACCOUNTANT...
SHOULD I HAVE THE SOUP? 27 OUT OF THE 36 TIMES I'VE HAD IT, IT WAS PRETTY GOOD... BUT IS MONDAY THE REGULAR CHEF'S NIGHT OFF? AND WHAT IF ALL THE AIR MOLECULES IN THE ROOM SUDDENLY FLY UP TO THE CEILING?

Aims

Define statistics ?

A **decision aid** tool : a set of methods to help **summarizing data** (extract relevant information) from the observation of a **random phenomenon**.

This course aims at providing a **general mathematical framework** for these methods in order to define general principles.

The tools are "models" (to simplify the complexity of data) and "principles" (to guide the choices).

This course is not about a particular method like descriptive statistics, experimental design, data mining, scoring, survey sampling etc.

Fundamental assumption

The statistician assumes that the random phenomenon (or experiment) that generates the data (DGP for data generating process) can be described by :

- the space of observations Ω : the different issues (or outcomes) of the experiment
- a sigma-algebra $\mathcal{A}$ of measurable events : they are the subsets of Ω which can be assigned a probability.
- a family $\mathcal{P}$ of probability laws P defined on $(\Omega, \mathcal{A})$.

Later on, we will specify a bit this set $\mathcal{P}$.

Outline

Fundamental assumption

Fundamental assumption : the data set is made of one or several realizations/observations x of a random variable X with values in Ω with unknown law P (called the true distribution) belonging to $\mathcal{P}$.

The three elements $(\Omega, \mathcal{A}, \mathcal{P})$ compose the **statistical model**.

The objective is to use the data x to draw some conclusions about the unknown probability distribution (or "population") P.

Different types of conclusions about P :

- mean/expectation
- variance
- entire distribution function
- ...

A first example

Experiment : at the end of the semester, a reporter interviews (with replacement) a subset of n M1 students with the following question : did you enjoy the Proba&Stat class ?

Possible answers are

- yes (coded 1)
- no (coded 0)

What is the statistical model ?

If X_i is the answer of the i-th student, then :

- the answers are i.i.d.
- the answers are some realizations of some r.v. X_i with Bernoulli probability law with parameter

$$p = \mathbb{P}(X_1 = 1) = \cdots = \mathbb{P}(X_n = 1),$$

- p represents the proportion of satisfied students in the entire class.

A first example

Experiment : at the end of the semester, a reporter interviews (with replacement) a subset of n M1 students with the following question : did you enjoy the Proba&Stat class ?

Possible answers are

- yes (coded 1)
- no (coded 0)

One can take $\Omega = \{0,1\}^n$ and $X = (X_1, \cdots, X_n)$.

The law of X is then $\mathcal{B}(1,p)^{\otimes n}$.

p is unknown because the reporter does not interview all the students (origin of the randomness).

A first example

The data is the collection of 0-1 answers $(x_1, \cdots, x_n)$ from which one can calculate the observed proportion of satisfied students

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

The family of probability laws $\mathcal{P}$ is the set of laws :

$$\mathcal{P} = \left\{ \mathcal{B}(1, p)^{\otimes n}, p \in [0, 1] \right\}$$

The aim is to make inference on p from the observation of the empirical frequency $\hat{p}$.

Another choice for the same experiment

With the same experiment and the same data, on could choose

$\Omega = \{0, 1, \cdots, n\}$, $X = \sum_{i=1}^{n} X_i$ and $P = \mathcal{B}(n, p)$

We will prove in a forthcoming lecture that there is
no difference in terms of quantity of information about p between the two options.

A second example

Experiment :

- An insurance company counts the number X of claims that occurred a given year.
- The randomness of this number comes from the fact that X varies from year to year in a non deterministic way.
- The set Ω of possible values of X is equal to the set of integers $\mathbb{N}$.

From the description of the experiment and without additional assumptions, the distribution of X is not entirely defined apart from the fact that X is a random variable with integer values.
The family $\mathcal{P}$ of probability laws $P \equiv P_X$ of X satisfies
$P(k) = P_X(k) = \mathbb{P}(X = k)$, with $\sum_{k\in\Omega} P(k) = \sum_{k=0}^{+\infty} P(k) = 1$.

A second example

Experiment :

- An insurance company counts the number X of claims that occurred a given year.
- The randomness of this number comes from the fact that X varies from year to year in a non deterministic way.
- The set Ω of possible values of X is equal to the set of integers $\mathbb{N}$.

A commonly used assumption is that X is Poisson distributed with mean λ :

$$\mathbb{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \quad \forall k \in \mathbb{N}.$$

Here $\mathcal{P}$ is the set of Poisson laws on $\mathbb{N}$ with unknown parameter $\lambda > 0$, and the data is the observed value x_i of X for year i, with $i = 1, \ldots, n$.

Type of inference

Several types of inference can be considered, for example

- in example 1 where $X_i \rightsquigarrow \mathcal{B}(1, p)$: guess the value of p
  $\twoheadrightarrow$ point estimation
- in example 1 : give a bracket for the value of $p \in [p_\ell, p_u]$ =?
  $\twoheadrightarrow$ confidence interval
- in example 1 : decide whether p is larger (or smaller) than (or equal to) a given threshold p_0
  $\twoheadrightarrow$ test *e.g.* $p > p_0$ against $p \leq p_0$
- in example 2 (counts of the number of claims) : predict the value of X for next year
  $\twoheadrightarrow$ prediction

The latitude of the statistician

In general, X and Ω are determined by the experiment up to some finite number of options as in example 1 where one can take $X = (X_1, \dots, X_n) \in \Omega = \{0,1\}^n$ or $X = \sum_{i=1}^n X_i \in \Omega = \{0, 1, \dots, n\}$.

But it is most of the time the case that, from the description of the experiment, the probability family is not completely specified :
the statistician has some latitude/leeway and may introduce more or less additional assumptions to work with a tractable model.

However, he has to pay this freedom by justifying his model choices :
testing validity of his assumptions.

He also has to justify his choice of inference method :
use principles to stand up for his choice (*e.g.* how to compare two estimation methods and select the most efficient one ?)

What you are expected to be able to do !

Given a statistical model $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \times\}$

- Compute an expectation of a random variable.
- Compute a variance, or a variance/covariance matrix
- Compute some quantiles when possible
- Use the transfer theorem : for any measurable function f, identify the distribution of $Y = f(X)$.
- Find or manipulate the probability density function
- Compute or manipulate the cumulative distribution function

The exponential model family

Let $(\Omega, \mathcal{P})$ be a parameteric model for a scalar valued random variable. Here, the r.v. of interest $X \in \Omega \subset \mathbb{R}$ with distribution

$$P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

The model is equipped with the Borel σ-algebra
(when the σ-algebra is evident, it is omitted).

$(\Omega, \mathcal{P})$ is an exponential model if there exist known real functions $h(x)$, $C(\theta)$, $T_j(x)$, $Q_j(\theta)$ such that the density f_θ of X can be written as

$$f_\theta(x) = h(x)\, C(\theta)\, \exp\left[\sum_{j=1}^{r} Q_j(\theta)\, T_j(x)\right].$$

Note : for a given such family, the functions $h(x)$, $C(\theta)$, $T_j(x)$, $Q_j(\theta)$ are not unique.

The exponential model family

The family comprises the following models :

- Gaussian
- Poisson
- Binomial,
- Exponential
- Gamma
- Chi-square, etc.

Counterexample : The family does not comprise the set of uniform laws on $[\alpha, \beta]$ when α and β vary in $\mathbb{R}$ (can be discrete or continuous).
Exercise : Identify for each of the previous family h, Q and T.

Reparameterization of a model

Reparameterizing a model consists in replacing an original parameter θ by a new parameter $\lambda = \phi(\theta)$, for a one-to-one transformation ϕ.

This does not change the given family of distributions.

Example : exponential distribution

- pdf $f_\theta(x) = \frac{1}{\theta} \exp(-\frac{x}{\theta}), \quad \theta > 0$
- pdf $f_\lambda(x) = \lambda \exp(-\lambda x), \quad \lambda = \frac{1}{\theta} > 0$

The mean of this distribution is $\theta = \mathbb{E}(X)$ so that :

- in the first parameterization, the parameter (*i.e.* θ) can be interpreted as the mean ;
- in the second parameterization, the parameter (*i.e.* λ) can be interpreted as the hazard rate (density divided by survival function).

The exponential model family

If moreover the functions Q_j satisfy $Q_j(\theta) = \theta_j$, then we say that θ is the **natural parameter** of the model.

The statistics $T_j(X)$ are called **"sufficient statistics"** of the model : they play a particular role.

This role and the name (sufficiency) will be further explained in the chapter "Elements of information theory" (second semester).

The exponential model family

The role in statistics of the exponential family is to deal *e.g.* with :

- estimation theory : chapter "Point estimation" (second semester)
- hypotheses testing : chapter "Confidence intervals and Testing" (second semester)
- theory of generalized linear models : logistic regression, scoring (Master 2)
- bayesian statistics : existence of a conjugate prior distribution (Master 2)

Outline

Transfer theorem

You are expected to be able to solve the following problems.

Transfer theorem - 1D example

Exercise : For any random variable X of density f w.r.t. the Lebesgue measure, identify the distribution of $Y = e^X$.
When $X \sim \mathcal{N}(\mu, \sigma^2)$, Y is a log-normal random variable.

Transfer theorem - 1D example

Exercise : For any random variable X of density f w.r.t. the Lebesgue measure, identify the distribution of $Y = X^{-1}$.
When $X \sim \mathcal{N}(0,1)$, identify the density of Y. When X is a centered Cauchy random variable, show that Y is also a Cauchy random variable.

Transfer theorem - 1D example

Exercise : When $X \sim \mathcal{N}(0,1)$, identify the density of $Y = X^2$ (chi-square distribution).

Transfer theorem - 2D example

Exercise : When X and Y are *independent* $\mathcal{N}(0,1)$ random variable, identify the density of $(U,V) = \left(\frac{X+Y}{\sqrt{2}}, \frac{X-Y}{\sqrt{2}}\right)$.

Transfer theorem - 2D example

Exercise : Consider X and Y two random variables, identify the density of $U = XY$ and the density of $V = \frac{X}{Y}$. Simplify a bit the results when X and Y are independent.

Transfer theorem - 2D example

Exercise : When X and Y are *independent* $\mathcal{N}(0,1)$ random variable, identify the density of $U = \frac{X}{Y}$.

Transfer theorem - 2D example

Exercise : When X and Y are *independent* $\mathcal{N}(0,1)$ random variable, identify the density of $(R,\theta) = (\sqrt{X^2+Y^2}, \tan^{-1}(Y/X))$.

Outline

Goal

We aim to introduce some key tools that allow to make some easier computations. Among them :

- The generating function G_X, defined for any integer valued random variable X :

$$G_X : s \longmapsto \sum_{k=0}^{\infty} \mathbb{P}(X = k) s^k = \mathbb{E}[s^X]$$

- The moment generating function M_X :

$$M_X : s \longmapsto \mathbb{E}[e^{sX}]$$

  M_X also refers to the Laplace transform.

- The caracteristic function φ_X :

$$\varphi_X : s \longmapsto \mathbb{E}[e^{isX}],$$

  which is also the Fourier transform of the density of f.