MATH 562: STATISTICAL INFERENCE

Rafiki's Notes

Rafael Barroso
Ingenierie Mathematique
École Polytechnique Fédérale de Lausanne

September 20, 2025

Contents

1 Introduction & Basics

Data, dara and deitar; in short, statistics is the art of learning from data. We usually start by **considering a question or problem** with our goal in mind being to **solve this by using data**. Our main tasks are to provide enough evidence in order to address the question or problem at hand and ideally we'd love to **draw a conclusion** or **reach a decision**.

In this course, we will merely discuss **how express the evidence** attained from the data given (usually, in most cases, we recieve really shitty data), but we depend on the quality, choice and methodology of the data collection, which really impacts our evidence and clarity of decision.

Remark 1.1. The collection of data usually has **structure** (i.g. patterns) and **'random' variation** (i.g. noise), hence, any conclusion we might reach is *unbcertain*; therefore called an *inference*.

So, a *statistical inference* aims to use probability theory (big boy language; mathematical framework) in order to explain the variation in the data and to quantify the uncertainty in our conclusions.

As most mathematical fields, statistical theory has many branches, the main ones being:

- **Desctiptive statistics/ Exploratory Data Analysis (EDA)**: which is the art of **summarizing** and **visualizing** data.
- **Inferential statistics**: a way of using probability theory in order to provide evidence for the statements we *infer* abotu a certain situation.
- **Algorithmic methods**: often used togetrher with inferential statistics, these are methods that are used to **fit models** to data in order to make predictions or decisions. Nowadays, these methods are often associated with **machine learning** and **artificial intelligence**.

Remark 1.2. Note that all of the above end up using **probability theory**. It is also important to define which type of analysis is being conduted in order to prevent horribly biased conclusions.

Circling back to the elephant in the room, **data collection**, we have two main types of data:

- **Observational data**: where we merely observe the data, and try to make sense of it (little or no control whatsoever as for the data collection). This type of data is usually plagued with confounding variables (lurking variables) that might bias our conclusions.
- **Experimental data**: where we have more control over the data collection process, e.g. we can plan an investigaton, design the experiment and gather the data. This method typically provides "stronger" data, hence stronger conclusions.

1.1 Some definitions

Since statisticians use a lot of jargon, here are some important definitions that will be used throughout the course.

Definition 1.3. Main jargon used in statistics:

1. **Population:** The entire collection of individuals or objects about which information is desired.

2. **Sample:** A subset of the population that is used to gain information about the entire population.

3. **Random variable:** A variable whose value is subject to variations due to chance (i.e. randomness). It can take on different values, each with an associated probability. It can be *discrete* (e.g. number of heads in 10 coin flips) or *continuous* (e.g. height of a person). It will be more formally defined later on, and we will use it more than a hight-strung porn addict uses pornhub.

4. **(Probability) density function:** A function f that describes the probability for a random variable X to take on a given value x i.e. $f(x) = \mathbb{P}(X = x)$.

   - For a discrete random variable, it is called a **probability mass function (pmf)**.
   - For a continuous random variable, it is called a **probability density function (pdf)**.

5. **(Cumulative) distribution function:** A function F that describes the probability for a random variable X to take on a value less than or equal to x i.e. $F(x) = \mathbb{P}(X \leq x)$.

6. **Statistical model:** For data y, a statistical model is a density function $f(y)$ defined for $y \in \mathcal{Y}$.

7. **Parametric model:** $f \equiv f(y; \theta)$ where f is determined by parameters $\theta \in \Theta \subset \mathbb{R}^d$. If no such θ exists, the model is called **non-parametric**.

8. **Family of models:** Sometimes used to stress the idea that there might be many posibilities: $\{f(y;\theta) | \theta \in \Theta\}$.

There are many other definitions that will be introduced as we go along the course. Also there are some caveats to the above definitions.

Remark 1.4. A parameter θ is usually split into two types $\theta = (\psi, \lambda)$ where ψ is the parameter of interest and λ is a nuisance parameter (not of interest, but still has to be accounted for). i.g. whenever we wish to estimate the variance of a population we must also estimate the mean; in this case, the nuisance parameter.

We must also denote the notation that will be used throughout the course as follows:

- Vectors are allways **column vectors**, with row vectors being denoted by a superscript T (transpose).

- Lowercase letters denote **scalars**, e.g. $c, d, n \in \mathbb{C}$ one may also take $\mathbb{C}$ to be any arbitrary field k (if you have no regard for beauty, you may as well fix $k = \mathbb{R}$); for a non-mathematical person, scalars = constants.

- Uppercase letters denote **random variables**, e.g. X, Y, Z and their realizations (actual values) are denoted by lowercase letters x, y, z.

- Greek letters denote **parameters**, e.g. $\theta, \lambda, \mu, \sigma$; α is usually reserved for significance levels.

1.2 ProBABYlity theory basics

In order to understand statistical inference, we must first understand probability theory. Probability theory is a mathematical framework that allows us to quantify uncertainty and randomness. It provides the tools and concepts needed to model and analyze random phenomena.

Definition 1.5. A probabulity space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is the sample space, i.e. the set of all possible outcomes of a random experiment.

- $\mathcal{F}$ is a sigma-algebra on Ω, i.e. a collection of subsets of Ω that:

  1. includes the empty set i.e. $\emptyset \in \mathcal{F}$
  2. is closed under *complementation* i.e. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ (this actually tells us that $\emptyset^C = \Omega \in \mathcal{F}$).
  3. is closed under countable unions i.e. if $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- $\mathbb{P}$ is a probability measure, i.e. a function that assigns a probability to each event in $\mathcal{F}$. More formally, $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that:

  1. For any event $A \in \mathcal{F}$, we have $0 \leq \mathbb{P}(A) \leq 1$.
  2. $\mathbb{P}(\Omega) = 1$ (the probability of the entire sample space is 1).
  3. For any countable collection of *disjoint* events $A_1, A_2, \ldots \in \mathcal{F}$, we have $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(\mathcal{A}_i)$.

Well, that was a mouthful, but it is important to understand the above definition as it is the foundation of probability theory and to thank daddy Kolmogorov for it. We would like to specify some important notions stated above. A collection of sets $\{A_i\}_{i \in \mathcal{I}}$ is said to be **disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$. The **complement** of a set A is defined as $A^c = \Omega \setminus A = \Omega - A$.

Remark 1.6. We call $(\Omega, \mathcal{F})$ a measurable space, and the elements of $\mathcal{F}$ are called measurable sets or events. Also, from the definitions give, we may deduce the **inclusion-exclusion principle**: for any two events $A, B \in \mathcal{F}$, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

If we have that $\mathbb{P}(\mathcal{B}) > 0$ where $\mathcal{B} \subset \mathcal{F}$ we can define *conditional* probabilities $\mathbb{P}(\mathcal{A}|\mathcal{B})$ and is read as "the probability of $\mathcal{A}$ given $\mathcal{B}$".

Definition 1.7. We define the conditional probability of an event $\mathcal{A}$ given an event $\mathcal{B}$ with $\mathbb{P}(\mathcal{B}) > 0$

$$\mathbb{P}(\mathcal{A}|\mathcal{B}) := \frac{\mathbb{P}(\mathcal{A} \cap \mathcal{B})}{\mathbb{P}(\mathcal{B})}$$

A really important result in probability theory is the notion of the multiplication rule, which is a direct consequence of the definition of conditional probability and one may derive one of the 'holy grail' theorems in probability theory known as **Bayes Theroem** (which in my opinion is like applauding a baby for saying the word 'mum'; with a bit of thought and playfulness, anyone can derive this).

Remark 1.8. For any two events $\mathcal{A}, \mathcal{B} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{B}) > 0$, we have the multiplication rule

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A}|\mathcal{B})\mathbb{P}(\mathcal{B})$$

Please pause and ponder on this definition. If you for some reason have not yet thought, i'll spoil you (fucking lazy ass): If these two events $\mathcal{A}, \mathcal{B}$ are independent, then $\mathbb{P}(\mathcal{A}|\mathcal{B}) = \mathbb{P}(\mathcal{A})$. Please now think about the multiplication rule of two independent events.

$$\text{Bayes' theorem:} \quad \mathbb{P}(\mathcal{A}|\mathcal{B}) = \frac{\mathbb{P}(\mathcal{B}|\mathcal{A})\mathbb{P}(\mathcal{A})}{\mathbb{P}(\mathcal{B})} \tag{1}$$

Whooooow, lot's of definitions; these are all we offer on this brief part of the section, but get your butt cheeks ready for more as the following parts contain even more vicious definitions.

1.3 Even more definitions

Let us now introduce a few more notions that will be central throughout this course. We have already seen what a random variable is (a way to attach numbers to random outcomes), and what a probability space is. Now we refine these ideas by specifying how random variables distribute their randomness.

Definition 1.9 (Distribution and density)**.** Given a random variable X, its **cumulative distribution function** (cdf) is

$$F_X(x) := \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

It tells us how much probability mass has been accumulated up to x.
If F_X is differentiable, we can define the **probability density function** (pdf)

$$f_X(x) := \frac{d}{dx} F_X(x).$$

This is why we say the density is the “derivative” of the distribution function. Intuitively, $f_X(x)$ tells us how tightly probability is packed around x. For continuous random variables,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

If you think of F_X as the big ol’ soda machine that pours probability, then f_X is how fast the soda is coming out at a given point. Slow pour (flat density) means spread out probability; fast pour (spike) means a lot of probability concentrated nearby.

Definition 1.10 (Joint, marginal, and conditional densities)**.** For a pair of random variables (X, Y) with joint density $f_{X,Y}(x, y)$ we define:

- The **marginal density** of X:

  $$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy,$$

  and similarly for $f_Y(y)$. Intuitively, you “forget” about one variable by integrating it out.

- The **conditional density** of X given $Y = y$ (when $f_Y(y) > 0$):

  $$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

  This expresses how X behaves once we fix $Y = y$.

Definition 1.11 (Independent and identically distributed samples)**.** A collection of random variables $X_1, \dots, X_n$ is called **independent and identically distributed (i.i.d.)** if:

1. Independence: their joint distribution factors as

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i).$$

2. Identically distributed: each X_i has the same distribution F_X.

This means we are drawing observations from the same population, and none of them knows (or cares) about the others.

We now introduce the notion of quantiles, these are basically the "percentile checkpoints" of your distribution. The median is just the 50% checkpoint — the Switzerland of statistics, perfectly neutral.

Definition 1.12 (Quantiles and order statistics)**.** For $0 < \alpha < 1$, the α**-quantile** of a random variable X with cdf F_X is

$$q_\alpha := \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}.$$

Special cases:

- The median is $q_{0.5}$.
- Quartiles are $q_{0.25}$ and $q_{0.75}$.

Quantiles generalize the idea of medians to any percentage split of the probability mass.
Related to quantiles are the notions of **minimum**, **maximum**, **infimum**, and **supremum** for sets $A \subset \mathbb{R}$:

- $\min A$: the smallest element of A (if it exists).
- $\max A$: the largest element of A (if it exists).
- $\inf A$: the greatest lower bound (might not belong to A).
- $\sup A$: the least upper bound (might not belong to A).

Note that the median of X is simply a number m such that half of the probability lies to the left of m and half to the right. It's the statistical equivalent of sitting right in the middle of a bus, no matter how many people get in.

1.4 Moments (a.k.a. expectations and friends)

Time to arm ourselves with the workhorse concepts of inference: expectations, variances, covariances, and their conditional cousins. Everything here lives on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables are measurable maps into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as defined earlier.

Definition 1.13. Expectation (a.k.a. the first moment). Let X be a real-valued random variable. We say X is *integrable* if $\mathbb{E}[|X|] < \infty$; in that case the (Lebesgue) expectation is

$$\mathbb{E}[X] \;=\; \int_\Omega X(\omega)\, d\mathbb{P}(\omega) \;=\; \int_\mathbb{R} x\, dF_X(x),$$

where F_X is the cdf of X. In common cases:

$$\text{(discrete)} \quad \mathbb{E}[X] = \sum_{x \in \mathsf{supp}(X)} x\, p_X(x), \qquad \text{(continuous)} \quad \mathbb{E}[X] = \int_\mathbb{R} x\, f_X(x)\, dx,$$

with p_X the pmf and f_X the pdf (when it exists). More generally, for a measurable function $g : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}[\,|g(X)|\,] < \infty$,

$$\mathbb{E}[g(X)] \;=\; \int_\mathbb{R} g(x)\, dF_X(x) = \begin{cases} \sum_x g(x)\, p_X(x), & \text{discrete,} \\ \int_\mathbb{R} g(x)\, f_X(x)\, dx, & \text{continuous.} \end{cases}$$

Indicator trick: for $A \in \mathcal{B}(\mathbb{R})$, $\mathbb{E}[\mathbf{1}_{\{X \in A\}}] = \mathbb{P}(X \in A)$.

Some key properties (whisper them before every exam) we must keep close to our hearts are the following:

- **Linearity:** $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$ (no independence needed).
- **Monotonicity:** If $X \le Y$ a.s., then $\mathbb{E}[X] \le \mathbb{E}[Y]$.
- **Jensen:** If φ is convex and $\mathbb{E}[|X|] < \infty$, then $\varphi(\mathbb{E}[X]) \le \mathbb{E}[\varphi(X)]$.

Worked micro-examples.

- If $X \sim \text{Bernoulli}(p)$, then $\mathbb{E}[X] = p$ and $\mathbb{E}[g(X)] = g(0)(1-p) + g(1)p$.
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $g(x) = e^x$, then $\mathbb{E}[e^X] = \exp\big(\mu + \frac{1}{2}\sigma^2\big)$. (One-line proof with the mgf below or by completing the square. Proof contained in ex. sheet 2)

Definition 1.14. Variance and higher (central) moments. The *variance* of an integrable X with finite second moment is

$$\operatorname{Var}(X) \;:=\; \mathbb{E}\big[(X-\mathbb{E}[X])^2\big] \;=\; \mathbb{E}[X^2]-\mathbb{E}[X]^2.$$

Properties:

- $\operatorname{Var}(aX+b)=a^2\operatorname{Var}(X)$.
- If X, Y have finite second moments, then $\operatorname{Var}(X+Y)=\operatorname{Var}(X)+\operatorname{Var}(Y)+2\operatorname{Cov}(X,Y)$; if $X \perp\!\!\!\perp Y$, this reduces to $\operatorname{Var}(X+Y)=\operatorname{Var}(X)+\operatorname{Var}(Y)$.

The k-th *moment* is $\mathbb{E}[X^k]$ (when finite); the k-th *central moment* is $\mathbb{E}[(X-\mathbb{E}[X])^k]$. Skewness and kurtosis are (scaled) third and fourth central moments, respectively.

Covariance and correlation. For square-integrable X, Y,

$$\operatorname{Cov}(X,Y) \;:=\; \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] \;=\; \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y].$$

As you might have guessed, these objects also have some neat properties.

- Bilinear and symmetric:

  $$\operatorname{Cov}(aX+bY,Z)=a\operatorname{Cov}(X,Z)+b\operatorname{Cov}(Y,Z)$$

  and

  $$\operatorname{Cov}(X,Y)=\operatorname{Cov}(Y,X)$$

- Cauchy–Schwarz: $|\operatorname{Cov}(X,Y)| \le \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$.
- Independence $\Rightarrow$ zero covariance, but not conversely (uncorrelated $\not\Rightarrow$ independent).